August 2022

Medical Image Segmentation with Deep Convolutional Neural Networks

Chuanbo Wang
University of Wisconsin-Milwaukee

MEDICAL IMAGE SEGMENTATION WITH DEEP CONVOLUTIONAL NEURAL NETWORKS

by

Chuanbo Wang

A Dissertation Submitting in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin-Milwaukee

August 2022

ABSTRACT

MEDICAL IMAGE SEGMENTATION WITH DEEP CONVOLUTIONAL NEURAL NETWORKS

by

Chuanbo Wang

The University of Wisconsin-Milwaukee, 2022
Under the Supervision of Zeyun Yu

Medical imaging is the technique and process of creating visual representations of the body of a patient for clinical analysis and medical intervention. Healthcare professionals rely heavily on medical images and image documentation for proper diagnosis and treatment. However, manual interpretation and analysis of medical images are time-consuming, and inaccurate when the interpreter is not well-trained. Fully automatic segmentation of the region of interest from medical images has been researched for years to enhance the efficiency and accuracy of understanding such images. With the advance of deep learning, various neural network models have gained great success in semantic segmentation and sparked research interests in medical image segmentation using deep learning. We propose two convolutional frameworks to segment tissues from different types of medical images. Comprehensive experiments and analyses are conducted on various segmentation neural networks to demonstrate the effectiveness of our methods. Furthermore, datasets built for training our networks and full implementations are published.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Zeyun Yu, who guided me throughout my program and help me overcome obstacles. I would also like to acknowledge Dr. Gopalakrishnan, my lab mates, and friends who helped me and offered deep insight into the study. Special thanks to my family, especially my wife for being my spiritual prop along the journey.

Chapter 1

Introduction

1.1 Semantic Segmentation

Semantic segmentation is one of the most important tasks in computer vision. The goal of semantic segmentation is to find the boundary between the object of interest and the background. More specifically, semantic segmentation aims at classifying each image pixel with a class label. Since the achievements AlexNet [13] accomplished in the ImageNet large-scale visual recognition challenge [14] in 2012, the success of deep learning in the domain of computer vision sparked interest in semantic segmentation [15] using deep convolutional neural networks (CNN) [16].

1.1.1 Supervised Learning Methods

As one of the early segmentation models based on CNN, Long et al. proposed fully convolutional networks (FCN) [17] for pixel-wise semantic segmentation. One of their main contributions is using up-sampling layers to generate the output activation maps. The output is fused with the output of shallower layers to preserve the contextual spatial information of an image. The next high-impact segmentation model is SegNet [68] which introduces the encoder-decoder architecture. As the name suggested, this architecture consists of two sub-networks: the encoder network and the decoder network. The encoder network consists of convolutional layers and pooling layers like [17]. The decoder network is mapping the low-resolution encoder feature to input resolution feature maps. The main contribution of this work is the way that the lower resolution feature maps being upsampled in the decoder. Pooling indices computed in the encoder's max-pooling layer are used by the corresponding decoder to upsample feature maps into higher resolution. Next, the skip connection is introduced to the encoder-decoder architecture by U-Net [45] and V-Net [69]. Skip connections skip some of the layers in the neural network and feed the output of one layer as the input to the next layers. In U-Net, the output of a layer in the encoder is fed into the corresponding layer in the decoder to improve segmentation accuracy and solve the vanishing gradients problem. Additionally, V-Net added skip connections to the 3D variant of U-Net and proposed a novel objective function based on the Dice coefficient, which is widely used and adopted in our proposed segmentation models. DenseNet [70] push the idea of skip connection further where each layer obtains additional inputs from all preceding layers and passes on its feature-maps to all subsequent layers. Following these works, several variants [72] [73] [74] [75] [76] [77] [78] of the encoder-decoder architecture have been applied to semantic segmentation with modifications like deeper or shallower models, and the addition of extra attention blocks.

Another well-adopted improvement in the deep semantic segmentation models is the spatial pyramid pooling (SPP) module. Following this idea, DeepLabv3+ is proposed in 2018 by Chen, Liang-Chieh, et al. [79] and outperformed state-of-the-art networks in the PASCAL VOC 2012 segmentation challenge [36] and the cityscapes segmentation challenge [80] [81]. The SPP module was initially introduced in [83] inserted before the output layer of CNN for object detection. When applied to semantic segmentation, the SPP module is used to obtain multi-scale context information by concatenating the feature maps in multiple scales or rates. DeepLabv3 [84] and its successor, DeepLabv3+, explore the multi-scale context information in a different way by using four parallel atrous convolutions with different atrous rates applied to handle segmenting the object at different scales. The benefits of SPP are twofold: 1. It removes the need for fixed-size image input. 2. It improves the robustness of object deformation.

Another research area in deep segmentation networks is reducing the high computational resources. Over the years, semantic segmentation neural networks evolve into deeper, wider, and more complicated models at the cost of high computational complexity and training time. One of the first modern efforts in this area is network pruning proposed in [85] by Han, Song, et al., in 2015, following the idea [86] [87] [88] proposed earlier in the 1990s. Next, [89] proposed a context-aware guiding module (CAGM) to improve existing pruning methods. CAGM allows the model to preserve the channels that always provide useful contextual clues under diverse inputs given that semantic segmentation emphasizes more on local-to-global features aggregation than image classification where most existing pruning methods ignore such channels. Moreover, Zhou, Zongwei, et al. proposed U-Net++

[90], a redesign of the skip connections of the U-Net architecture. U-Net++ focuses on model pruning during the inference time where the segmentation output is selected from only one of the segmentation branches formed from the new design of the skip connections. The DeepLabv3 model is also modified to reduce its computational complexity in [91]. The main contribution is the inverted residual with a linear bottleneck. This module takes as a low-dimensional compressed input that is first expanded to high-dimension representation and fed into a depth-wise convolution, which is more lightweight than normal convolution operations. Features are then subsequently projected back to a low-dimensional representation with a linear convolution.

The attention mechanism is one of the most studied methods improving the performance of the encoder-decoder model for machine translation. The basic idea of the attention mechanism is to utilize the most relevant input sequence in the decoder by adding weights to the encoded input vectors, where the highest weights are applied to the most relevant vectors. Many recent works explore combining the attention mechanism with semantic segmentation networks. Chen, Liang-Chieh, et al. [92] propose to jointly learn an attention model that softly weights the features from different input scales when predicting the semantic label of a pixel. The final output of our model is produced by the weighted sum of score maps across all the scales. Li, Hanchao, et al. [93] propose to combine attention mechanism and spatial pyramid to extract precise dense features for pixel labeling instead of complicated dilated convolution and artificially designed decoder networks. Fu, Jun, et al. [94] propose the dual attention network to adaptively integrate local features with their global dependencies. Two attention modules are introduced where the position attention module selectively aggregates the feature at each position by a weighted sum of the features at all positions, and the channel attention module selectively emphasizes interdependent channel maps by integrating associated features among all channel maps.

1.1.2 Semi-supervised Learning Methods

Deep learning models have been rapidly growing in capacity in the past decade, requiring a larger amount of data. However, it is very hard to annotate pixel-level ground truth labels for semantic segmentation at this scale. Semi-supervised learning (SSL) combines supervised learning and unsupervised learning to fully utilize the data including labeled and unlabeled data. More specifically, SSL models are trained with a small amount of labeled data and a large amount of unlabeled data, which is the case in many real-world applications. Over the recent years, many SSL methods have been proposed and can be roughly categorized into three groups: consistency-based methods, self-training methods, and generative (GAN) methods.

Consistency-based methods utilize the unlabeled data to enforce the cluster assumption in the trained model where it is assumed that the decision boundary must lie in low-density regions. The basic idea of consistency-based methods is that the prediction of an unlabeled sample should not be significantly different from the prediction of a realistic perturbation of this sample. Following the early works [95] [96] [97] [98] on consistency training, a cross-consistency unsupervised loss is proposed in [62]. The method is based on the observation that cluster assumption is violated at the input level but maintained where the class boundaries have high average distances in the output of the encoder for semantic segmentation networks. The main contribution is the introduction of a series of auxiliary decoders, each uses a perturbation function to generate synthesized data. Another line of research uses a data augmentation technique [101] that composites new images by mixing two original images. The new image contains pixels from one original image and other pixels from another original image. Following this idea, Olsson, Viktor, et al. propose Classmix [100] which uses the unlabeled data to synthesize new data and corresponding artificial labels, which are used to enforce the consistency regularization.

Self-training methods [102] [103] [104] are SSL algorithms that generate proxy labels on unlabeled data using the segmentation model trained on the labeled data. The best proxy labels and the corresponding unlabeled data are then selected to train the segmentation model. Applying this idea to semantic segmentation, a self-training method is proposed in [105]. Their main contribution is training the supervised model for multiple rounds. For each round, newly selected proxy labels and unlabeled data are added to the training dataset to train the model in the next round.

GANs introduce a smart way of training a generative model as a supervised learning problem with two sub-models: the generator model that is trained to generate new samples, and the discriminator model that tries to tell if the sample is generated or real. The two models are trained together where generated samples mixed with real samples are provided to the discriminator model. The discriminator is trained to be better at telling generated samples from real ones. Meanwhile, the generator is trained based on how well the generated samples fooled the discriminator. In this competition, the goal is that the generator can generate fake samples that the discriminator cannot tell, in other words, the discriminator classifies the provided samples as 50% real and 50% generated. Like almost every new development in CNN architectures, GANs are first applied to image classification as the DCGAN [110]. GANs are recently applied to semantic segmentation. For example, Dong, Xue, et al. propose a GAN model [106] that segments multiple organs on thoracic CT images. A set of U-Nets are used as generators and FCNs are utilized as discriminators. Following the Pix2Pix model [107], Cirillo M D, et al. propose the Vox2Vox model [108] that segments brain tumors from multi-modal MRI scans in the BraTS Challenge 2020 [109]. A CNN combining U-Net and ResNet is used as the generator, and a discriminator is adopted from Pix2Pix with 6 3D convolutional layers.

1.2 Semantic Segmentation for Medical Images

Acute and chronic nonhealing wounds represent a heavy burden to healthcare systems, affecting millions of patients around the world [48]. In the United States, Medicare cost projections for all wounds are estimated to be between $28.1B and $96.8B [49]. Unlike acute wounds, chronic wounds fail to predictably progress through the phases of healing in an orderly and timely fashion, thus requiring hospitalization and additional treatment adding billions in cost for health services annually [50]. Intervertebral discs (IVDs) are spine components that provide cushioning between adjacent vertebrae and absorb pressure on the spine. IVD disease is a common condition that affects about 5% of the population in developed countries each year [51], causing pain in the lower back and frequently in the neck and limbs as well. For accurate diagnosis and proper treatment planning of chronic wounds and IVD diseases, healthcare professionals rely heavily on medical images, including computed tomography (CT) images, magnetic resonance imaging (MRI) scans, and natural images taken in clinical settings. Such images are further measured and analyzed to provide quantitative parameters for the diagnosis and treatment. Traditionally, this process is performed manually by specialists. However, this process is tedious and time-consuming given the large number of images involved for each patient. Furthermore, the shortage of medical resources and clinicians in primary and rural healthcare settings decreases the access and quality of care for millions of Americans. Consequently, research interests in automatic segmentation and measurement from medical images were captured, especially in the fields of intervertebral disc segmentation from 3D MRI scans and wound segmentation from 2D images. Such studies can be roughly categorized into two groups: traditional methods and deep learning methods.

1.2.1 Traditional Machine Learning Methods

Studies in the first group focus on combining computer vision techniques and traditional machine learning approaches. These studies apply manually designed feature extraction to build a dataset that is later used to support machine learning algorithms. [1] proposed an algorithm to segment the wound region from 2D images. 49 features are extracted from a wound image using K-means clustering, edge detection, thresholding, and region growing in both grayscale and RGB. These features are filtered and prepared into a feature vector that is used to train a Multi-Layer Perceptron (MLP) and a Radial Basis Function (RBF) neural network to identify the region of a chronic wound. [2] proposed an IVD segmentation method applied to chest MRI scans. The method solves an energy minimization problem by graph-cuts algorithms where the graph edges are divided into two types: terminal edges that connect the voxels and non-terminal edges that connect neighbor voxels. [3] proposed to generate a Red-Yellow-Black-White (RYKW) probability map of an input image with a modified hue-saturation-value (HSV) model. This map then guides the region of interest (ROI) segmentation process using either optimal thresholding or region growing. [4] and [5] applied an atlas-based method that first proposes atlas candidates as initialization and then utilizes label fusion to combine IVD atlases to generate the segmentation mask. However, to generate the initialization, [4] registers IVD atlases to the localization obtained by integral channel features and a graphical parts model. Whereas [5] uses data-driven regression to create a probability map, which further defines an ROI as the initialization for segmentation. [6] demonstrated a wound segmentation method using an energy-minimizing discrete dynamic contour algorithm applied to the saturation plane of the image in its HSV color model. The wound area is then calculated from a flood fill inside the enclosed contour. Another regression-based IVD segmentation method [7] was proposed to address the segmentation of multiple anatomic structures in multiple anatomical planes from multiple imaging modalities with a sparse kernel machines-based regression. A 2D segmentation method proposed in [8] applied an Independent Component Analysis (ICA) algorithm to the pre-processed RGB images to generate hemoglobin-based images, which are used as input of K-means clustering to segment the granulation tissue from the wound images. These segmented areas are utilized as an assessment of the early stages of ulcer healing by detecting the growth of granulation tissue on the ulcer bed. [9] proposed a similar system to segment the burn wounds from 2D images. Cr-Transformation and Luv-Transformation are applied to the input images to remove the background and highlight the wound region. The transformed images are segmented with a pixel-wise

Fuzzy C-mean Clustering (FCM) algorithm. [10] proposes an automatic method using a conditional random field (CRF) based on super-voxels generated from a variant of simple linear iterative clustering (SLIC). A support vector machine (SVM) is then used to perform super-voxels classification, which is later integrated into the potential function of the CRF for final segmentation using graph cuts. [11] builds an automatic IVD segmentation framework that localizes the vertebral bodies using regression-forests-based landmark localization and optimizes the landmarks by a high-level Markov Random Field (MRF) model of global configurations. The IVD segmentation mask is then generated from an image processing pipeline that optimizes the convex geodesic active contour based on the geometrical similarity to IVDs. In [12], IVD segmentation is performed by iteratively deforming the corresponding average disc model towards the edge of each IVD, in which edge voxels are defined by a 26-dimension feature vector including intensity, gradient orientation and magnitude, self-similarity context (SSC) descriptor, and Canny edge descriptor, etc. This group of methods suffers from at least one of the following limitations: 1) As in many traditional computer vision systems, the computation complexity is high in the segmentation pipeline, 2) They depend on manually tuned parameters and empirically handcrafted features which does not guarantee an optimal result. Additionally, they are not immune to severe pathologies and rare cases, which are very impractical from a clinical perspective, and 3) The performance is evaluated on a small, biased dataset.

1.2.2 Deep Learning Methods

1.2.2.1 Supervised Learning

Traditional computer vision and machine learning methods typically make decisions based on feature extraction. Thus, to find the segmentation mask, one must guess which wound features are important and then design sophisticated algorithms that capture these features. However, in CNN, feature extraction and decision-making are integrated. The features are extracted by convolutional kernels and their importance is determined by the network during the training process. A typical CNN architecture consists of convolutional layers and a fully connected layer as the output layer, which requires fixed-size inputs. One successful variant of CNN is fully convolutional neural networks (FCN) [17]. Networks of this type are composed of convolutional layers without any fully connected layer at the end of the network. This enables the network to take arbitrary input sizes and prevent the loss of spatial information caused by the fully connected layers in CNNs. Several FCN-based methods have been proposed to solve the wound segmentation problem. [18] estimated the wound area by segmenting wounds with the vanilla FCN architecture [17]. With time-series data consisting of the estimated wound areas and corresponding images, wound healing progress is predicted using a Gaussian process regression function model. However, the mean Dice accuracy of the segmentation is only evaluated to be 64.2%. [19] proposed to employ the FCN-16 architecture on the wound images in a pixel-wise manner that each pixel of an image is predicted to which class it belongs. The segmentation result is simply derived from the pixels classified as a wound. By testing different FCN architectures, they can achieve a Dice coefficient of 79.4% on their dataset. However, the network's segmentation accuracy is limited in distinguishing small wounds and wounds with irregular borders as the tendency is to draw smooth contours. [20] proposed a new FCN architecture that replaces the decoder of the vanilla FCN with a skip-layer concatenation up-sampled with bilinear interpolation. A pixel-wise SoftMax layer is appended to the end of the network to produce a probability map, which is post-processed to be the final segmentation. A dice accuracy of 91.6% is achieved on their dataset with 950 images taken under an uncontrolled lighting environment with a complex background. However, images in their dataset are semi-automatically annotated using a watershed algorithm. This means that the deep learning model is learning how the watershed algorithm labels wounds as opposed to human specialists.

FCNs are also adopted to solve the IVD segmentation problem. [21] extends the 2D FCN into a 3D version with end-to-end learning and inference. [22] proposes a 3D multi-scale FCN that expands the typical single-path FCN to three pathways where each pathway takes volumetric regions on a different scale. Features from three pathways are then concatenated to generate a probability map, from which the final 3D segmentation mask is generated by simple thresholding. More recently, a modified FCN, U-Net [45], and its variants have outperformed the state-of-art in many biomedical image segmentation tasks. The pertinacious architecture and affluent data augmentation allow U-Net to quickly converge to the optimal model from a limited number of annotated samples. Compared to CNN and vanilla FCN, U-Net uses skip connections between contraction and expansion and a concatenation operator instead of a sum, which could provide more local information to global information while expanding. Moreover, U-Net is symmetric such that feature maps in an expansive path facilitate to transfer more information. U-Net has been widely applied to the IVD segmentation problem. [23] applies a conventional 3D U-Net [24] on the IVD dataset provided by the 3rd MICCAI Challenge [25] of Intervertebral Discs Localization and Segmentation. [26] designs a new network architecture based on U-Net, boundary specific U-Net (BSU). The architecture consists of repeated application of BSU pooling layers and residual blocks, following the idea of residual neural networks (RNN). [27] extends the conventional U-Net by adding three identical pathways in the contracting path to process the multi-modality channels of the input. These pathways are interconnected with hyper-dense connections to better model relationships between different modalities in the multi-modal input images. [28] proposes an IVD segmentation pipeline that first segments the vertebral bodies using a conventional 2D U-Net to find the spine curve and IVD centers. Transverse 2D images and sagittal 3D patches are cropped around the centers to train an RNN fusing both 2D and 3D convolutions. However, the effectiveness of data augmentation and multi-modality input images are not fully explored in these works.

1.2.2.2 Semi-supervised Learning Methods

Tissue growth is a key indicator of wound healing and there has been growing research interest in monitoring the growth of different types of tissue. Due to limited annotation, SSL is a popular option that combines a small amount of labeled data with a large amount of unlabeled data during training. SSL makes use of unlabeled data to improve deep learning models to an extent that the performance of SSL methods is close to supervised methods trained with significantly more labeled data.

Recent efforts in SSL spark interest in medical image segmentation with SSL. Zhou, et al. proposed an SSL model [57] based on a generative adversarial network (GAN) with an attention mechanism to segment lesions from retina images and detect diabetic retinopathy. Bortsova, et al. proposed to segment abnormality from chest X-Ray images using a consistency-based SSL model [58] with U-Net as the backbone network. Peng, et al. proposed to use a clustering loss based on mutual information [59] that explicitly enforces prediction consistency between random perturbations of the unlabeled input. Chen, et al. proposed an attention-based SSL method [60] to segment brain tumors from MRI scans. In this method, an autoencoder is trained to reconstruct synthetic segmentation masks created by the attention mechanism. Sedai, et al. proposed an uncertainty guided SSL method [61] based on student-teacher learning for retinal layer segmentation from optical coherence tomography scans. Motivated by abundant unlabeled data and limited annotations, we propose to tackle the segmentation of different types of tissues in wound images using an SSL model based on cross-consistency training (CCT) [62], which is barely applied to deal with wound images.

To better explore the capacity of deep learning in the wound segmentation, tissue segmentation, and IVD segmentation problem, we propose three frameworks to automatically segment ROI from medical images. The first framework proposed is built upon a 3D network, 3D U-Net [24], to segment IVD from MRI scans. We adopted a two-stage pipeline: localizing the IVDs followed by segmenting IVDs based on the localization. To examine the effectiveness of different combinations of modalities, various modalities are analyzed with respect to image properties of the input data based on our analysis in the conducted experiments. The second framework is built above a 2D network, MobileNetsV2 [29], to tackle the wound segmentation problem. This network is lightweight and computationally efficient since significantly fewer parameters are used during the training process. We built a large dataset of wound images with segmentation annotations done by wound specialists. This is by far the largest dataset focused on wound segmentation to the best of our knowledge. The third model is proposed based on the cross consistency training model where we propose new auxiliary decoders with realistic perturbation functions in wound images.

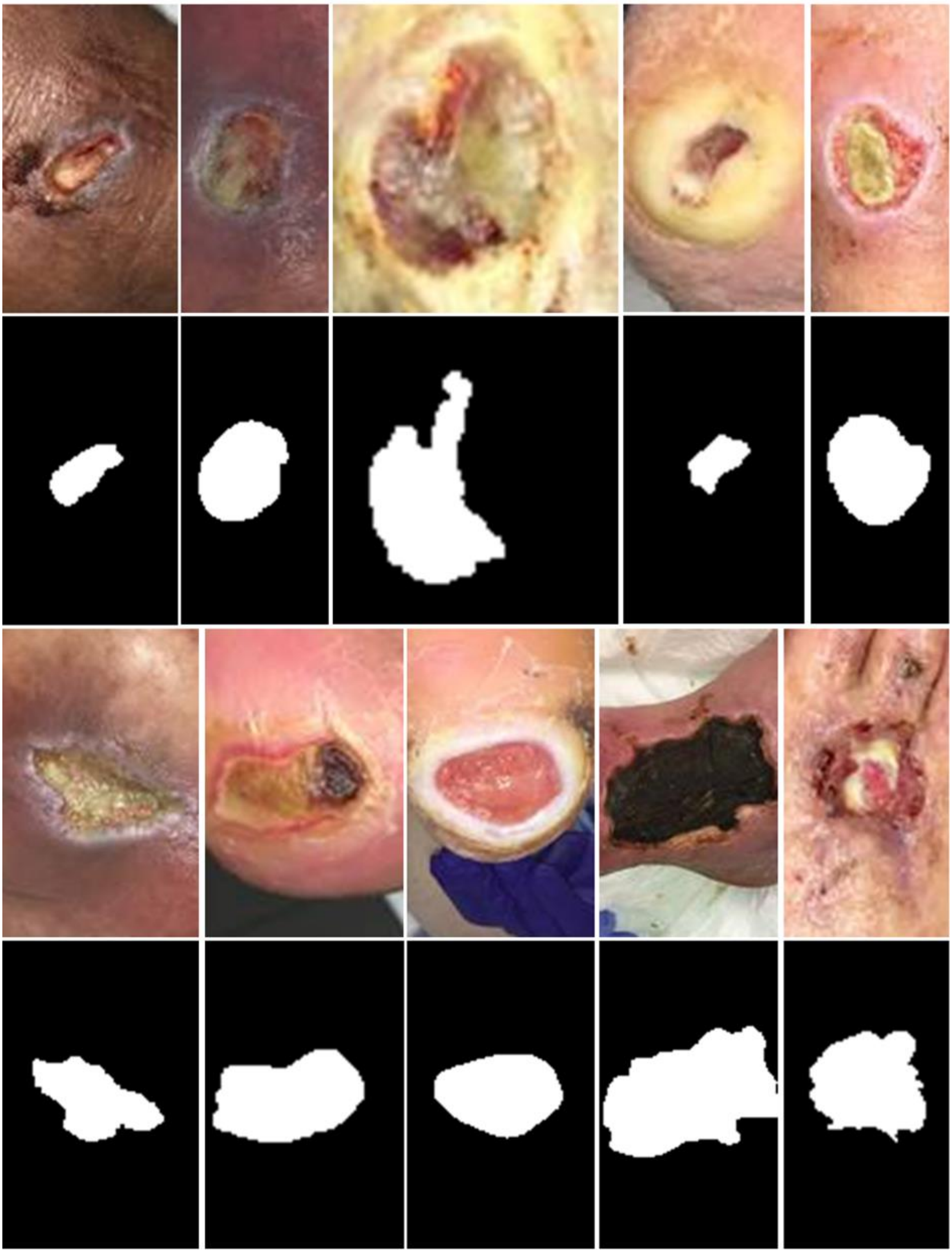


Figure 1 An illustration of images in our dataset. The first row contains the raw images collected. The second row consists of segmentation mask annotations we create with the AZH wound and vascular center

Chapter 2

Datasets

In this section, we introduce the datasets used in wound segmentation, tissue segmentation, and IVD segmentation. The wound dataset and the tissue dataset are created by our lab and contain wound images collected from our collaborating wound clinic. Thus, the construction methods and the annotation process are described in detail.

2.1 The Wound Dataset

2.1.1 Dataset Construction

There is currently no public dataset large enough for training deep-learning-based models for wound segmentation. To explore the effectiveness of wound segmentation using deep learning models, we collaborated with the Advancing the Zenith of Healthcare (AZH) Wound and Vascular Center, Milwaukee, WI. Our chronic wound dataset was collected over 2 years at the center and includes 1,210 foot ulcer images taken from 889 patients during multiple clinical visits. The raw images were taken by digital single-lens reflex cameras and iPads under uncontrolled illumination conditions, with various backgrounds. Figure 1 shows some sample images in our dataset.

The raw images collected are of various sizes and cannot be fed into our deep learning model directly since our model requires fixed-size input images. To unify the size of images in our dataset, we first localize the wound by placing bounding boxes around the wound using an object localization model we trained de novo, YOLOv3 [31]. Our localization dataset contains 1,010 images, which are also collected from the AZH Wound and Vascular Center. We augmented the images and built a training set containing 3645 images and a testing set containing 405 images. For training our model we have used LabelImg [32] to manually label all the data (both for training and testing). The YOLO format has been used for image labeling. The model has been trained with a batch size of 8 for 273 epochs. With an intersection over union (IoU) rate of 0.5 and non-maximum suppression of 1.00, we get the mean Average Precision (mAP) value of 0.939. In the next step, image patches are cropped based on the bounding boxes resulting from the localization model. We unify the image size (224 pixels by 224 pixels) by applying zero padding to these images, which are regarded in our dataset data points.

2.1.2 Data Annotation

During training, a deep learning model is learning the annotations of the training dataset. Thus, the quality of annotations is essential. Automatic annotation generated with computer vision algorithms is not ideal when deep learning models are trained to learn how human experts recognize the wound region. In our dataset, the images were manually annotated with segmentation masks that were further

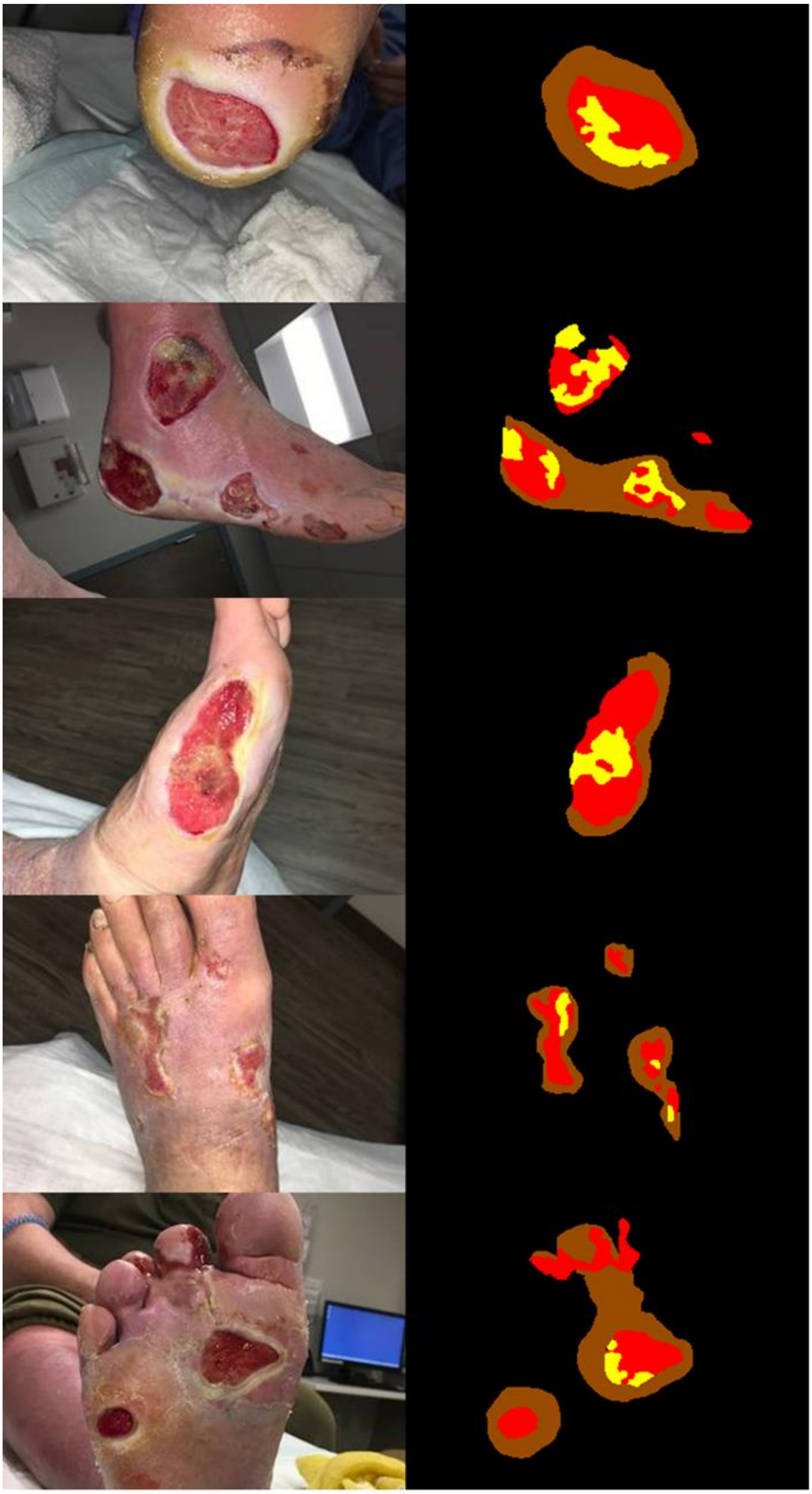


Figure 2 Images and annotations of the tissue dataset.

reviewed and verified by wound care specialists from the collaborating wound clinic. Initially, only foot ulcer images were annotated and included in the dataset as these wounds tend to be smaller than other types of chronic wounds, which makes it easier and less time-consuming to manually annotate the pixel-wise segmentation masks. In the future, we plan to create larger image libraries to include all types of chronic wounds, such as venous leg ulcers, pressure ulcers, and surgery wounds as well as non-wound reference images. The AZH Wound and Vascular Center, Milwaukee, WI, had consented to make our dataset publicly available.

2.2 The Tissue Dataset

Under the same annotation protocol, we also start to label different types of tissue inside our wound images (Figure 2). It is more challenging to annotate wound tissue due to its complexity and unpredictable shape. Until now, there are 110 labeled images and 1358 unlabeled images in our tissue dataset. Table 1 shows the appearance count of each type of tissue and we can see that three major types of tissue are granulation, callous, and fibrin tissue. This is the main reason that we focus our tissue segmentation model on these three types of tissue, which will be discussed in more detail in Chapter 5.

Table 1 The appearance count for each type of tissue in 110 labeled images.

Granulation	Callous	Fibrin	Necrotic	Eschar	Neodermis	Tendon	Dressing
93	86	74	24	11	11	2	2

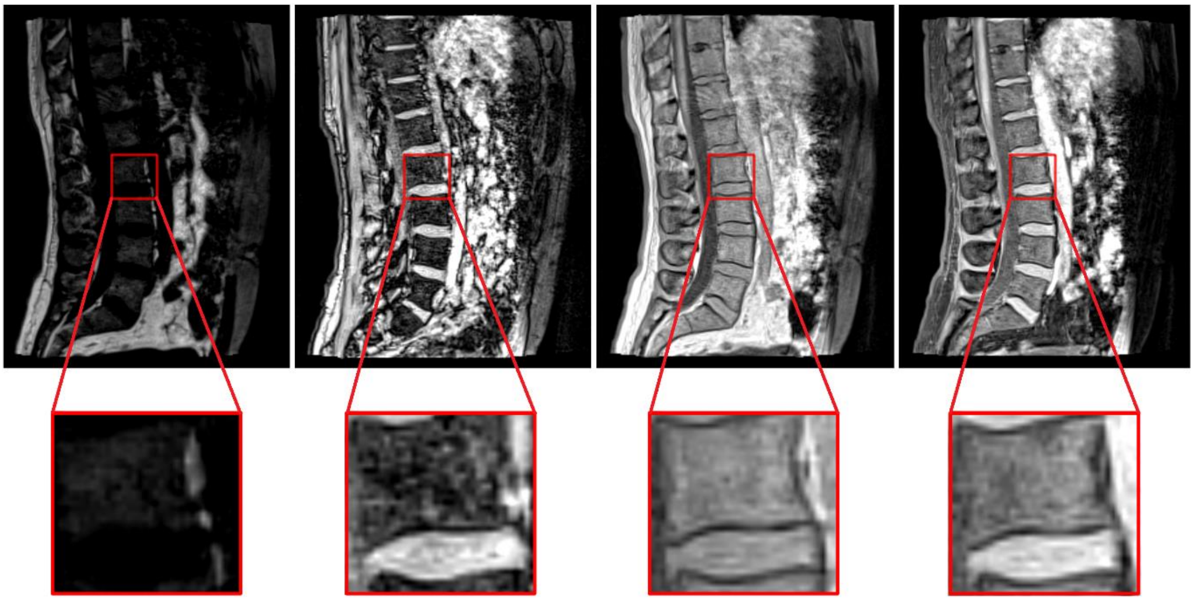


Figure 3 The comparison of contrast in different modalities of images (The modalities from left to right are respectively fat, opposed-phase, in-phase, water)

2.3 The IVD Dataset

The IVD dataset [30], by courtesy of Prof. Guoyan Zheng from the University of Bern, consists of 8 sets of 3D multi-modality MRI spine images collected from 8 patients in 2 different stages of prolonged bed tests. Each spine image contains at least 7 IVDs of the lower spine (T1-L5) and four modalities following Dixon protocol: in-phase (inn), opposed-phase (opp), fat, and water (wat) images as illustrated in Figure 3. In detail, water images are spin echo images acquired from water signals. fat images are spin echo images acquired from water signals. In-phase images are the sum of water images and fat images. Opposed-phase images are the difference between water images and fat images. In total, there are 32 3D single-modality volumes and 66 IVDs. For each IVD, the ground truth is composed of binary masks manually labeled by three trained raters under the guidance of clinicians.

Chapter 3

IVD Segmentation

3.1 Methods

3.1.1 Multi-modality Analysis

The traditional multi-modality deep learning methods conventionally fed all modalities images into different channels as input to the neural network. Zhang et. al. use multi-modality images to segment infant brains [112]. Li et. al. trained the 3D FCN separately on every single modality (fat, in-phase, opposed-phase, and water) and then on a merged full-modality of the spine dataset to validate the superiority of training with multi-modality data. Both works show that training on the full-modality images could yield more accurate IVD segmentation results than the single-modality strategy [111]. Moreover, Li's research indicates that the single-modality of opposed-phase images and water images could enhance the performance than using the modality of fat images and in-phase images. Nonetheless, the full-modality-fused images cannot guarantee to achieve their full capacity efficiently in all experiments. The dependency between diverse modalities could lead to data co-adaption, which means the same features are detected repeatedly. In addition, the significant difference between modalities at the same locations could trigger the data corruption problem to misdirect the corresponding neurons.

To alleviate these problems, we made a further exploration of the intensity distribution of various modalities and analyzed the mean gradient on the boundary of the intervertebral discs. As shown in Figure 3, the images of the fat modality have obviously low contrast at the edges compared with the images of the other three modalities. In more detail, assume the target intervertebral discs are foreground and the other areas are background and Table I lists the average intensities and the standard deviations for the foreground and background samples using the provided label maps as a reference. Moreover, Table 2 also shows the absolute Weber contrast coefficients of in-phase, opposed-phase, fat and water images [3], which are computed by:

$$C = \frac{I - I_b}{I_b}$$

Where I is the mean intensity value of the foreground voxels and I_b is the mean intensity value of the voxels in the background. Compared to other modalities, fat and in-phase images have relatively low Weber contrast values. To take full advantage of multiple modalities of this dataset, we take a step further from [112] and train our network on different combinations of multi-modality images to examine the effectiveness of the fat and in-phase images. The comprehensive results will be presented later in the results section.

Table 2 Absolute Weber contrast

Modality	Mean ± SD Foreground Intensity	Mean ±SD Background Intensity	Absolute Weber Contrast
fat	15.9 ± 12.7	35.2 ± 47.5	0.57
in-phase	172.1 ± 39.9	97.9 ± 89.5	0.75
opposed-phase	155.4 ±47.5	63.2 ± 65.1	1.46
water	163.4 ± 41.1	67.9 ± 67.5	1.43

3.1.2 Two-stage strategy

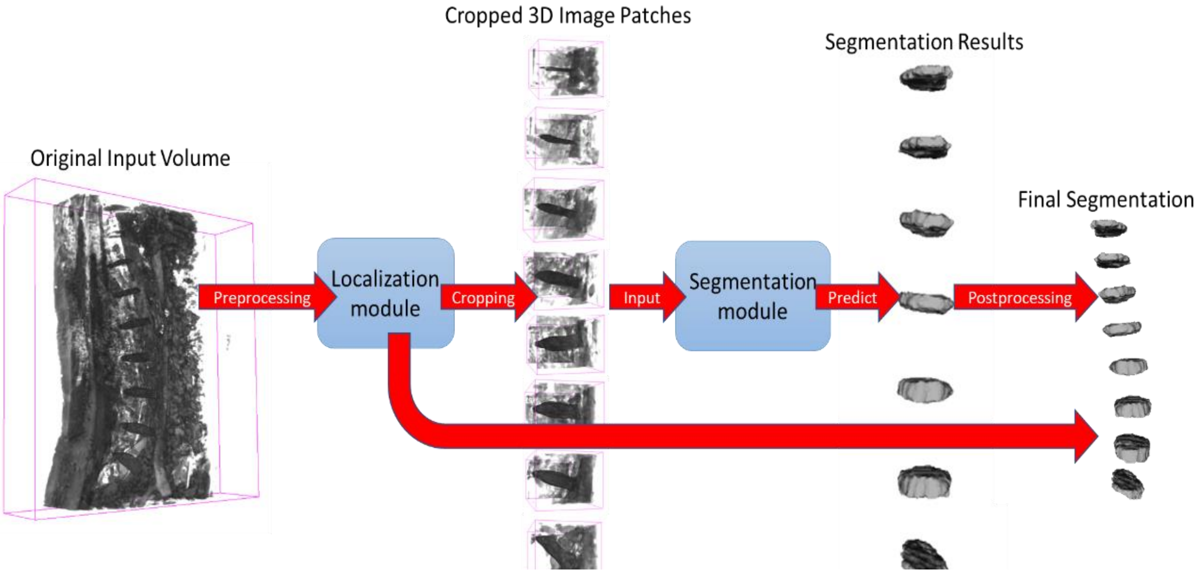


Figure 4 Workflow of the proposed 3D method.

In our proposed 3D method, a two-stage coarse-to-fine strategy is used to tackle the segmentation problem directly on 3D volumes. The general workflow is illustrated in Figure 4. In the first stage, each IVD is localized and a voxel is assigned as its center. These centers are used to divide the volume into small 3D patches, each of which contains a single IVD. In the second stage, a multimodal deep learning model is trained on the patches for precise segmentation.

Medical images are often complex and noisy where ROI is relatively small compared to the background. We first localize the IVDs in the image and then crop 3D patches based on the localization. This not only gets rid of some background but simplifies the problem for the segmentation stage and reduces the computational cost as well. It has been shown that 3D U-Net achieves the best localization result but not the best segmentation result [25] We use this two-stage strategy to work around this problem. In the end, post-processing is performed to generate the final segmentation.

3.1.3 Localization Network

For the localization of IVDs, we train a localization network, which is a conventional 3D U-net, on the IVD dataset to roughly locate the IVDs from the volume. From this segmentation, we have a good estimate of IVD centers by finding the center of each connected component after removing small regions.

From our observation, IVDs are generally sparsely located in 3D space at a distance from each other and share a common disc-like morphological profile. Thus, we simply put a 35*35*25 bounding box around each estimated center to crop a 3D patch. Then we zero-pad the patches to 36*36*28 so they can be nicely fed into the segmentation network in the next stage described below.

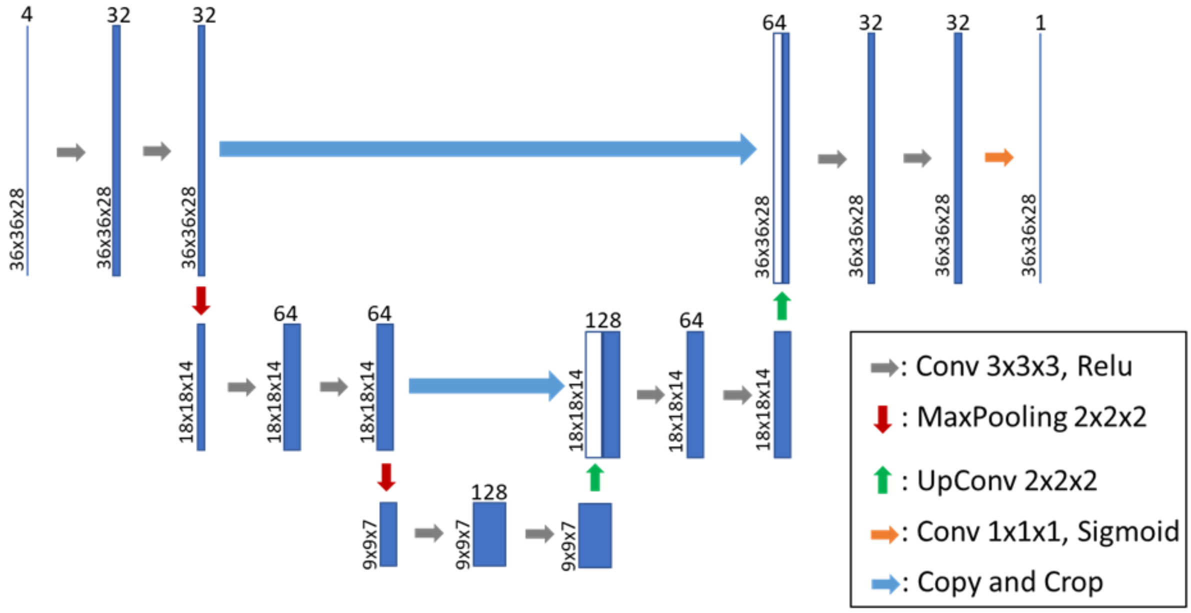


Figure 5 The network architecture of our 3D segmentation network.

3.1.4 3D Segmentation Network

For IVD segmentation from the 3D patches, we employ a modified 3D U-Net architecture that essentially looks at IVD segmentation as a regression problem. This network takes 3D patches as input and predicts 3D patches where the intensity value on each voxel stands for how confident the network is in the voxel belongs to an IVD. Figure 5 presents the network architecture of our 3D segmentation network. Each step in the contracting path consists of repeated application of two 3x3x3 unpadded 3D convolutions followed by a Relu. A dropout operation is inserted between the two convolutions to reduce the dependence on the training dataset and increase the accuracy. A dropout rate of 0.2 is used following the analysis [38] on the dropout effect in CNN. We also apply batch normalization to speed up and stabilize the training process and a 2x2x2 max pooling layer with stride 2 for down-sampling after every two convolutional layers. At each down-sampling step, we double the number of feature channels. Every step in the expansive path consists of an up-sampling of the feature map followed by a 2x2x2 up convolution that halves the number of feature channels, a concatenation with the corresponding feature map from the contracting path, and two 3x3x3 convolutions, each followed by a Relu. The output layer is a 1x1x1 convolution layer with sigmoid activation used to generate the segmentation mask for each modality. In total the network has 12 convolutional layers and 1.4 million parameters.

3.1.5 Data Augmentation

When only a few training samples are available, data augmentation is essential to increase the size of the desired dataset and increase the network's robustness to data variances. With only 8 sets of spine data provided, our networks demand more data for training. Especially for the 2D U-net, augmenting the training samples could significantly enhance the performance. A variety of conventional 3D image processing techniques such as translations, rotations, flip, and scaling are adopted in our method. In addition to these affine transformations, as a particularly important augmentation technique in biomedical segmentation tasks, elastic deformation is also used in combination with other transform functions since the most common variation in tissue is deformation [30]. We fabricate smoothly deformed surfaces by generating random displacement fields and the fields are convolved with a Gaussian of standard deviation δ (in pixels). And the displacement fields are then multiplied by a scaling factor α that controls the intensity of the deformation [30]. The per-pixel displacements are computed using bicubic interpolation. Based on the augmentation functions mentioned above, a random combination of operations is selected in an arbitrary order and applied to the original spine dataset. Consequently, the size of the training dataset is increased from 6 to 24 (2 of the 8 sets of original spine data are not used for augmentation but for validation of the prediction). To sufficiently take advantage of the information between adjacent layers, besides using the augmented spine data in the 3D framework, we also use the stack of images sliced along each axis of the 3D augmented data as the training set of the 2D network rather than directly applying 2D augmentation techniques on the sliced images from the original dataset.

3.1.6 Training

Both 2D and 3D networks in the presented work are implemented in python with Keras [41] and TensorFlow [42] backend. The batch size is set to 1 for the 3D network and 16 for the 2D network to obtain optimized training accuracy. The models are trained on a PC with an 8-core 3.4GHz CPU and a single NVIDIA GTX TITAN XP GPU. The training time of a single epoch takes about 3 s in the 3D model and 45 s in the 2D model. Eventually, the validation loss will stop increasing at around 10000 epochs for the 3D model and 500 epochs for the 2D model before overfitting.

The convolutional kernels of our networks are initialized with the HE initialization [44] to speed up the training process. For updating the parameters in the networks, we employ the Adam optimization algorithm [43], which has been popularized in the field of stochastic optimization due to its fast convergence compared to other optimization functions. The learning rate is set to 1e-5 for accurate predictions and reasonable training time. The energy function is computed by the Dice-coefficient loss function defined as:

$$\text{Loss} = \frac{2 \times |Graound\ truth \cap Prediction|}{|Graound\ truth| + |Prediction| + S}$$

Where S is a smoothing factor with a value of 1.

3.1.7 Post Processing

The prediction from the segmentation stage contains 3D patches with continuous voxel intensity values that represent how confident is the network in the voxel belonging to an IVD. The final segmentation mask for each patch is obtained by binary thresholding with a threshold of 0.5, which means voxels that are predicted more likely to be IVD voxels than background voxels are included in the segmentation mask. Then the mask patches are assembled back to a 3D volume of the lower spine, with the same size as the IVD dataset, using the IVD center locations from the localization stage and zero-padding.

3.2 Results

3.2.1 3D Evaluation Metrics

To evaluate the performance, two metrics are adopted from the 2015 MICCAI Challenge [25]. In addition to the Dice coefficient mentioned in section 4.1.1, we also calculated the Hausdorff distance (HD) that measures the distance between two surface meshes. We compute HD for surfaces reconstructed from the ground true segmentation mask and our segmentation result. Surfaces are generated using Iso2mesh [47] from binary segmentation masks. The closest distance from each vertex on the source surface mesh to the target surface mesh is found and HD is then computed. A smaller HD value indicates better segmentation performance.

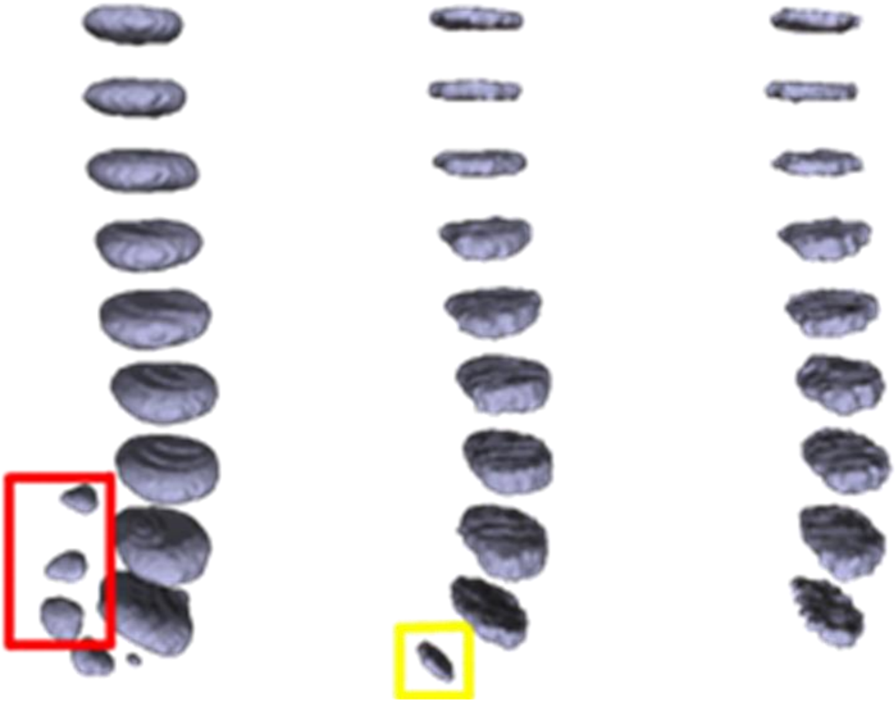


Figure 6 Training without in-phase images. Left: segmentation results of the baseline model trained with all modalities in the dataset. In the middle: segmentation results trained without in-phase images. Right: the ground true label.

3.2.2 Effectiveness of the multi-modality data

The segmentation results achieved by excluding the in-phase images from the training dataset are more accurate and less noisy near the lower IVDs than that in the original full-modality data as shown in Figure 6. Moreover, using the training dataset without in-phase images, our localization network can learn more details and make much more accurate predictions about the IVD centers. This makes the localization of centers more stable and allows us to simply remove small regions (marked by yellow boxes) and then crop a fix-size 3D patch for each IVD in the volume to train the segmentation network.

From the multi-modality analysis, we found that the fat and in-phase images have a significantly lower contrast among all the modalities. To analyze the effectiveness of the multi-modality input data, we train our 3D network on 4 different combinations of input modalities: 1) we train the network on full-modality images as the baseline, 2) we exclude the fat images from all 4 modalities to build the second training dataset, 3) the fat images are excluded from all 4 modalities, and 4) we only include oppose-phase and water images in the last training dataset. The mean Dice scores of the segmentation results predicted by the network trained on each dataset are presented in Table 2. Among all the different training settings, the network trained on full-modality images shows the worst segmentation performance. The reason is that the fat and in-phase images have a lower contrast, which means that the input values of the network are closer to each other and make it more difficult for the convolutional kernels to distinguish between them. It is worth pointing out that input normalization does not help with this situation because it is performed over the values of all the modalities. In other words, if we treat these 4 types of images equally during the training process, the fat and in-phase images confuse the network with their low image contrast.

Table 3. Segmentation performance of our 3D method using different combinations of modalities as the training dataset

Training dataset	Combination	Mean Dice ± SD
1)	opp, wat, fat, and inn	87.9% ± 1.7%
2)	opp, wat, and fat	89.0% ± 1.4%
3)	opp, wat, and inn	88.0% ± 1.6%
4)	opp, and wat	88.5% ± 1.6%

3.2.3 Comparison with state-of-the-art methods

To evaluate the performance of the proposed methods, we compare the segmentation results achieved by our methods with those by 3D U-Net[15], the CNN-based team UNICHK [23], and the winning team UNIJLU [12] in the test1 dataset of the 2015 MICCAI Challenge [25]. Quantitative results evaluated with the different architectures are presented in Table 3. The mean Dice score obtained by our 3D method is 89.0% with a standard deviation (SD) of 1.4%. We bring a 1.5% boost compared to the conventional 3D U-Net by training our network on 3D image patches extracted from opposed-phase, water, and fat images. This result is still 2.5% behind the state-of-the-art performance achieved by UNIJLU. The Mean HD of our 3D method reached 0.8 mm with an SD of 0.3 mm, which indicates that our method is slightly better when the segmentation results are reconstructed to 3D models. The strength of deep learning methods is the computation time. The Theano-based implementation of 3D U-Net from UNICHK takes 3.1s to process one 40 × 512 × 512 volume. Our network is implemented based on TensorFlow and it takes about 0.5s to segment all the IVDs in a 36 × 256 × 256 input volume. Overall, the computation time of our end-to-end segmentation is about 10s including localization, preprocessing, segmentation and postprocessing. Whereas it takes 5 min on average to segment all IVDs for a patient by UNIJLU. It is also worth mentioning that the training dataset used in our study only contains data from 6 patients while UNICHK and UNIJLU have access to a training dataset from 16 patients i.e., our network can learn the 3D geometric morphometrics of IVDs with much fewer data to learn from.

Table 4. Evaluation of the conventional 3D U-Net, UNICHK, UNIJLU, and our method.

Methods	Mean Dice ± SD	Mean HD ± SD
3D U-Net	87.5% ± 0.9%	1.1 ± 0.2
UNICHK	88.4% ± 3.7%	1.3 ± 0.2
UNIJLU	91.5% ± 2.3%	1.1 ± 0.2
Our method	89.0% ± 1.4%	0.8 ± 0.3

3.2.4 Effectiveness of Data Augmentation

To explore the effectiveness of data augmentation when applying 3D CNNs to IVD segmentation, we conducted experiments by training our 3D network with and without the proposed data augmentation method. The results do not suggest obvious differences in terms of Dice and HD. However, we found that data augmentation enables the network to learn more details in the boundary area. More specifically, segmentation results are improved in the regions between an IVD and the adjacent vertebral body. Examples are presented in Figure 7, the regions well segmented by our 3D network trained with augmented dataset are marked by red boxes. Segmentation of these regions with sharp boundaries is a difficult task for convolutional kernels since CNNs tend to predict smooth contour. The core of an IVD is composed of jelly-like material and the smooth elastic deformation technique in our data augmentation mimics the real-world deformation of IVDs and enriches the training dataset by adding more morphological variants. This makes CNNs more capable to deal with unseen IVD data and make better predictions.

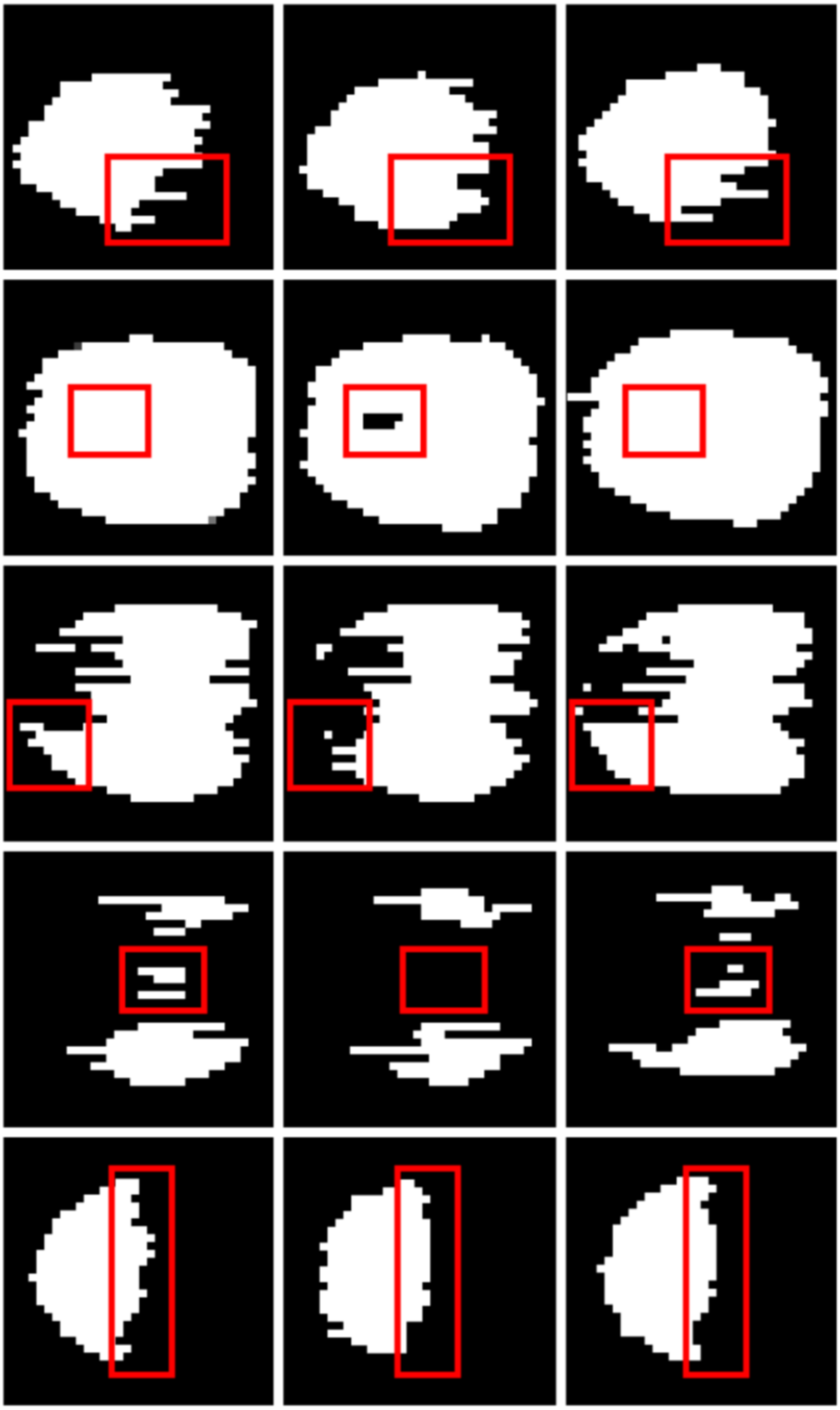


Figure 7 Effectiveness of data augmentation. Each column represents a MRI slice. The segmentation mask predicted by 3D U-Net with data augmentation, 3D U-Net without data augmentation and the ground truth label of the slice are shown, respectively, in the transverse view.

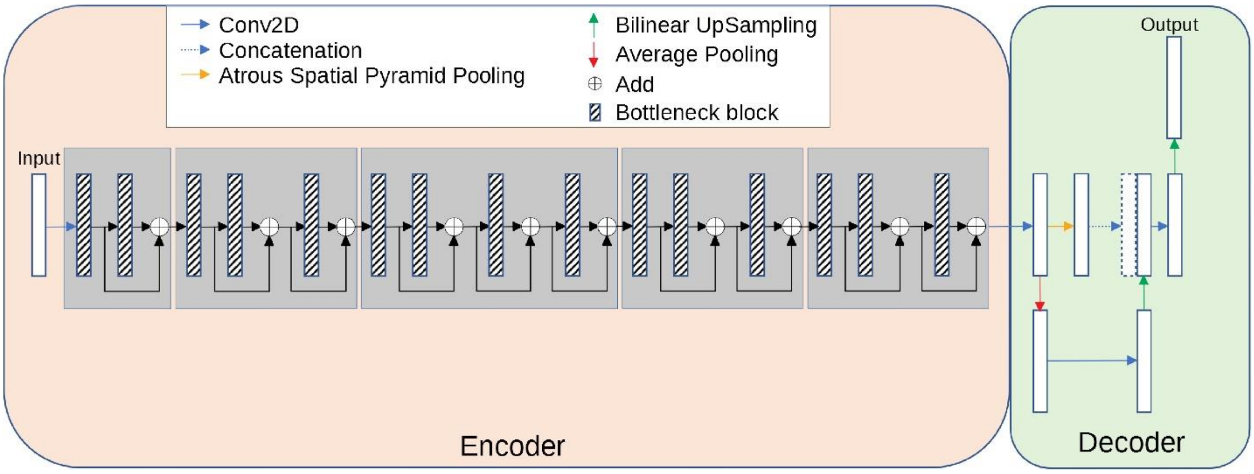


Figure 8 The proposed encoder-decoder architecture.

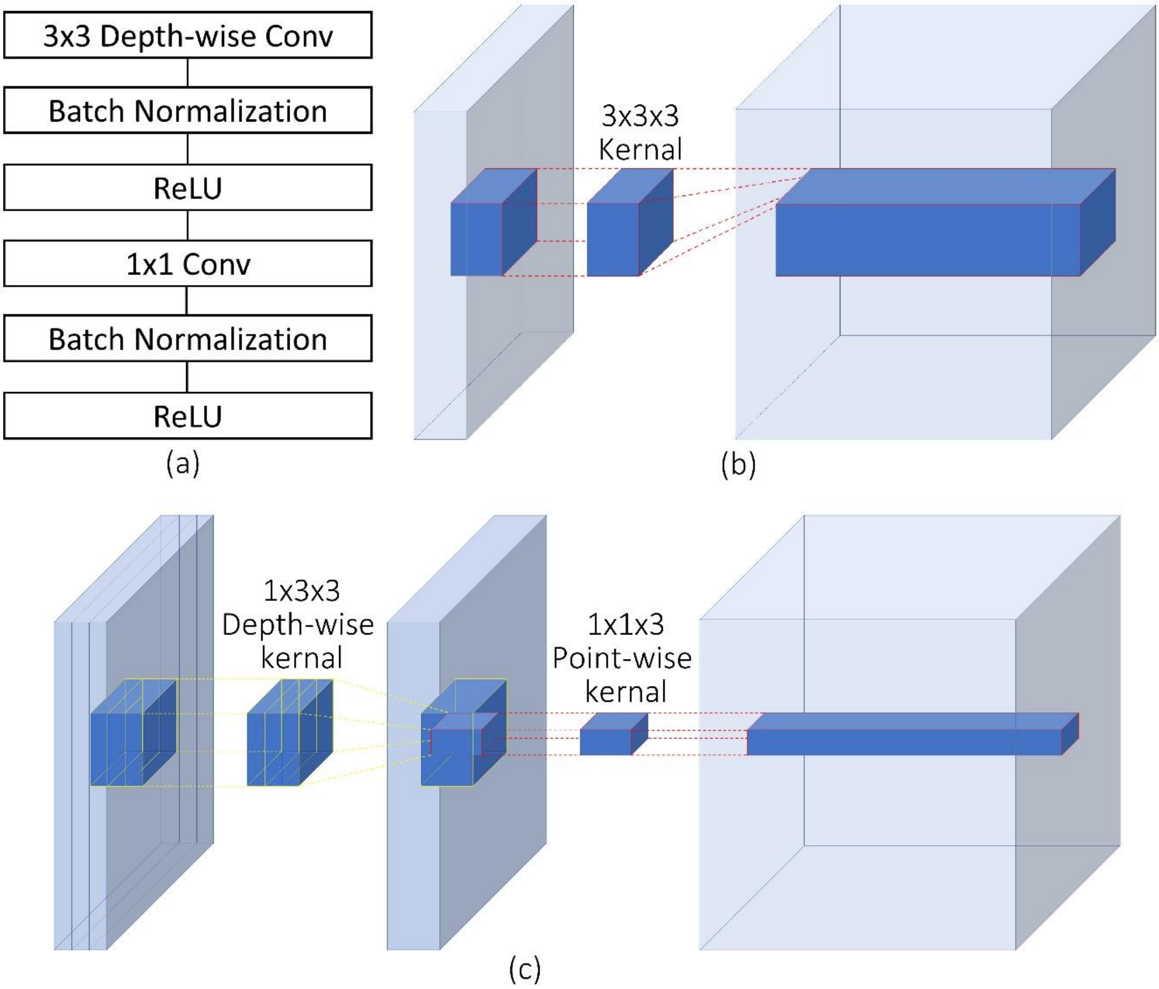


Figure 9 (a). A depth-separable convolution block. The block contains a 3 × 3 depth-wise convolutional layer and a 1 × 1 point-wise convolution layer. Each convolutional layer is followed by batch normalization and Relu6 activation. (b) An example of a convolution layer with a 3 × 3 × 3 kernel. (c) An example of a depth-wise separable convolution layer equivalent to (b).

Chapter 4

Wound Segmentation

4.1 Methods

In this section, we describe the methods we used with the architectures of the deep learning models for wound segmentation. The transfer learning used during the training of our model and the post-processing methods including hole filling and removal of small noises is also described.

4.1.1 Pre-processing

Besides cropping and zero-padding discussed in the dataset construction section, standard data augmentation techniques are applied to our dataset before being fed into the deep learning model. These image transformations include arbitrary rotations in the range of +25 to -25 degrees, random left-right and top-down flipping with a probability of 0.5, and random zooming within 80% of the original image area. Random zooming is performed as the only non-rigid transformation because we suspect that other non-rigid transformations like shearing do not represent common wound shape variations. Eventually, the training dataset is augmented to around 5000 images. We keep the validation dataset unaugmented to generate convincing evaluation outcomes.

4.1.2 Model Architecture

A convolutional neural network (CNN), MobileNetV2 [29], is adopted to segment the wound from the images. Compared with conventional CNNs, this network substitutes the fundamental convolutional layers with depth-wise separable convolutional layers [33] where each layer can be separated into a depth-wise convolution layer and a point-wise convolution layer. A depth-wise convolution performs lightweight filtering by applying a convolutional filter per input channel. A point-wise convolution is a 1 × 1 convolution responsible for building new features through linear combinations of the input channels. This substitution reduces the computational cost compared to traditional convolution layers by almost a factor of k2 where k is the convolutional kernel size. Thus, depth-wise separable convolutions are much more computationally efficient than conventional convolutions suitable for mobile or embedded applications where computing resource is limited. For example, the mobility of MobileNetV2 could benefit medical professionals and patients by allowing instant wound segmentation and wound area measurement immediately after the photo is taken using mobile devices like smartphones and tablets. An example of a depth-wise separable convolution layer is shown in Figure 9 (c), compared to a traditional convolutional layer shown in Figure 9 (b).

The model has an encoder-decoder architecture as shown in Figure 8. The encoder is built by repeatedly applying the depth-separable convolution block (marked with diagonal lines). Each block, illustrated in Figure 9 (a), consists of six layers: a 3 × 3 depth-wise convolutional layer followed by batch normalization and rectified linear unit (Relu) activation [34], and a 1 × 1 point-wise convolution layer followed again by batch normalization and Relu activation. To be more specific, Relu6 [35] was used as the activation function. In the decoder, shown in Figure 8, the encoded features are captured in multiscale with a spatial pyramid pooling block, and then concatenated with higher-level features generated from a pooling layer and a bilinear up-sampling layer. After the concatenation, we apply a few 3 × 3 convolutions to refine the features followed by another simple bilinear up-sampling by a factor of 4 to generate the final output. A batch normalization layer is inserted into each bottleneck block and a dropout layer is inserted right before the output layer. In MobileNetV2, a width multiplier α is introduced to deal with various dimensions of input images. we let α = 1 thus the input image size is set to 224 pixels × 224 pixels in our model.

4.1.3 Transfer Learning

To make the training more efficient, we used transfer learning for our deep learning model. Instead of randomly initializing the weights in our model, the MobileNetV2 model, pre-trained on the Pascal VOC segmentation dataset [36] is loaded before the model is trained on our dataset. Transfer learning with the pre-trained model is beneficial to the training process in the sense that the weights converge faster and better.

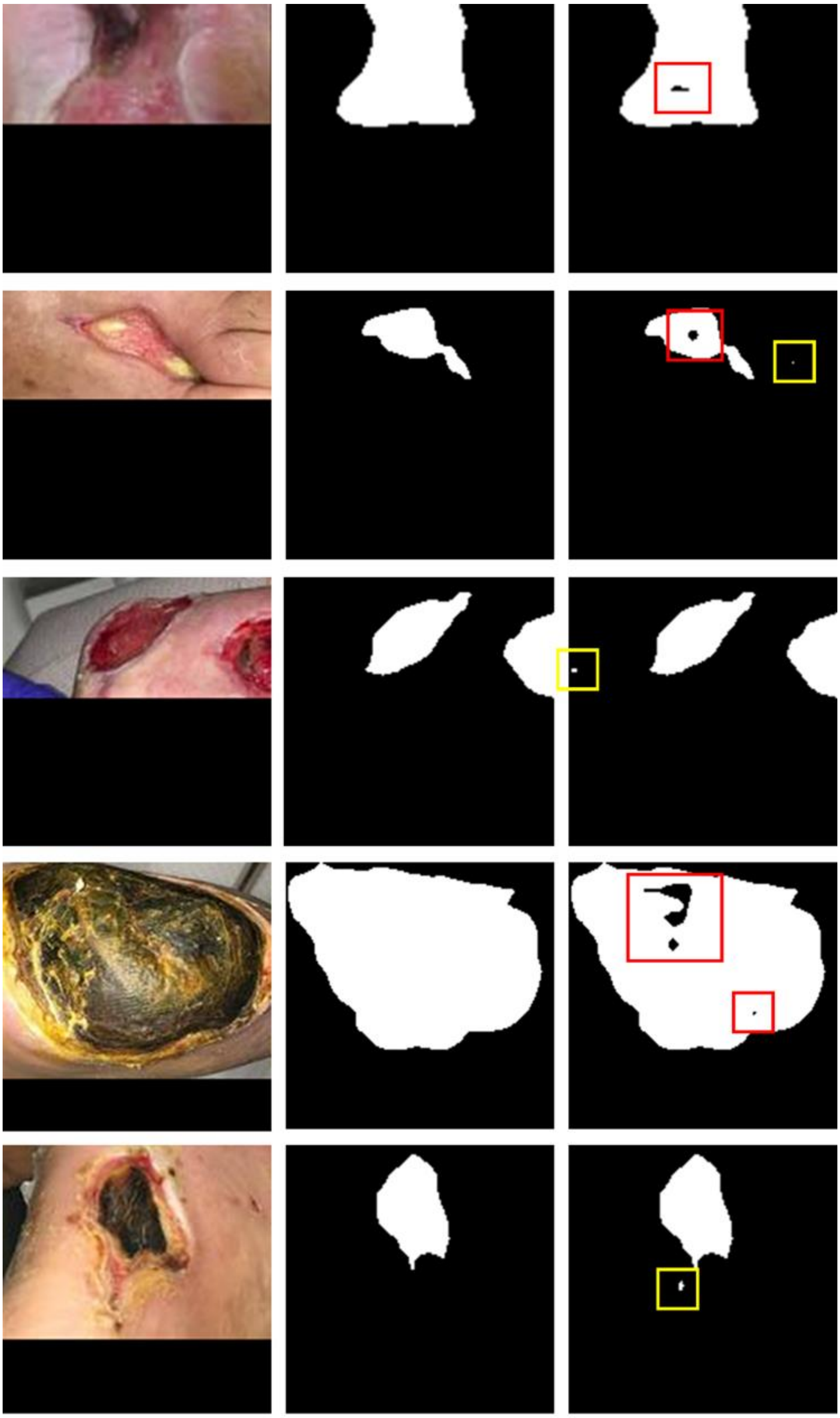


Figure 10 An illustration of the segmentation result and the post processing method. The first row illustrates images in the testing dataset. The second row shows the segmentation results predicted by our model without any post processing. The holes are marked with red boxes and the noises are marked with yellow boxes. The third row shows the final segmentation masks generated by the post processing method.

4.1.4 Post Processing

Post Processing, including hole filling and removal of small regions, is performed to improve the segmentation results as shown in Figure 10. We notice that abnormal tissue like fibrinous tissue within chronic wounds could be identified as non-wound and cause holes in the segmented wound regions. Such holes are detected by finding small connected components in the segmentation results and filled to improve the true positive rate using connected component labeling (CCL) [37]. The small noises are removed in the same way. The images in the dataset are cropped from the raw image for each wound. So, we simply remove noises in the results by removing the connected component small enough based on adaptive thresholds. To be more specific, a connected region is removed when the number of pixels within the region is less than a threshold, which is adaptively calculated based on the total number of pixels segmented as wound pixels in the image.

4.2 Results

In this section, we describe the evaluation metrics and compare the segmentation performance of our method with several popular and state-of-the-art methods. Our deep learning model is trained with data augmentation and preprocessing. Extensive experiments were conducted to investigate the effectiveness of our network. FCN-VGG-16 is trained as the baseline model [19] [39]. For fairness of comparison, we used the same training strategies and data augmentation strategies throughout the experiments.

4.2.1 2D Evaluation Metrics

To evaluate the segmentation performance, Precision, Recall, and the Dice coefficient are adopted as the evaluation metrics [40]:

Precision: Precision shows the accuracy of segmentation. More specifically, Precision measures the percentage of correctly segmented pixels in the segmentation and is computed by:

$$\text{Precision} = \frac{True\ positives}{True\ positives + False\ positives}$$

Recall: Recall also shows the accuracy of segmentation. More specifically, it measures the percentage of correctly segmented pixels in the ground truth and is computed by:

$$\text{Recall} = \frac{True\ positives}{True\ positives + False\ negtives}$$

Dice coefficient (Dice): Dice shows the similarity between the segmentation and the ground truth. Dice is also called the F1 score as a measurement balancing Precision and Recall. More specifically, Dice is computed by the harmonic mean of Precision and Recall:

$$\mathrm{Dice} = \frac{2 \times True\ positives}{2 \times True\ positives + False\ negtives + False\ positives}$$

4.2.2 Training

The deep learning model in the presented work was implemented in python with Keras [41] and TensorFlow [42] backend. To speed up the training, the models were trained on a 64-bit Ubuntu PC with an 8-core 3.4GHz CPU and a single NVIDIA RTX 3090 GPU. For updating the parameters in the network, we employed the Adam optimization algorithm [43], which has been popularized in the field of stochastic optimization due to its fast convergence compared to other optimization functions. Binary cross entropy was used as the loss function and we also monitored Precision, Recall, and the Dice score as the evaluation matrices. The initial learning rate was set to 0.0001 and each minibatch contained only 2 images for balancing the training accuracy and efficiency. The convolutional kernels of our network were initialized with HE initialization [44] to speed up the training process and the training time of a single epoch took about 77 seconds. We used early stopping to terminate the training so that the best result was saved when there was no improvement for more than 100 epochs in terms of the Dice score. Figure 11 shows the training history of our model. Eventually, our deep learning model was trained for around 1200 epochs.

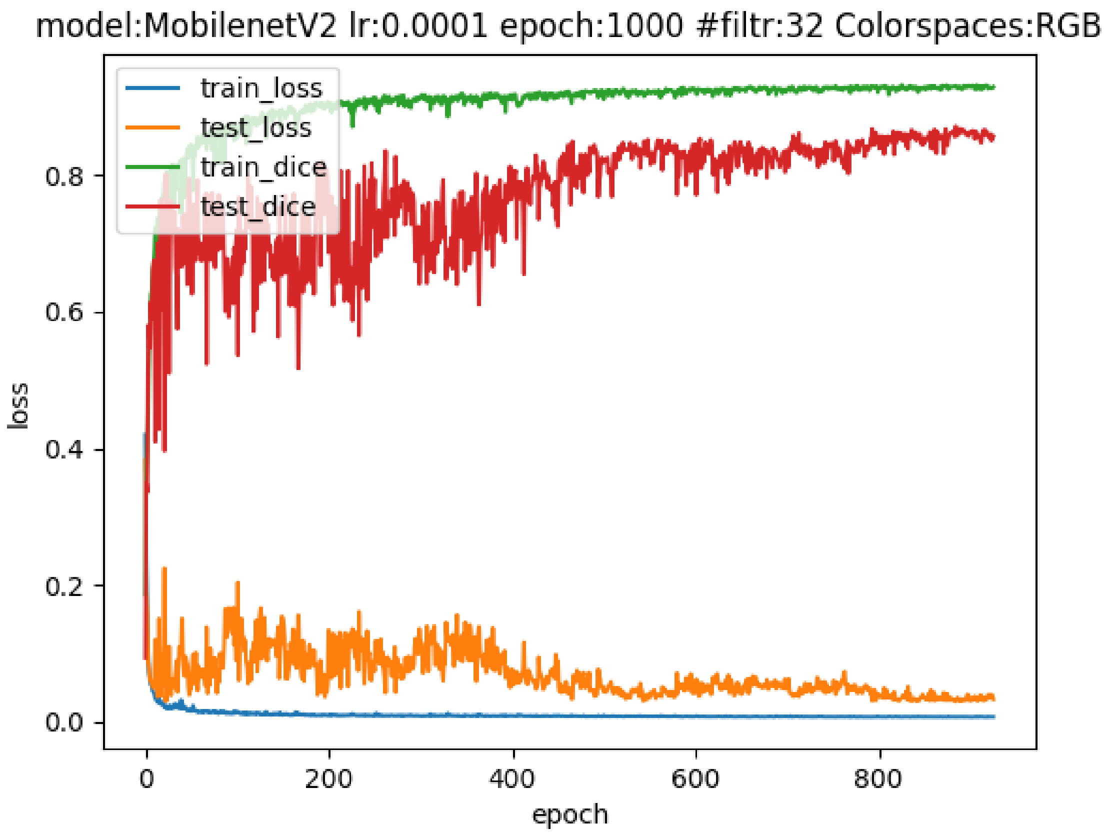


Figure 11 Training history of our model.

4.2.3 Comparing our method to the others

To evaluate the performance of the proposed method, we compared the segmentation results achieved by our methods with those by FCN-VGG-16 [19] [39], SegNet [68], and Mask-RCNN [55][56]. We also added 2D U-Net [45] to the comparison due to its outstanding segmentation performance on biomedical images with a relatively small training dataset. The segmentation results predicted by our model are demonstrated in Figure 10 along with the illustration of our post-processing method. Quantitative results evaluated with the different networks are presented in Table 4 where bold numbers indicate the best results among all four models. To better demonstrate the accuracy of the models, the numbers shown in the table are the highest possible number reached among various training.

Table 5. Evaluation on our dataset.

Model	VGG16	SegNet	U-Net	Mask-RCNN	MobileNetV2	MobileNetV2+CCL
Precision	83.91%	83.66%	89.04%	**94.30%**	90.86%	91.01%
Recall	78.35%	86.49%	**91.29%**	86.40%	89.76%	89.97%
Dice	81.03%	85.05%	90.15%	90.20%	90.30%	**90.47%**

In the performance measures, the Recall of our model was evaluated to be the second-highest among all models, at 89.97%. This was 1.32% behind the highest Recall, 91.29%, which was achieved by U-Net. Our model also achieved the second-highest Precision of 91.01%. Overall, the results show that our model achieves the highest accuracy with a mean Dice score of 90.47%. the VGG16 was shown to have the worst performance among all the other CNN architectures. Mask-RCNN achieved the highest Precision of 94.30%, which indicates that the segmentation predicted by Mask-RCNN contains the highest percentage of true positive pixels. However, the Recall is only evaluated to be 86.40%, meaning that more false-negative pixels are undetected compared to U-Net and MobileNetV2. Our accuracy was slightly higher than U-Net and Mask-RCNN, and significantly higher than SegNet and VGG16.

Comparing our model to VGG16, the Dice score is boosted from 81.03% to 90.47% tested on our dataset. Based on the appearance of chronic wounds, we know that wound segmentation is complicated by various shapes, colors, and the presence of different types of tissue. The patient images captured in clinic settings also suffer from various lighting conditions and perspectives. In MobileNetV2, the deeper architecture has more convolutional layers than VGG16, which makes MobileNetV2 more capable to understand and solve these variables. MobileNetV2 utilizes residual blocks with skip connections instead of the conventional convolution layers in VGG16 to build a deeper network. These skip connections bridging the beginning and the end of a convolutional block allow the network to access earlier activations that weren't modified in the convolutional block and enhance the capacity of the network.

Another comparison between U-Net and SegNet indicates that the former model is significantly better in terms of mean Dice score. Similar to the previous comparison, U-Net also introduces skip connections between convolutional layers to replace the pooling indices operation in the architecture of SegNet. These skip connections concatenate the output of the transposed convolution layers with the feature maps from the encoder at the same level. Thus, the expansion section which consists of a large number of feature channels allows the network to propagate localization combined with contextual information from the contraction section to higher resolution layers. Intuitively, in the expansion section or “decoder” of the U-

Net architecture, the segmentation results are reconstructed with the structural features that are learned in the contraction section or the “decoder”. This allows the U-Net to make predictions at more precise locations.

Besides the performance, our method is also efficient and lightweight. As shown in Table 5, the total number of trainable parameters in the adopted MobileNetV2 was only a fraction of the numbers in U-Net, VGG16, and Mask-RCNN. Thus, the network took less time during training and could be applied to mobile devices with less memory and limited computational power. Alternatively, higher-resolution input images could be fed into MobileNetV2 with less memory size and computational power compared to the other models.

Table 6. Comparison of total numbers of trainable parameters.

Model Name	FCN-VGG16	SegNet	U-Net	Mask-RCNN	MobileNetV2
Number of parameters	134,264,641	902,561	4,834,839	63,621,918	2,141,505

4.2.4 Comparison within the Medetec Dataset

Apart from our dataset, we also conducted experiments on the Medetec Wound Dataset [46] and compared the segmentation performance of these methods. The results are shown in Table 6. We annotated the dataset in the same way that our dataset was annotated and trained the networks with the same experimental setup. The highest Dice score is evaluated to be 94.05% using MobileNetV2+CCL. The performance evaluation agrees with the conclusion drawn from our dataset where our method outperforms the others regardless of which chronic wound segmentation dataset is used, thereby demonstrating that our model is robust and unbiased.

Table 7 Evaluation on the Medetec dataset.

Model	VGG16	SegNet	U-Net	Mask-RCNN	MobileNetV2
Dice	79.24%	72.94%	84.01%	93.20%	93.88%

Chapter 5

Tissue Segmentation

5.1 Problem

Accurate measurement of different types of tissues within the wound is also critical to the assessment and management of chronic wounds to monitor the wound healing trajectory and to determine future interventions. We characterize wound tissues into three types as wound healing indicators: granulation, slough, and necrotic tissues. Granulations are light red or dark pink tissues featuring bumpy surfaces, which is a sign of healing and development of new tissues. Slough or fibrin tissues have yellowish color ranging from white to yellow or yellow green to brown depending on the bacterial colonization and hemoglobin content. Necrotic tissues contain dead cells that often appeal in black. In our dataset, wound care professionals did not remove dry necrotic tissues that protect the wound underneath. However, wet necrotic tissues, indicating the presence of bacteria, were salvaged from the wound. Like whole wound segmentation, public datasets with pixel-level annotations of wound tissues are not sufficient for training supervised deep learning models. Annotation of wound tissues is more complicated and requires a higher level of expertise because tissue boundaries are less clear than wound boundaries. Moreover, tissues like granulation develop randomly in shape during the healing process. With a large number of images and a limited number of annotations, we propose an SSL model based on cross-consistency training (CCT) that takes advantage of both the labeled and unlabeled data.

As an SSL method, our CCT model is trained with a small number of labeled images and a large number of unlabeled images. To formulate the problem, let $D_l = \{ \left(x_i^l, y_i\right) \}_{i=1}^{N_l}$ represent the labeled data and $D_u = \{ x_i^u \}_{i=1}^{N_u}$ represent the unlabeled data. Here x_i^l represents the i-th labeled input image and y_i represents its annotation. x_i^u represents the i-th unlabeled input image. The total number of labeled and unlabeled input images are N_l and N_u, respectively.

5.2 Methods

5.2.1 The Cluster Assumption

The basic idea of CCT methods is that the prediction of an unlabeled sample should not be significantly different from the prediction of a realistic perturbation of this sample. This allows the model to make consistent predictions over similar inputs with a limitation that the model works better only when decision boundaries locate in low density regions of the input, i.e., the cluster assumption. For semantic segmentation problems, the assumption holds in the output of the last layer in the encoder [65][66][67] instead of the input. This is one of the fundamental hypotheses of encoder-decoder-like semi-supervised segmentation architectures. To investigate this in our dataset, we extracted the feature map from the last layer of the encoder and calculate the average distance between each pixel and its neighbors. These

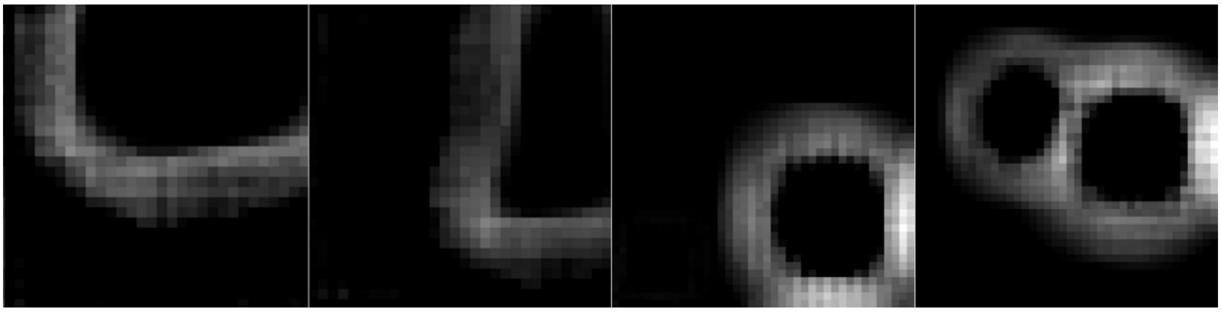


Figure 12 The average distance between each pixel and its neighbors in the feature map outputted from the last layer of the encoder.

feature maps are further processed with Gaussian blur for visualization in Figure 12. We can see that the high intensity regions are aligned with the boundary of the wounds. From here we can conclude that the cluster assumption is preserved in the segmentation of our wound images, which means that unsupervised perturbations applied to these feature maps can be used to enforce consistency through the unsupervised consistency loss function.

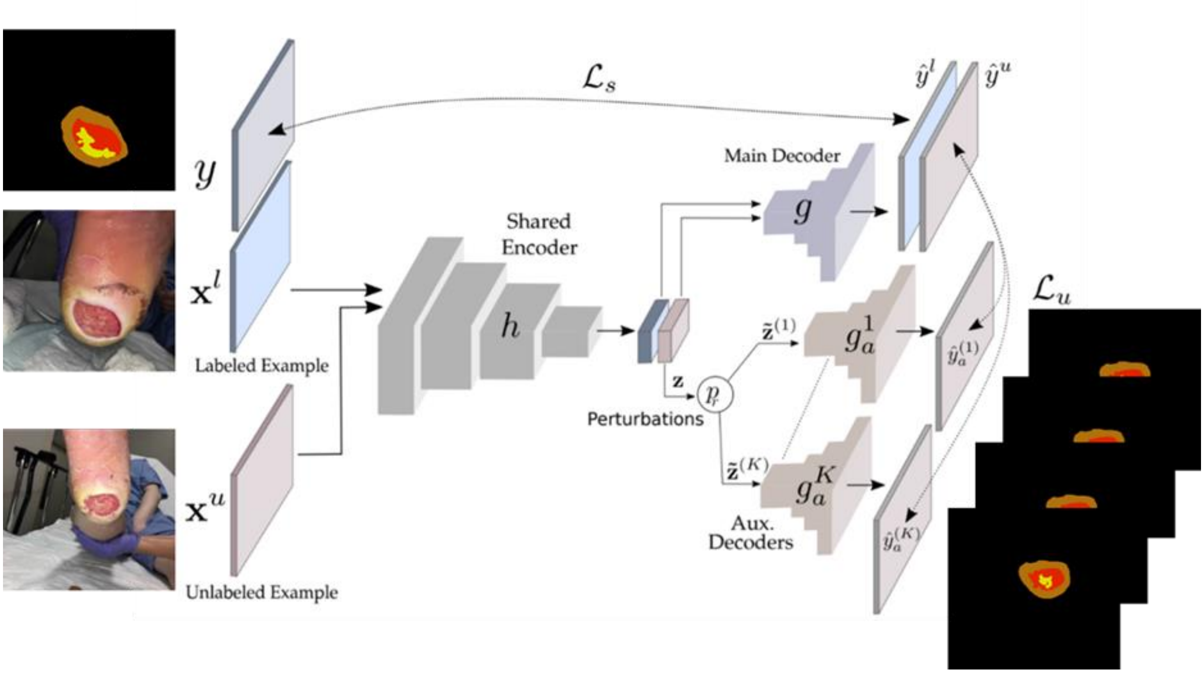


Figure 13 The proposed semi-supervised CCT model.

5.2.2 Semi-supervised Cross-Consistency Training

Our proposed network (illustrated in Figure 13) employs a widely used encoder-decoder architecture. The encoder h is shared by the labeled and unlabeled inputs. We have tested different variations of the ResNet and ResNet50 is selected as the backbone for the encoder due to its slightly better performance. Figure 14 shows the architecture of the encoder. Our architecture consists of a series of decoders, the main decoder g and a set of K auxiliary decoders, $\{g_a^k\}_{k=1}^K$. The loss function is composed of two parts: supervised loss L_S, and unsupervised loss L_U:

$$Loss = L_S + \omega L_U$$

, where the weight ω is a Gaussian function that increases from 0 to a constant λ during training. The supervised loss is computed as the Cross-Entropy loss (CE) between the ground truth annotation y and the segmentation mask $\hat{y} = f(x_i^l)$ that is predicted from the outpur of the main decoder through the network $f = h \circ g$:

$$L_S = \frac{1}{N_l} \sum_{i=1}^{N_l} H(y_i, f(x_i^l))$$

For unlabeled images, the output of the encoder z is slightly modified by K different perturbation functions P_r and perturbed into K modified versions, $z^{(1)}$ to $z^{(K)}$. Each of

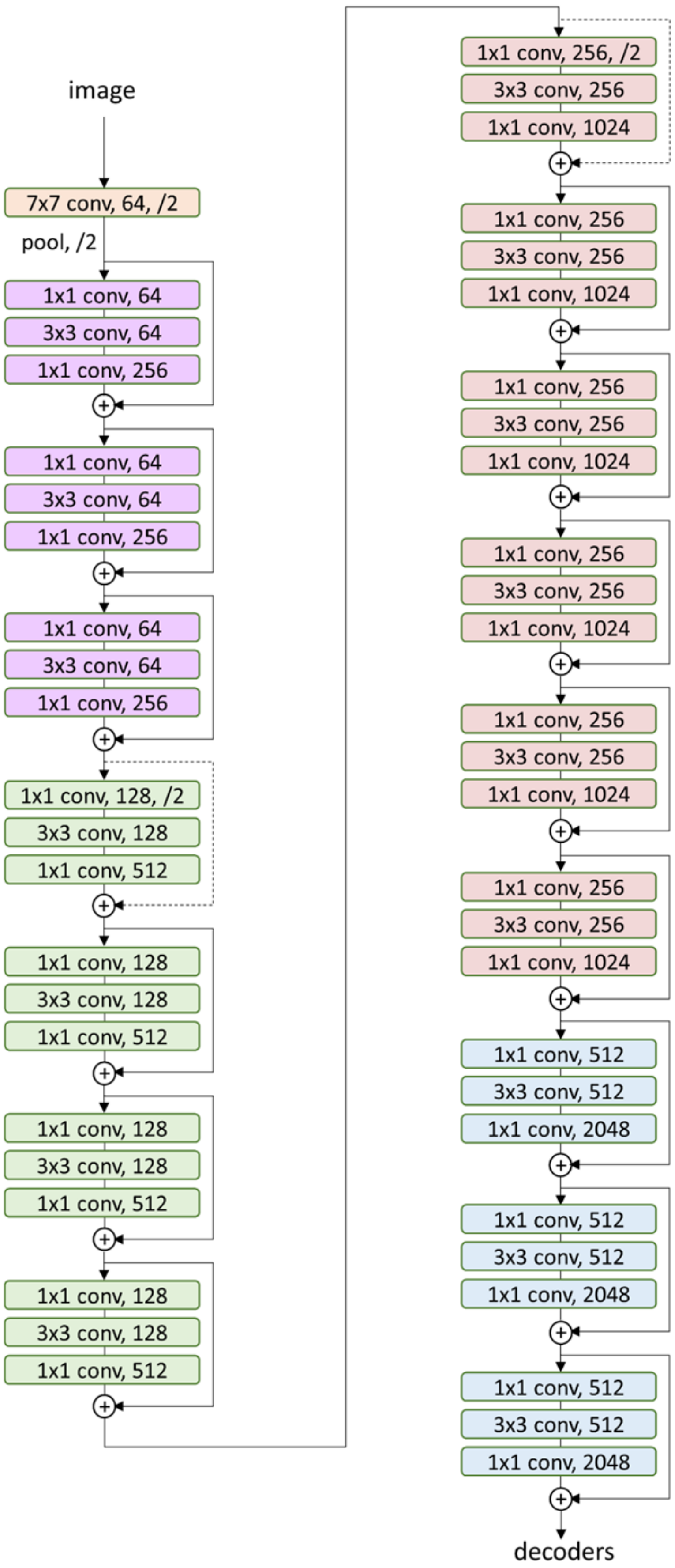


Figure 14 The encoder network of our CCT model.

these modified versions are fed into the corresponding auxiliary decoder, yielding K predictions, $\hat{y}_a^{(1)}$ to $\hat{y}_a^{(K)}$. The unsupervised loss L_U is then computed from the mean square error (MSE) between the output of the main decoder g and the auxiliary decoders g_a^{k}:

$$L_U = \frac{1}{N_u}\frac{1}{K}\sum_{i=1}^{N_u}\sum_{k=1}^{K} MSE(g(z_i),\ g_a^k(z_i))$$

The unsupervised loss is big when the distance between the outputs of the main decoder and the auxiliary decoders. In other words, the unsupervised loss measures the consistency of the model predicting the wound tissues given similar wounds. The training goal is to maximize the supervised loss the minimize the unsupervised loss. The total loss is then back-propagated to train the network including the auxiliary networks.

5.2.3 Perturbation Functions

In our CCT model, a key component is the perturbation functions applied to the feature maps outputted from the main encoder. Following the principle that the perturbation functions should be meaningful and realistic for the wound images, we propose 5 perturbation functions including random rotation, random zoom, random blurring, random normalization, and random elastic transformation. In addition to the above mentioned functions, we also adopted 7 perturbation functions suggested in

[62] including adversarial perturbations (VAT), guided cutout, random cutout, context masking, object masking, feature-based masking, and feature-based noise. The perturbation functions are categorized into four types: transformation-based, prediction-based, feature-based, and random-based perturbations.

Transformation-based perturbations

These perturbation functions map the feature map z to another feature map by performing geometric operations. To compute the unsupervised loss, we geometrically restore the output of auxiliary decoders by applying a corresponding inverse transformation to $\hat{y}_a^{(k)}$ before calculating the MSE between $g(z_i)$ and $g_a^k(z_i)$.

- **Random rotation and zoom.** We rotate the feature map z with a random degree between -90° and 90° and scale it with a random factor between 0.8 and 1.2. In clinical settings, the wound images are taken from various angles and positions. Since CNNs are not rotation and scaling invariant, we add these two perturbation functions to train our model to be consistent when predicting from a wider range of wound images. Note that we did not use projection transformation because most wound images are taken right in front of the wounds.
- **Stochastic elastic deformation.** Given the fact that skins are deformable, we apply elastic deformation [64] to the feature map z to reduce the reliance on certain shapes of common wounds and improve the robustness and consistency. We use a standard deviation of (32, 32) to limit the pixel displacement within a reasonable range. Bilinear interpolation is used to regenerate the deformed

pixels with a 64 by 64 Gaussian kernel and a scaling factor of (1.0, 1.0).

Prediction-based perturbations

These perturbation functions mask, zero out, or add noise to the feature map z based on the predictions from the main decoder or auxiliary decoders, i.e., $\hat{y}$ or $\hat{y}_a^{(k)}$.

- **Virtual adversarial perturbation (VAP).** Based on the prior work [62], the adversarial perturbation effectively alters the prediction of a given auxiliary decoder. To investigate the effectiveness of VAT in smoothing the output distribution, we adopt VAP [63] to generate a perturbed version of the feature map z.
- **Guided cutout.** We also adopt the guided cutout method as a perturbation function to limit the model's reliance on a particular feature of the wounds. In detail, a bounding box localizing the wound is predicted from $\hat{y}$. Then a random mask is zeroed out within the bounding box of the wound in the feature map z.
- **Guided masking.** CNN relies heavily on learning spatial relations to make predictions. To limit the model's reliance on spatial relations, we adopt two guided-masking perturbation functions, object masking and context masking. An object mask is predicted from $\hat{y}$ first and then the context mask masking the background area is extracted by excluding the object mask from $\hat{y}$. These two masks are applied to the feature map z to create two perturbed versions.

Feature-based perturbations

These perturbation functions add noise to or drop out some of the activations from the feature map z.

- **Feature-based masking.** The objective of this perturbation function is to drop a small portion of the most active regions in the feature map z. A threshold γ with a minimum of 0.6 and a maximum of 0.9 is uniformly sampled first. After summing over the channel dimension, z is normalized and a feature-based mask is defined as $M = \{z < \gamma\}$. The perturbed version is then generated by taking the product of z and M.
- **Feature-based noise.** To generate a feature-based noise, we multiply the feature map z by a uniformly sampled noise tensor of the same size with a minimum of -0.3 and ma a maximum of 0.3. Next, a perturbed version of z is generated by adding the noise to z itself.

Random-based perturbations

- **Gaussian blur.** We uniformly sample the variance with a minimum of 0.2 and a maximum of 1.0 for the Gaussian smoothing applied to the feature map z to create a perturbed version.
- **Random normalization.** An adjustment factor is uniformly sampled with a minimum of 0.8 and a maximum of 1.2. A perturbed version of z is generated by taking the product of z and the factor.
- **Random cutout.** A random perturbation of z is generated by spatial dropout of a random mask in z.

5.3 Results

To evaluate the CCT model and investigate its effectiveness in tissue segmentation, we conduct extensive experiments in various settings. In detail, we carry out the experiments on our CCT model in three parts, the evaluation of encoders, decoders, and the overall performance. The mean Dice coefficient described in section 4.2.1 is adopted as the evaluation matrix and all comparisons are conducted as controlled experiments where we use the same software, hardware, and training parameters.

The first question we try to answer is the choice of encoders. A prior work [62] suggests using an encoder based on ResNet-50 for the PASCAL VOC 2012 dataset. To investigate this on tissue segmentation, we tested ResNet-50, ResNet-101, and ResNet-152 on the tissue dataset. The results show that ResNet-50 outperforms other architectures by a small margin. Given that ResNet-50 is more lightweight, we use an encoder based on ResNet-50 in our CCT model.

One of the key components of our CCT model is the auxiliary decoders associated with 12 different perturbation functions categorized into four groups, i.e., transformation-based, prediction-based, feature-based, and random-based functions. Since the random-based, prediction-based, and feature-based functions are either widely proven or extensively studied in [62], We focus our investigation on the proposed transformation-based functions. Figure 15 shows the evaluation of the CCT model’s performance on granulation, fibrin, and eschar tissue trained with different auxiliary decoder settings. We use 2 VAP decoders, 6 random cutout decoders, 6 feature-based noise decoders, 6 feature-based masking decoders, 4 guided masking decoders including 2 object masking, 2 context masking decoders, 2 Gaussian blur decoders, and 2 random normalization decoders for all the settings. The difference between different settings is in the transformation-based decoders: setting 1 has 6 random rotation decoders, 6 random zoom decoders, and 6 stochastic elastic deformation decoders; setting 2 has 6 random rotation decoders, 6 random zoom decoders, and 2 stochastic elastic deformation decoders; setting 3 has no transformation-based decoders. We can see that stochastic elastic deformation improves the performance of our CCT model by comparing setting 1 and 2. For example, the segmentation accuracy of fibrin tissue is improved from 81.2% to 88.2% by adding 4 additional stochastic elastic deformation decoders. Moreover, the effectiveness of random zoom and random rotation is verified by comparing setting 2 and 3. The segmentation accuracy of fibrin and eschar tissue is improved by 3.9% and 2.8% through the addition of 12 random zoom and rotation decoders.

Overall, the performance of the semi-supervised CCT model is better than a supervised ResNet-50 segmentation network. As shown in Figure 16, our CCT model is evaluated to yield a larger mean Dice coefficient under each tissue type. Some of the best predictions from our CCT model are illustrated in figure 17 along with the corresponding input images and ground truth annotations.

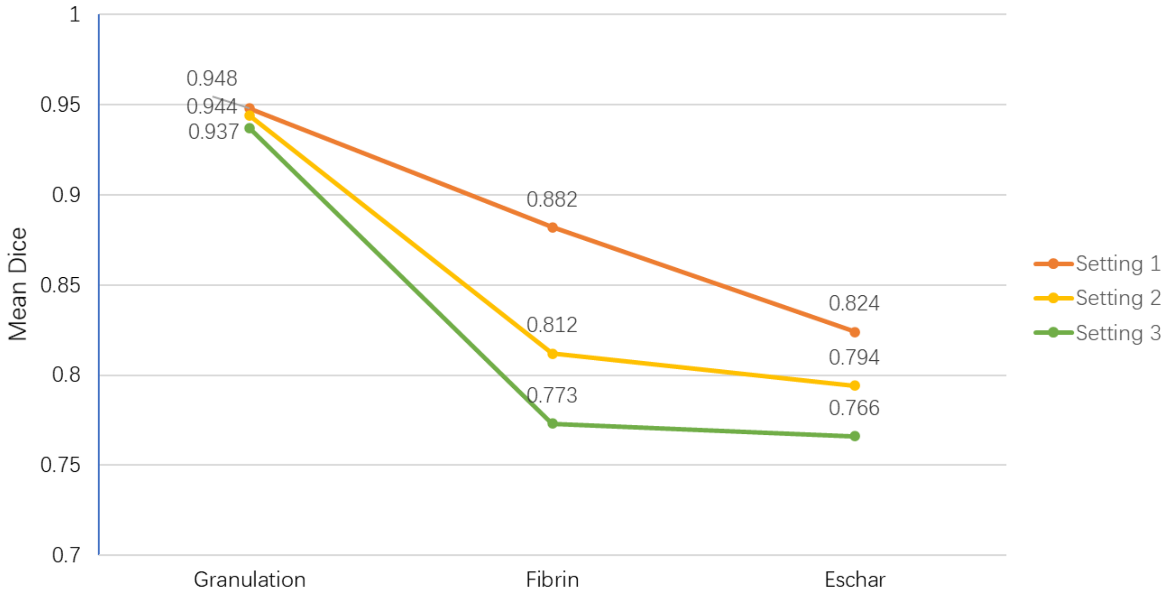


Figure 16 The Dice coefficient of our CCT model on granulation, fibrin, and eschar tissue trained with different auxiliary decoder settings

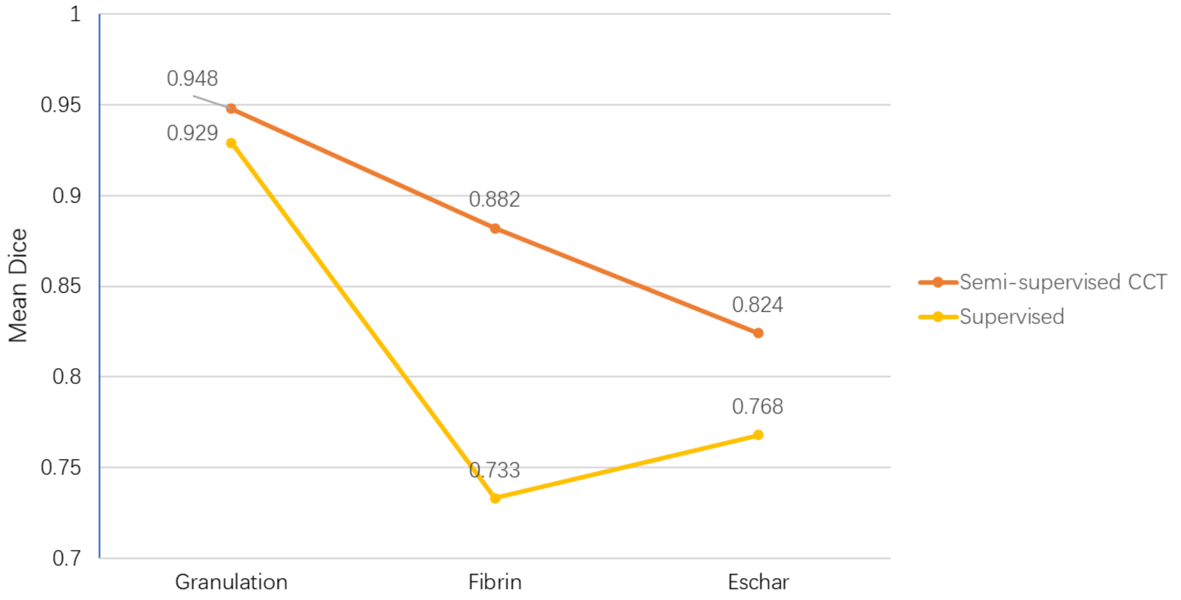
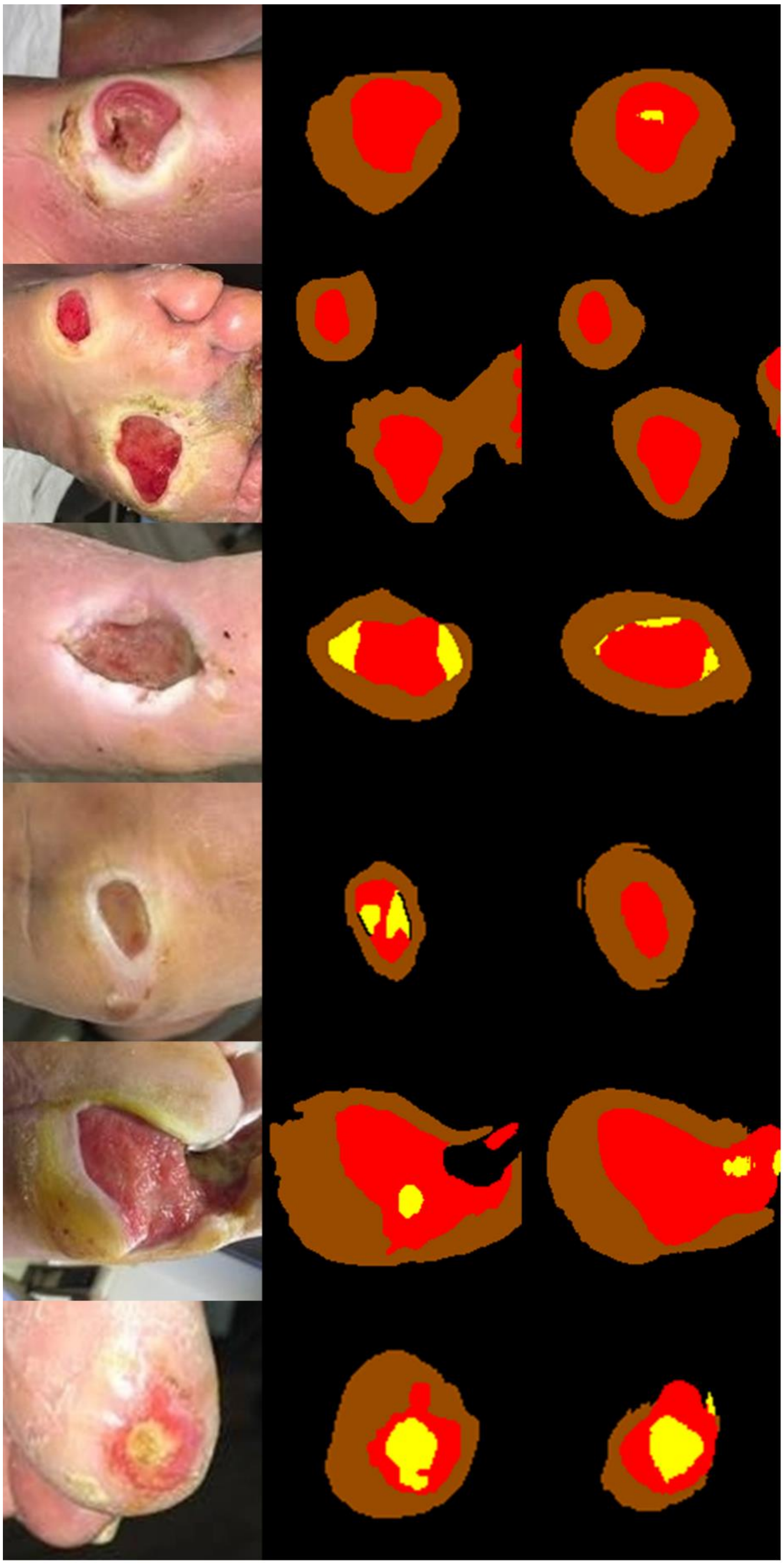


Figure 15 Comparison between the semi-supervised CCT model and the supervised model in terms of Dice coefficient for each type of wound tissue.

Figure 17 Illustration of the predictions from our CCT model. On the top are the original input images. In the middle are the ground truth annotations. On the bottom are the predictions from our model.



5.4 Discussions

In Figure 17, the results show that the segmentation of the granulation tissue and the eschar tissue are more robust than the fibrin tissue. This can also be confirmed in Figure 18 where the curves represent the training history of validation Dice accuracy for the granulation tissue (red curve), fibrin tissue (orange curve), and the eschar tissue (brown curve). Our interpretation is that this is caused by the class imbalance of these types of tissues, i.e., the average number of pixels for the fibrin tissue is significantly less than the other two types of tissues in our dataset. This means we need more samples of the fibrin tissue in our dataset to train a robust CCT model. Again, the images in our dataset are randomly collected without any filtering or interference. The class imbalance problem is from the nature of foot ulcers.

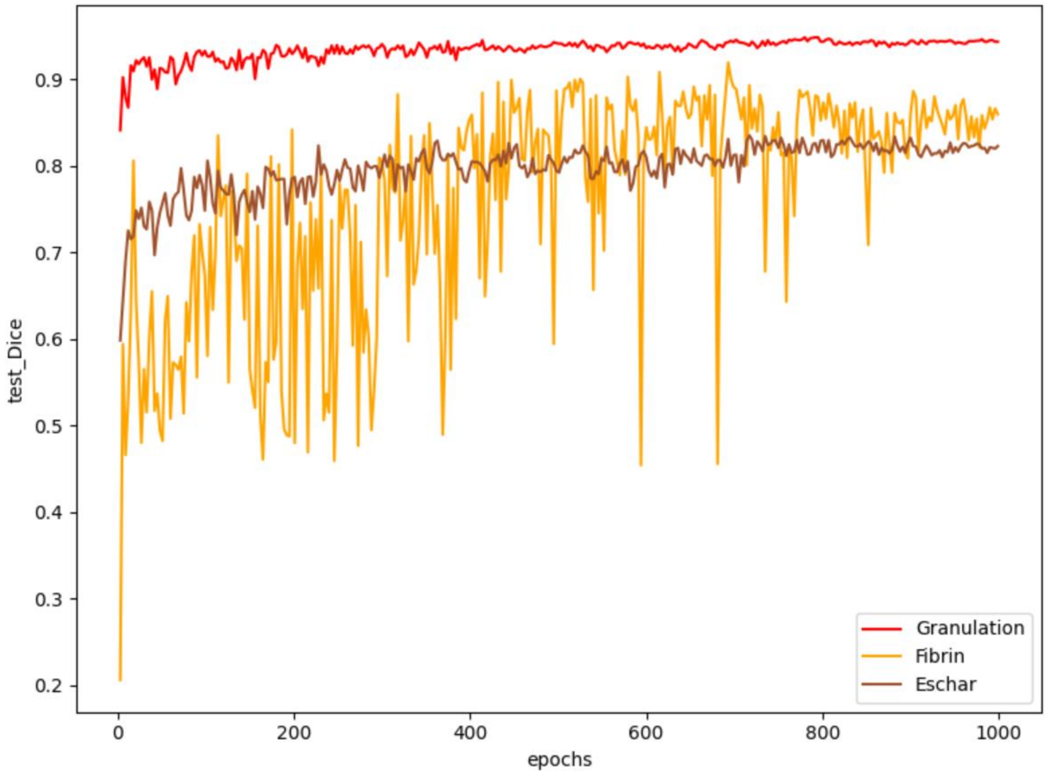


Figure 18 The training history of our CCT model. The curves mark the validation dice accuracy of the granulation tissue, the fibrin tissue, and the eschar tissue.

Chapter 6

Conclusions

We hope our wound dataset reaches a wider audience and serves academia as a benchmark to compare the performances of wound segmentation algorithms. Based on the dataset, we organized the Foot Ulcer Segmentation (FUSeg) challenge in conjunction with the 2021 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). Teams around the world submitted automated methods to predict wound segmentations on our dataset. Each participant was asked to submit a docker container that contains their algorithm and the prediction code. We have re-produced the results to be evaluated and ranked on our GPU server. Thus, one of our future directions is to further evaluate and analyze the top-3 submitted algorithms, namely the ensemble network with U-Net and LinkNet [52], the HardNet-MSEG [53], and the double encoder-decoder network [54].

The major problem of tissue segmentation is that limited labeled data is available since labeling tissues is more time-consuming and requires a higher level of expertise. To solve the problem, we investigated the CCT model with various numbers of auxiliary decoders. Besides labeling more images, a future direction is to explore the effectiveness of more perturbation functions for the auxiliary decoders. Another future work is to conduct more experiments on more auxiliary decoder settings to investigate the effectiveness of each decoder.

We attempted to solve three problems using deep learning: 1) the automated segmentation of chronic foot ulcers in a dataset we built on our own. 2) the automated segmentation of wound tissues 3) The automated segmentation of IVDs from 3D MRI scans. For evaluating the performance, we conducted comprehensive experiments and analyses on SegNet, VGG16, 2D U-Net, Mask-RCNN, 3D U-Net, our model based on modified 3D U-Net, a model based on MobileNetV2 and CCL, and a model based on cross-consistency training. In the comparison of various neural networks, our methods have demonstrated their effectiveness in the field of medical image segmentation due to their fully convolutional architectures. We also demonstrated the robustness of our models by testing them on publicly available datasets where our model still achieves the highest Dice score. In the future, we plan to improve our work by extracting the shape features separately from the pixel-wise convolution in the deep learning model. Also, we will include more data in the dataset to improve the robustness and prediction accuracy of our method.

References

[1] Song, Bo, and Ahmet Sacan. “Automated wound identification system based on image segmentation and artificial neural networks,” In 2012 IEEE International Conference on Bioinformatics and Biomedicine. 1-4 (2012).

[2] A. I. Lopez, B. Glocker, “Complementary classification forests with graph-cut refinement for IVD localization and segmentation,” in Proc. the 3rd MICCAI Workshop & Challenge on Computational Methods and Clinical Applications for Spine Imaging (2015).

[3] Fauzi, Mohammad Faizal Ahmad et al. “Computerized segmentation and measurement of chronic wound images,” Computers in biology and medicine. 60, 74-85 (2015).

[4] D. Forsberg, "Atlas-based registration for accurate segmentation of thoracic and lumbar vertebrae in CT data," in Recent Advances in Computational Methods and Clinical Applications for Spine Imaging, pp. 49-59, Springer, Cham (2015).

[5] C. Chen, D. Belavy, W. Yu, C. Chu, G. Armbrecht, M. Bansmann, D. Felsenberg, and G. Zheng, "Localization and segmentation of 3D intervertebral discs in MR images by data driven estimation," IEEE transactions on medical imaging, vol. 34, no. 8, pp. 1719-1729 (2015).

[6] Hettiarachchi, N. D. J., R. B. H. Mahindaratne, G. D. C. Mendis, H. T. Nanayakkara, and Nuwan D. Nanayakkara. “Mobile based wound measurement,” In 2013 IEEE Point-of-Care Healthcare Technologies (PHT). 298-301 (2013).

[7] Z. Wang, X. Zhen, K.Y. Tay, S. Osman, W. Romano, and S. Li. "Regression segmentation for M3 spinal images," IEEE transactions on medical imaging, vol. 34, no. 8, 1640-1648 (2015).

[8] Hani, Ahmad Fadzil M., Leena Arshad, Aamir Saeed Malik, Adawiyah Jamil, and Felix Yap Boon Bin. “Haemoglobin distribution in ulcers for healing assessment,” In 2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012). 1, 362-367 (2012).

[9] Wantanajittikul, Kittichai, Sansanee Auephanwiriyakul, Nipon Theera-Umpon, and Taweethong Koanantakool. “Automatic segmentation and degree identification in burn colour images,” In The 4th 2011 Biomedical Engineering International Conference. 169-173 (2012).

[10] H. Hutt, R. Everson, and J. Meakin, "3d intervertebral disc segmentation from MRI using supervoxel-based crfs," in International Workshop on Computational Methods and Clinical Applications for Spine Imaging, pp. 125-129, Springer, Cham (2015).

[11] M. Urschler, K. Hammernik, T. Ebner, and D. Štern, "Automatic intervertebral disc localization and segmentation in 3d mr images based on regression forests and active contours," In International Workshop on Computational Methods and Clinical Applications for Spine Imaging, pp. 130-140, Springer, Cham (2015).

[12] R. Korez, B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec. "Deformable model-based segmentation of intervertebral discs from MR spine images by using the

SSC descriptor," in International Workshop on Computational Methods and Clinical Applications for Spine Imaging, pp. 117-124, Springer, Cham (2015).

[13] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. “Imagenet classification with deep convolutional neural networks,” In Advances in neural information processing systems. 1097-1105 (2012).

[14] Russakovsky, Olga et al. “Imagenet large scale visual recognition challenge,” International journal of computer vision. 115, no. 3, 211-252 (2015).

[15] Garcia-Garcia, Alberto, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. “A review on deep learning techniques applied to semantic segmentation,” arXiv preprint arXiv:1704.06857 (2017).

[16] LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition,” Proceedings of the IEEE. 86, no. 11, 2278-2324 (1998).

[17] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation,” In Proceedings of the IEEE conference on computer vision and pattern recognition. 3431-3440 (2015).

[18] Wang, Changhan et al. “A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks,” In 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). 2415-2418 (2015).

[19] Goyal, Manu, Moi Hoon Yap, Neil D. Reeves, Satyan Rajbhandari, and Jennifer

Spragg. “Fully convolutional networks for diabetic foot ulcer segmentation,” In 2017 IEEE international conference on systems, man, and cybernetics (SMC). 618-623 (2017).

[20] Liu, Xiaohui et al. “A framework of wound segmentation based on deep convolutional networks,” In 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 1-7 (2017).

[21] H. Chen, Q. Dou, X. Wang, J. Qin, J. CY Cheng, and P. A. Heng, "3D fully convolutional networks for intervertebral disc localization and segmentation," in International Conference on Medical Imaging and Virtual Reality, pp. 375-382, Springer, Cham (2016).

[22] X. Li, Q. Dou, H. Chen, C. Fu, X. Qi, D. L. Belavý, G. Armbrecht, D. Felsenberg, G. Zheng, and P. A. Heng, "3D multi-scale FCN with random modality voxel dropout learning for Intervertebral Disc Localization and Segmentation from Multi-modality MR Images," Medical image analysis, vol. 45, pp. 41-54 (2018).

[23] H. Chen, Q. Dou, X. Wang, P. A. Heng, “Deepseg: Deep segmentation network for intervertebral disc localization and segmentation,” in Proc. 3rd MICCAI Workshop & Challenge on Computational Methods and Clinical Applications for Spine Imaging (2015).

[24] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in International

Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 424-432. Springer, Cham (2016).

[25] G. Zheng, C. Chu, D. L. Belavý, B. Ibragimov, R. Korez, T. Vrtovec, and H. Hutt et al, "Evaluation and comparison of 3D intervertebral disc localization and segmentation methods for 3D T2 MR data: A grand challenge," Medical image analysis, vol. 35, pp. 327-344 (2017).

[26] S. Kim, W. Bae, K. Masuda, C. Chung, and D. Hwang, "Fine-grain segmentation of the intervertebral discs from MR spine images using deep convolutional neural networks: BSU-Net," Applied Sciences, vol. 8, no. 9, pp. 1656 (2018).

[27] J. Dolz, C. Desrosiers, and I. B. Ayed, "IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal Unet," arXiv preprint arXiv:1811.08305 (2018).

[28] J. T. Lu, S. Pedemonte, B. Bizzo, S. Doyle, K. P. Andriole, M. H. Michalski, R. G. Gonzalez, and S. R. Pomerantz. "DeepSPINE: Automated Lumbar Vertebral Segmentation, Disc-level Designation, and Spinal Stenosis Grading Using Deep Learning," arXiv preprint arXiv:1807.10215 (2018).

[29] Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. “Mobilenetv2: Inverted residuals and linear bottlenecks,” In Proceedings of the IEEE conference on computer vision and pattern recognition. 4510-4520 (2018).

[30] C. Chen, D. Belavy, and G. Zheng, "3D intervertebral disc localization and

segmentation from MR images by data-driven regression and classification," in International Workshop on Machine Learning in Medical Imaging, pp. 50-58 (2014).

[31] Redmon, Joseph, and Ali Farhadi. “Yolov3: An incremental improvement,” Preprint at arXiv:1804.02767 (2018).

[32] Tzutalin. LabelImg. Git code https://github.com/tzutalin/labelImg (2015).

[33] Chollet, François. Xception: “Deep learning with depthwise separable convolutions,” In Proceedings of the IEEE conference on computer vision and pattern recognition. 1251-1258 (2017).

[34] Nair, Vinod, and Geoffrey E. Hinton. “Rectified linear units improve restricted boltzmann machines,” In Proceedings of the 27th international conference on machine learning (ICML-10). 807-814 (2010).

[35] Krizhevsky, Alex, and Geoff Hinton. “Convolutional deep belief networks on cifar-10,” Unpublished manuscript. 40, 1-9 (2010).

[36] Everingham et al. “The pascal visual object classes challenge: A retrospective,” International journal of computer vision. 111, no. 1, 98-136 (2015).

[37] Pearce, David J. “An improved algorithm for finding the strongly connected components of a directed graph,” Victoria University, Wellington, NZ, Tech. Rep (2005).

[38] S. Park, and N. Kwak, "Analysis on the dropout effect in convolutional neural networks," in Asian Conference on Computer Vision, pp. 189-204. Springer,

Cham (2016).

[39] Li, Fangzhao, Changjian Wang, Xiaohui Liu, Yuxing Peng, and Shiyao Jin. “A composite model of wound segmentation based on traditional methods and deep neural networks,” Computational intelligence and neuroscience. (2018).

[40] Zou, Kelly H. et al, “Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports,” Academic radiology. 11, no. 2, 178-189 (2004).

[41] F. Chollet et al, Keras, chollet2015keras, https://keras.io.

[42] S. S. Girija, Tensorflow: Large-scale machine learning on heterogeneous distributed systems. (2016).

[43] Kingma, Diederik P., and Jimmy Ba. “Adam: A method for stochastic optimization,” Preprint at arXiv:1412.6980 (2014).

[44] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” In Proceedings of the IEEE international conference on computer vision. 1026-1034 (2015).

[45] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation,” In International Conference on Medical image computing and computer-assisted intervention. 234-241 (2015).

[46] Thomas, Stephen. Stock Pictures of Wounds. Medetec Wound Database http://www.medetec.co.uk/files/medetec-image-databases.html (2020).

[47] Q. Fang, and D. A. Boas. "Tetrahedral mesh generation from volumetric binary and grayscale images," in Biomedical Imaging: From Nano to Macro, ISBI'09, IEEE International Symposium, pp. 1142-1145, 2009.

[48] Frykberg, Robert G., and Jaminelli Banks. "Challenges in the treatment of chronic wounds. Advances in wound care. 4, no. 9, 560-582 (2015).

[49] Sen, Chandan K. "Human wounds and its burden: an updated compendium of estimates," Advances in Wound Care. 8, 39-48 (2019).

[50] Branski, Ludwik K., Gerd G. Gauglitz, David N. Herndon, and Marc G. Jeschke. "A review of gene and stem cell therapy in cutaneous wound healing," Burns. 35, no. 2, 171-180 (2009).

[51] Intervertebral Disc Disease. National Institutes of Health https://ghr.nlm.nih.gov/condition/intervertebral-disc-disease.

[52] Mahbod, Amirreza, Rupert Ecker, and Isabella Ellinger. "Automatic Foot Ulcer Segmentation Using an Ensemble of Convolutional Neural Networks," arXiv preprint arXiv:2109.01408 (2021).

[53] Huang, Chien-Hsiang, Hung-Yu Wu, and Youn-Long Lin. "Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps," arXiv preprint arXiv:2101.07172 (2021).

[54] Galdran, Adrian, Gustavo Carneiro, and Miguel A. González Ballester. "Double Encoder-Decoder Networks for Gastrointestinal Polyp Segmentation," International Conference on Pattern Recognition. Springer, Cham (2021).

[55] He, Kaiming, et al. "Mask r-cnn," Proceedings of the IEEE international conference on computer vision (2017).

[56] Abdulla, Waleed. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. Git code https://github.com/matterport/Mask_RCNN (2017)

[57] Zhou Y, He X, Huang L, et al. “Collaborative learning of semi-supervised segmentation and classification for medical images,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2079-2088 (2019).

[58] Bortsova, Gerda, et al. "Semi-supervised medical image segmentation via learning consistency under transformations." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, (2019).

[59] Peng, Jizong, Marco Pedersoli, and Christian Desrosiers. "Mutual information deep regularization for semi-supervised segmentation." Medical Imaging with Deep Learning. PMLR, (2020).

[60] Chen, Shuai, et al. "Multi-task attention-based semi-supervised learning for medical image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, (2019).

[61] Sedai, Suman, et al. "Uncertainty guided semi-supervised segmentation of retinal layers in OCT images." International Conference on Medical Image

Computing and Computer-Assisted Intervention. Springer, Cham, (2019).

[62] Ouali, Yassine, Céline Hudelot, and Myriam Tami. "Semi-supervised semantic segmentation with cross-consistency training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020).

[63] Miyato, Takeru, et al. "Virtual adversarial training: a regularization method for supervised and semi-supervised learning." IEEE transactions on pattern analysis and machine intelligence 41.8, 1979-1993 (2018).

[64] Simard, Patrice Y., David Steinkraus, and John C. Platt. "Best practices for convolutional neural networks applied to visual document analysis." Icdar. Vol. 3. No. 2003 (2003).

[65] Athiwaratkun, Ben, et al. "There are many consistent explanations of unlabeled data: Why you should average." arXiv preprint arXiv:1806.05594 (2018).

[66] French, Geoff, et al. "Semi-supervised semantic segmentation needs strong, varied perturbations." arXiv preprint arXiv:1906.01916 (2019).

[67] Laine, Samuli, and Timo Aila. "Temporal ensembling for semi-supervised learning." arXiv preprint arXiv:1610.02242 (2016).

[68] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE transactions on pattern analysis and machine intelligence 39.12, 2481-2495 (2017).

[69] Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully

convolutional neural networks for volumetric medical image segmentation." 2016 fourth international conference on 3D vision (3DV). IEEE, (2016).

[70] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 4700-4708 (2017).

[71] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 770-778 (2016).

[72] Amirul Islam, Md, et al. "Gated feedback refinement network for dense image labeling." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017).

[73] Fu, Jun, et al. "Stacked deconvolutional network for semantic segmentation." IEEE Transactions on Image Processing (2019).

[74] Lin, Guosheng, et al. "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017).

[75] Peng, Chao, et al. "Large kernel matters--improve semantic segmentation by global convolutional network." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017).

[76] Pohlen, Tobias, et al. "Full-resolution residual networks for semantic segmentation in street scenes." Proceedings of the IEEE conference on

computer vision and pattern recognition. (2017).

[77] Wojna, Zbigniew, et al. "The devil is in the decoder." British Machine Vision Conference 2017, BMVC 2017. BMVA Press, (2017).

[78] Zhang, Zhenli, et al. "Exfuse: Enhancing feature fusion for semantic segmentation." Proceedings of the European conference on computer vision. (2018).

[79] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European conference on computer vision. (2018).

[80] Cordts, Marius, et al. "The cityscapes dataset." CVPR Workshop on the Future of Datasets in Vision. Vol. 2. sn, (2015).

[81] Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding." Proceedings of the IEEE conference on computer vision and pattern recognition. (2016).

[82] Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017).

[83] He, Kaiming, et al. "Spatial pyramid pooling in deep convolutional networks for visual recognition." IEEE transactions on pattern analysis and machine intelligence 37.9 (2015): 1904-1916.

[84] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).

[85] Han, Song, et al. "Learning both weights and connections for efficient neural network." Advances in neural information processing systems 28 (2015).

[86] LeCun, Yann, John Denker, and Sara Solla. "Optimal brain damage." Advances in neural information processing systems 2 (1989).

[87] Hanson, Stephen, and Lorien Pratt. "Comparing biases for minimal network construction with back-propagation." Advances in neural information processing systems 1 (1988).

[88] Hassibi, Babak, and David Stork. "Second order derivatives for network pruning: Optimal brain surgeon." Advances in neural information processing systems 5 (1992).

[89] He, Wei, et al. "Cap: Context-aware pruning for semantic segmentation." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021.

[90] Zhou, Zongwei, et al. "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation." IEEE transactions on medical imaging 39.6, 1856-1867 (2019).

[91] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. (2018).

[92] Chen, Liang-Chieh, et al. "Attention to scale: Scale-aware semantic image segmentation." Proceedings of the IEEE conference on computer vision and

pattern recognition. (2016).

[93] Li, Hanchao, et al. "Pyramid attention network for semantic segmentation." arXiv preprint arXiv:1805.10180 (2018).

[94] Fu, Jun, et al. "Dual attention network for scene segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019).

[95] Bachman, Philip, Ouais Alsharif, and Doina Precup. "Learning with pseudo-ensembles." Advances in neural information processing systems 27 (2014).

[96] Rasmus, Antti, et al. "Semi-supervised learning with ladder networks." Advances in neural information processing systems 28 (2015).

[97] Laine, Samuli, and Timo Aila. "Temporal ensembling for semi-supervised learning." arXiv preprint arXiv:1610.02242 (2016).

[98] Tarvainen, Antti, and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results." Advances in neural information processing systems 30 (2017).

[99] Xie, Qizhe, et al. "Unsupervised data augmentation for consistency training." Advances in Neural Information Processing Systems 33 (2020).

[100] Olsson, Viktor, et al. "Classmix: Segmentation-based data augmentation for semi-supervised learning." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (2021).

[101] Yun, Sangdoo, et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features." Proceedings of the IEEE/CVF international

conference on computer vision. (2019).

[102] Yarowsky, David. "Unsupervised word sense disambiguation rivaling supervised methods." 33rd annual meeting of the association for computational linguistics. (1995).

[103] Riloff, Ellen. "Automatically generating extraction patterns from untagged text." Proceedings of the national conference on artificial intelligence. (1996).

[104] Riloff, Ellen, and Janyce Wiebe. "Learning extraction patterns for subjective expressions." Proceedings of the 2003 conference on Empirical methods in natural language processing. (2003).

[105] Babakhin, Yauhen, Artsiom Sanakoyeu, and Hirotoshi Kitamura. "Semi-supervised segmentation of salt bodies in seismic images using an ensemble of convolutional neural networks." German Conference on Pattern Recognition. Springer, Cham, (2019).

[106] Dong, Xue, et al. "Automatic multiorgan segmentation in thorax CT images using U‐net‐GAN." Medical physics 46.5 (2019): 2157-2168.

[107] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017).

[108] Cirillo, Marco Domenico, David Abramian, and Anders Eklund. "Vox2Vox: 3D-GAN for brain tumour segmentation." International MICCAI Brainlesion Workshop. Springer, Cham, (2020).

[109] Menze, Bjoern H., et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." IEEE transactions on medical imaging 34.10 1993-2024 (2014).

[110] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

[111] X. Li, Q. Dou, H. Chen, C. Fu, X. Qi, D. L. Belavý, G. Armbrecht, D. Felsenberg, G. Zheng, and P. A. Heng, "3D multi-scale FCN with random modality voxel dropout learning for Intervertebral Disc Localization and Segmentation from Multi-modality MR Images," Medical image analysis, 45: 41-54 (2018).

[112] Zhang, Wenlu, et al. "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation." NeuroImage 108: 214-224 (2015).