Automatic Longitudinal Assessment of Tumor Responses

by

Tzu-Ming Harry Hsu

B.S., National Taiwan University (2016)
B.S.E., National Taiwan University (2016)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author .....
Department of Electrical Engineering and Computer Science
May 15, 2020

Certified by .....
Peter Szolovits
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Automatic Longitudinal Assessment of Tumor Responses

by

Tzu-Ming Harry Hsu

Submitted to the Department of Electrical Engineering and Computer Science
on May 15, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

Tumor response to therapy is assessed by measuring the changes in lesion sizes across consecutive imaging studies. This process is oftentimes inconsistent and time-consuming even if official guidelines like RECIST exist. In this work, I develop a pipeline, integrating 3D CNNs and conventional optimization algorithms to determine changes in tumor sizes in consecutive MRI exams for patients with neuroendocrine tumors.

The CNN is evaluated against a publicly available dataset – LiTS – that contains 201 abdominal CT scans, in terms of key design choices, including image augmentation, network complexity, and loss functions. The best design is used to assemble the system, which is finally trained and evaluated on a private MRI dataset with 145 total studies, each labeled by two board-certified radiologists. Metrics include Dice score, sensitivity, and specificity of lesion detections on per-lesion, per-liver-segment, and per-study bases. Concordance with the radiologists on the endpoint evaluation according to RECIST is also evaluated.

The system is able to agree with radiologists in 91% of the cases, having a sensitivity of 0.85 (95% CI: 0.77, 0.93) and specificity of 0.93 (95% CI: 0.87, 0.96) to classify liver segments as diseased or healthy. The experimental evidence suggests a potential for the automatic system to perform these routine tasks in conjunction with clinicians. Moreover, the volumetric tumor burden change assessments showcased in the work demonstrates extended capabilities of the system that are shown to correlate with clinical endpoints better but are not feasible for radiologists in clinics.

Thesis Supervisor: Peter Szolovits
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I always feel like there is a genie who magically pops out whenever I have a wish. The wish is oftentimes demanding, and by no means these wishes should come true – but they eventually do most of the time. With time, I gradually realize that this genie does not come out of nowhere – it is a manifestation of love and support from everyone around me: from my significant other, my family, my friends, my colleagues, and my supervisor.

Graduate school has been tough for me as I spent some decent time looking for directions, for both life and research, upon entering MIT. I would not have discovered my interest in medical imaging and joined Medical Decision Making Group were it not for the recognition given by professor Peter Szolovits. Pete guides me around the ridges and trenches that would have caused me trouble. I would love to write up another thesis about how Pete magically finds ways to deal with our requests, questions, and problems, but I am still collecting more data.

Life has been through unimaginable ups and downs during these few years for me in an emotional way. There were moments when I felt very supported but others when it threw devastation at me, and yet I do sincerely thank the existence of both. I thank everyone involved and hope they can continue on successful and fulfilled lives.

No matter how things have changed, my dearest friends have always been on my side for my support: Lily, Wei-Hung, Schrasing, and many others. Thank you for your presence in my life. I also would like to attribute my gain in knowledge to my labmates Geeticka, Elena, Willie, and Matthew. The philosophical and sometimes nonsensical discussions in the office certainly open up my mind for a world beyond me.

Alex and Ronilda, thank you for your diligence in mentoring me as medical experts and getting along with me as friends. I hope one day I will be able to help someone else as much as you have helped me.

Mom and dad, I know I do not often express my love, but it is really both of you who not only gave birth to me and encouraged me to step on this path that people

deem challenging. Thank you for your trust in me and all of your support.

Finally, I would like to thank Chelsea, who is willing to spend her time with me, laughing or crying. It is not easy to pour all one's effort into an uncertain plan for the future, but we will work together to get it on track.

I am more than lucky to be in the place I am in, and there are too many people to thank in the process – please continue to be the genie, and I promise I will be a good boy.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Contribution

Longitudinal assessment of tumor burden is a clinically important task to determine whether the tumors have positive responses to the therapy. Both the tumor size change and the time to develop disease progression are important endpoints in the assessment from a cancer clinical trial perspective. As treatment response has direct ties with mortality [4], it is becoming more and more imperative to establish an objective evaluation of tumor responses that is universally accepted and comparable across different studies.

Hence, in 1981 the World Health Organization (WHO) published the first tumor response criteria [5], introducing a concept that utilizes the sum of products of bi-dimensional measurements of lesions as a proxy for tumor burden, and that the responses can thus be determined by the change of this surrogate. However, with time, research groups had been encountering the need to re-interpret the original specifications as the documents were not concise enough [6].

In response to the problem, RECIST (Response Evaluation Criteria in Solid Tumours) guideline [7] was proposed in 2000, aiming at both maximizing reproducibility and clinical efficiency. RECIST features definitions that involve uni-dimensional measurements and a limit on both the size and number of lesions that are taken into account. Subsequently, the criteria were quickly adopted by academic institutions

and industrial trials, and further revised into RECIST 1.1 [8].

All criteria mentioned above, in order to allow a reasonable trade-off between evaluation effort and time, prescribe measurements of a limited number of lesions in the 2-dimensional plane. Yet, there is evidence [9, 10, 11] suggesting that volumetric assessments demonstrate superior performance in quantifying tumor burden. On top of the matter, studies show [12] that following RECIST guideline does not guarantee absolute agreement between evaluators as there are certain nuances and subtleties of RECIST.

With the premise that manual annotation in the 3-dimensional space is impractical to be employed in day-to-day clinics, and that existing guidelines still are open to some level of subjective evaluation, we ask the question:

Are automatic methods able to aid, if not replace, the diagnostic workflow of assessing tumor burden changes in time from 3-dimensional medical images?

There are multiple challenges ahead with the attempt. Specifically, inhomogeneity in operating equipment, different practices from operators and dissimilar processing software suites all lead to non-trivial dynamics in the images. To put this problem into the context of liver lesion detection, the intensity disparity between regular liver tissue and lesions is highly variable [13] from lesion to lesion, so there is not a trivial threshold we can apply directly.

Luckily, the aforementioned question has a positive outlook, and we all know that is owing to the rapid growth of *deep neural networks*. Recent insights from the computer vision domain of object detection in natural images have brought new ideas and improvements to deep learning segmentation models for medical imaging [14]. In the task *segmentation*, we wish to provide a classification for each and every image element so that the resulting output has the same size as the input image. In 2-dimensional (2D) images these elements are called *pixels* and in 3-dimensional (3D) images like computed tomography (CT) and magnetic resonance imaging (MRI) they are called *voxels*. One classic example is the segmentation of 2D cell microscopy

images in which we wish to distinguish the cells from the background, thus creating *two classes* for the classification problem. It is this problem formulation and the rise of 3D neural networks that are gradually enabling an automatic pipeline to first extract lesion detection results from imaging studies and then to compare the detection across consecutive imaging studies.

While it is fantastic for computer vision techniques to aim to improve the precision of lesion detection on computation-oriented metrics [15], clinical utility arises only from connecting the detection outcomes to diagnostic conclusions about the patient. Following these lines, I concentrate this research on medical imaging, specifically for 3D imagery, CT and MRI included. This work proposes and evaluates a system that detects liver lesions in neuroendocrine cancer patients, and provides diagnostic descriptions across consecutive studies in time. To be concise, the contributions of this work are that:

- I present a dataset of abdominal MRI scans with liver lesion annotations[1].
- I demonstrate a workflow combining 3D convolutional neural networks (CNNs) and conventional algorithms to speed up longitudinal lesion comparison studies.
- I offer an extensive ablation study on selected techniques used for 3D CNNs.
- I compare the performance of the proposed methods to expert interpretations and showcase new evaluation targets for automatic methods in disease progression descriptions.

1.2 Literature Review

1.2.1 Conventional Liver Lesion Detection

The medical imaging community has been taking different approaches to accelerate the process of liver lesion identification, be it semi-automatic or automatic. The main roadblock in sight is that it significantly differs from object detection in natural

[1] I did not prepare this private dataset. See the data chapter for more information.

images, in that the size of the lesions can easily take up a sizable portion of the organ, but it can also be too tiny to be noticed/measured, which is without a doubt not a scale of interest for natural object detection.

The very first works in this line are based on simple thresholding [16, 17, 18] of image values. The choice of the threshold could be vulnerable to image noises, and hence multiple refinements are proposed, each using variance maximization [19], cross-entropy minimization [20], histogram equalization [21], and Isodata thresholding [22].

Due to the relatively low accuracy of the thresholding methods, some researchers [23, 24, 25] shift their focus to region-growing methods which also allows injection of knowledge-based priors [26]. This family of methods are often used together with an initialization based on threshold bootstraping [27, 20].

As previously explained in Section 1.1, the variability in the images are oftentimes too great for unsupervised, or weakly-supervised method to generalize. Machine learning-based methods start to boom based off this observation, and the first few works are using conventional machine learning techniques such as Bayesian classification [28], k-means clustering [29], fuzzy c-means clustering [30], hidden Markov model [31], support vector machine (SVM) [32, 33, 34], or adaptive boosting (AdaBoost) [35, 36].

With the development of neural networks starting in 2012 [37], recent efforts have poured into neural network for image segmentation and lesion detection.

1.2.2 Deep Learning for Segmentation

The first few explorations focus on architectural advances such as U-Net [1] which enables the model to learn spatial features on multiple spatial resolutions, while allowing information to skip intermediate levels for faster training over a simple encoder-decoder architecture. Fully-convolutional networks (FCN) [38] came out around the same period and it allows inference to be made on variable-sized images as opposed to only taking fixed-sized image patches. A classical benchmark in segmentation is thus combining U-Net and FCN, the two earliest works in this field.

The overall network architecture as been through several generations of improve-

ments while the most of them are eager to deal with the multi-scale problem. Some [39, 40] extracts deep features for images at different *pyramid scales*; Pyramid Scene Parsing Network (PSPNet) [41] pools features into deeper levels of features with different kernel sizes; and Feature Pyramid Networks (FPN) [42] effectively combines both pursuits into a more performant model.

In the meantime, the *backbone network*, which is the network accounting for the down-sampling path in a U-Net or similar architectures, received major improvement owing to ResNet [2] that dramatically accelerates training furthermore. Following ResNet, there has been a few more refinements building on top of it: ResNeXt [43], Inception [44], and Inception-ResNet [45].

All of the frameworks mentioned above have been mainly applied on 2D images due to their deeply rooted origin in natural images. With that said, a big game changer found its way into the medical imaging society. 3D CNNs [46] were developed with applications to videos, which is essentially a series of stacked 2D images, in mind, but were adapted to deal with 3D medical images such as CT and MRI. DeepMedic [47] achieved state-of-the-art brain lesion segmentation results with its 3D architecture in the *Ischemic Stroke Lesion Segmentation (ISLES) Challenge (2015*, and then this model was extended [48] to win another *Brain Tumor Image Segmentation (BRATS) Challenge 2016*. Many works that followed were all based on a 3D U-Net framework and received countless prizes in similar brain tumor challenges [49, 50, 51, 52].

A similar storyline happened with liver lesion segmentation as well. The first work [53] in this line explored 3D segmentation for liver, heart, and blood vessel. The appearance of *Liver Tumor Segmentation (LiTS) Challenge* [15] sparked interest in advancing techniques for liver lesion detection [54, 55, 56, 57, 58, 59, 60]. The best entry in the 2017 LiTS competition, H-DenseUNet [61], combines a 2D-learned feature for faster learning in the 3D model.

1.2.3 Liver Lesion Segmentation Datasets

Compared to the brain and other organs, not too much effort has been put into the making of publicly available liver lesion detection datasets, and has been a root cause

of the slower growth in research works [13]. The major reason for this disparity is that brain imaging suffers less from both motion artifact and examination length, and hence the resulting quality is typically superior to abdominal imaging. To make matters worse, out of the imaging modalities, only CT is standardized across different machine manufacturers, meaning that we cannot easily combine data from different sources into a larger data repository, nor can we compare performances across datasets on a fair ground. Sliver'07 [62] was among the first liver datasets to be published, containing 30 CT volumes, yet did not have any lesion labelings. 3Dircadb [63] released 22 CT volumes with full annotations including the liver and the lesion. TCGA-LIHC [64] further extended the volume count to 116 but unfortunately did not attach any ground truth.

The LiTS challenge in 2017, in response to the lack of useful public dataset, published 201 volumes on CT. It is the latest and largest liver lesion dataset at the time of writing[2]. Details of the dataset can be found in Section 2.1.

[2]The year of 2020 to be exact.

Chapter 2

Data

This chapter will introduce two datasets used in this work: the *Liver Tumor Segmentation (LiTS)* challenge dataset in the public domain, and the private *Longitudinal Liver Lesion MRI* data. Both are datasets focused on liver imaging for lesion detection purposes, with one in computational tomography (CT) and one in magnetic resonance imaging (MRI). The former is used in a per-study evaluation since there are no paired studies that follow up patients, while the latter is specifically curated with longitudinal assessment in mind and is evaluated in terms of both per-study metrics but also per-patient treatment response.

2.1 Liver Tumor Segmentation

LiTS challenge is a dataset released in 2017, containing 131 and 70 contrast-enhanced 3D abdominal CT scans for training and testing, respectively. The dataset was acquired with varying scanners and protocols from six clinical sites. In-plane spatial resolution varies from 0.55 mm to 1.0 mm and inter-slice spacing varies from 0.45 mm to 6.0 mm. The dimensionality of the images is 512×512 across x- and y- but variable in the z-axis.

Some example images can be seen in Figure 2-1. The original challenge has multiple metrics to compete over, but in this work, for the sake of clinical efficacy, we focus on the Dice score introduced in later sections.

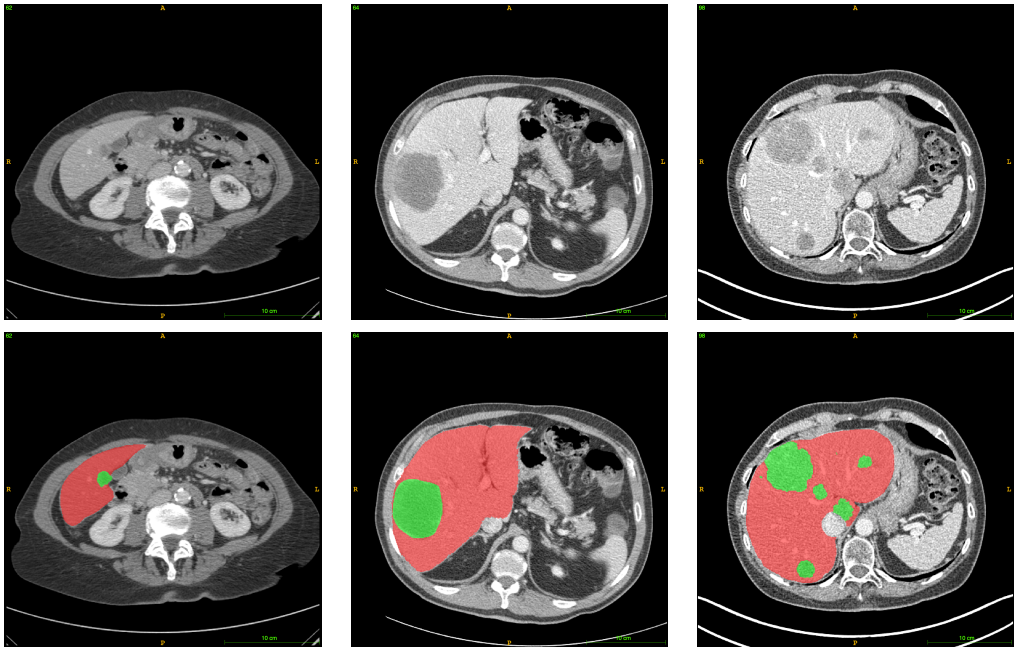


Figure 2-1: **LiTS Data Sample.** Red denotes the liver and green denotes the lesion.

2.2 Longitudinal Liver Lesion MRI

Collaborating with a board-certified radiologist[1], we curate a dataset of 128 abdominal MRIs from 64 patients with neuroendocrine tumors. These patients have undergone at least two consecutive scans (*pairs*) of liver MRI in a large academic medical center to be usable in our study.

To assemble this dataset, we first extract images from the institutional PACS for the cohort and manually review the studies to ensure no severe artifacts, including the motion of metallic implants, present. Note that during the extraction, aside from 64 pairs of studies, we involve 17 extra single studies that do not have follow-up scans, mainly for the purpose of providing more training data. Overall, it sums up to a total of 145 studies. The images are acquired using either Siemens or GE scanners, with their resolutions ranging between 640×320 and 512×512. Examinations are performed with a gadoxetate disodium contrast agent and four dynamic sequences at different times are captured, including pre-contrast, 30 seconds, 70 seconds, and 20 minutes post-contrast. The raw DICOM images from four sequences are then fed

[1]Alex Goehler, M.D., Radiology, Beth Israel Deaconess Medical Center, Boston MA.

		Training	Testing	p-value
Subjects	N	45	19	
Age (years)	mean ± std	58.7 ± 9.7	56.5 ± 11.8	0.43
Female	N (%)	23 (51)	11 (58)	0.61
Follow-up in months	median [min, max]	4 [1, 28]	4 [2, 37]	0.46
Studies	N	90	38	
Primary tumor site	N (%)			0.15
Small bowel		21 (47)	13 (68)	
Pancreas		13 (29)	5 (27)	
Other or unknown		11 (24)	1 (5)	
Study with any lesion	N (%)	86 (96)	37 (97)	0.98
Study with lesion > 1 cm	N (%)	78 (87)	31 (82)	0.46
Lesion per study	median [min, max]	8 [0, 88]	10 [0, 89]	0.22
Lesion per study (greater than 1cm)	median [min, max]	4 [0, 56]	5 [9, 23]	0.28
Number of lesions				0.47
< 10		49	19	
10 - 30		24	9	
> 30		13	9	
Diameter in mm (all lesions)	mean ± std	11.6 ± 11.2	8.4 ± 5.7	<0.05
Diameter in mm (largest lesion per study)	mean ± std	33.9 ± 22.3	19.4 ± 9.4	<0.05

Table 2.1: **Longitudinal Liver Lesion MRI Dataset Statistics.** The table shows the overview of our longitudinal MRI liver lesion dataset.

through a pre-processing pipeline as later described in detail in Section 3.2.1. Two board-certified radiologists (IG and AG) manually annotate the images with liver and lesion labels before finally arriving at a consensus, which we hereafter refer to as the ground truth.

These studies are then randomly split between the training set and test set, with the training set having 45 pairs of studies and 17 single studies and test set having 19 studies[2]. A summary of the statistics of the two splits is shown in Table 2.1. Note that since patients are randomly allocated into the training and test sets, their lesion diameters are not explicitly balanced, thus resulting in a discrepancy in the average lesion diameters across the two sets.

[2]One test patient is in fact dropped later due to their surgical liver resection, which is not in the training data.

Chapter 3

Methodology

3.1 Overview

The overall method can be broken into two major disjoint sections: *segmentation module* and *longitudinal correspondence module*. See Figure 3-1.

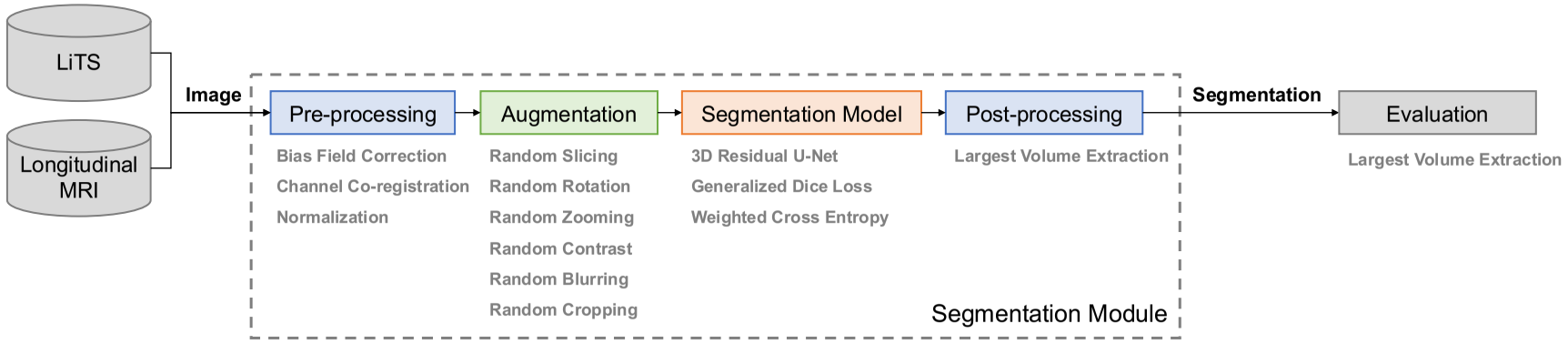
First of all, in the *segmentation module*, images have to go through a *pre-processing* step for dataset curation from raw data. We feed the images through the model, while optionally adding *augmentation* to training images for better model generalization. Finally, there are some heuristics we can apply to the resulting output in the *post-processing* step.

In the *longitudinal correspondence module*, we compare segmentation maps in the 3-dimensional space and identify individual corresponding lesion pairs.

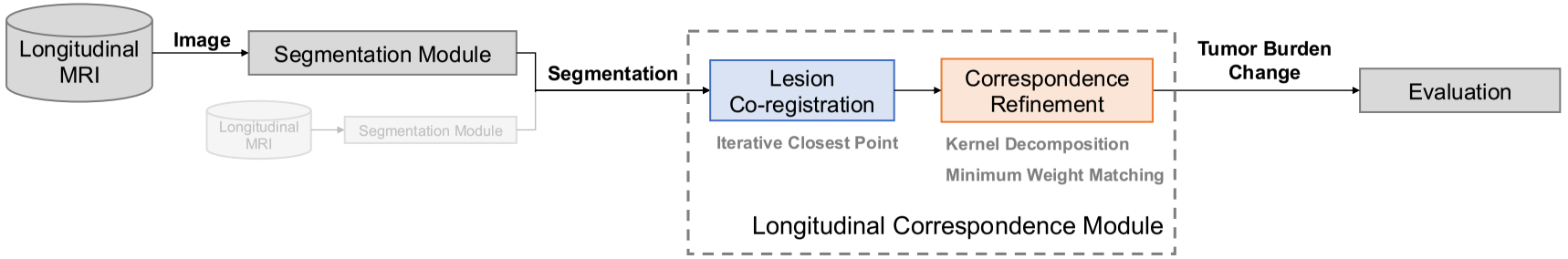
These two modules are evaluated differently. The *segmentation module* is evaluated on a single-image basis, whereas we compare the overall model against radiologists' annotation of longitudinal studies that each contains two images.

3.2 3D CNN for Segmentation

This part of the workflow is relatively similar to previous computational efforts in obtaining segmentation maps, taking the input in the form of 3-dimensional images and predicting the resulting voxel-wise map describing to which label each voxel in



(a) Overview of the *segmentation module*



(b) Overview of the *longitudinal correspondence module*, in combination with the segmentation module

Figure 3-1: **Overview of Methods.** Each row shows the workflow of the module and its key components. The segmentation module takes a single study as input and is evaluated on two datasets. The longitudinal correspondence module takes two studies as input and will be only evaluated against the longitudinal MRI dataset presented in this work.

the image belongs.

3.2.1 Pre-processing

As laid out in previous literature [65], after the acquisition of raw MRI/CT, there are some steps required to prepare the images for later models: *bias field correction* and *channel co-registration*.

Bias Field Correction

The *bias field* is a low-frequency intensity inhomogeneity across the whole image most significantly present in MR images. At low acquisition magnetic field strength around 0.5T (Tesla), this effect is almost nonexistent, yet higher intensity scanners at 1.5/3T introduce coupling effects in the reception coils [66]. The intensity inhomogeneity, while introducing only minor effects to medical experts [67], would inevitably affect the consequent numerical analysis. I apply *N4ITK* [68], a de facto standard in the field for such bias corrections.

Channel Co-registration

Image registration is the process of spatially aligning two or more images. Application scenarios include intra-patient image registration and inter-patient image registration. In intra-patient registration, we obtain images from one or more image modalities (e.g., MRI, CT, and/or PET) at different points in time, resulting in position differences due to substantial patient motions. Inter-patient registration is usually applied when we want to study and compare specific organs with fairly similar shapes (e.g., brain).

The co-registration of multiple MRI series is needed as they are typically acquired in sequences with different exposure times, and we refer to these series as *channels* since they will act as the channel dimension in later neural networks. This step is not required for CT, as we only obtain one series.

Registration involves finding a specific set of parameters that transforms a starting

image in space to align with a second. To achieve that, we require a parameter space for the transform, an initialization, and an objective for optimization [69]. Depending on the application, a *rigid transformation* with 6 parameters or an *affine transformation* of 12 parameters might suffice, and yet in the case of soft tissues such as the liver, we require non-rigid transformations such as free-form deformation [70, 71] and elastic deformation [72, 73]. Initialization of non-rigid transformations can be done with simpler rigid transformations. As for the objective for alignment, there are correlation-based methods [74, 75] which optimize for

$$\operatorname{Corr}(X, Y) = \frac{\mathbb{E}\left[\left(X - \bar{X}\right)\left(Y - \bar{Y}\right)\right]}{\sqrt{\mathbb{E}\left[\left(X - \bar{X}\right)^2\right]}\sqrt{\mathbb{E}\left[\left(Y - \bar{Y}\right)^2\right]}},$$

where X is one image and Y is the transformed version of the second image. $\bar{X}$ and $\bar{Y}$ are their respective means across the whole image.

Yet, most leading methods recently are based on the maximization of *mutual information* (MI) [76] where MI is defined as

$$\operatorname{MI}(X, Y) = H(X) + H(Y) - H(X, Y),$$

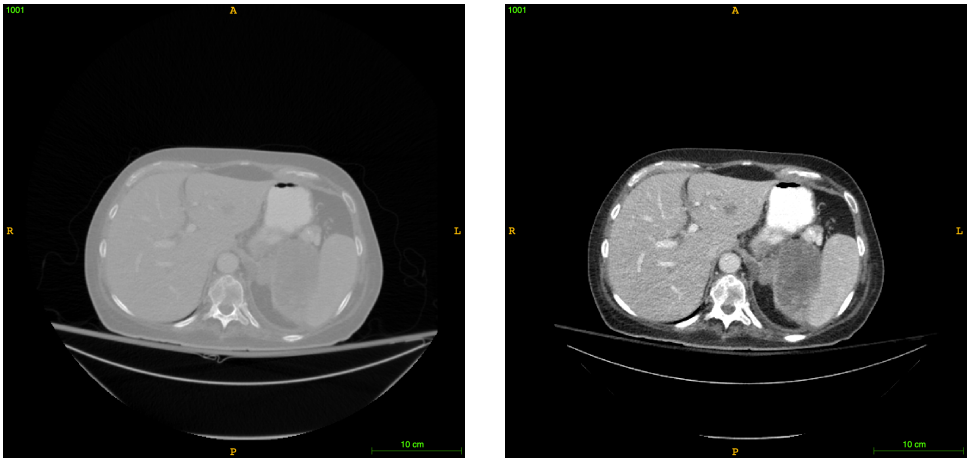
$H(X) = -\mathbb{E}[\log p(X)]$ is the entropy of the probability distribution p of the random variable X,

In this work, I use *elastix* [77], a toolbox for image registration with a sequence of rigid, affine, then b-spline free-form deformations aimed at maximizing mutual information [78].

Normalization

The span of values of the acquired images is the most prominent feature that distinguishes MR images and CT images.

In computational tomography (CT), *Hounsfield Unit* (HU) is a dimensionless unit defined in terms of physical properties of air and water and it standardizes the value across different machine manufacturers. Not only we can directly interpret data from



(a) Before level clipping (b) After level clipping

Figure 3-2: **Value Clipping for CT Images.** In clinical practice, radiologists clip values to $[-150, +250]$ Hounsfield Unit to inspect the soft tissue for pathology. We display the images before and after clipping, and it is clear that the image in (b) has more contrast than in (a).

different sources in a uniform manner, but we do not require explicit normalization except for z-scoring (i.e., rescaling to unit variance and zero mean). However, as the organ of interest in this work is the liver and the soft tissue around it, a typical range of interest in CT for radiologists is $[-150, +250]$ HU [79]. Hence we clip the CT images to this numerical range before z-scoring, as shown in Figure 3-2.

For MR images, we adopt a simple intensity scaling [80] with the minimum being zero intensity and the maximum being the 99^{th} percentile. Normalization is done channel-wise since each of the channels (sequences) may have different numerical ranges.

3.2.2 Segmentation Model

There are several design choices for the segmentation neural network. In general, the input to the network is a 4-dimensional tensor consisting of multiple channels of 3-dimensional images. In particular, we have four dynamic contrast-enhanced MR sequences or one CT sequence. On the other hand, the output should predict the discrete label of each voxel.

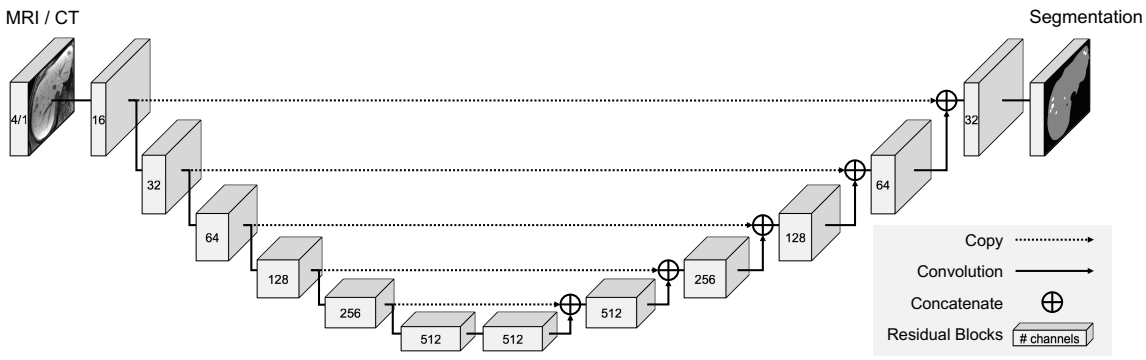


Figure 3-3: **Simplified Model Architecture.** The U-Net [1] employs Residual blocks [2] with padding in convolutions, hence the spatial sizes of feature maps stay the same. The number of channels is labeled on each block.

Model Architecture

U-Net [1, 81] architecture has proven tremendous success for biomedical segmentation tasks. In Figure 3-3, we follow its design and employ a *down-sampling* path and an *up-sampling* path. Different from the original U-Net, which uses 2D plain convolutions, I use 3D residual convolutions with zero-padding. As the depth of the network becomes deep enough to effectively hinder back-propagation, residual blocks allow the magnitude of gradients to stay fairly constant rather than diminishing. The padding allows the output feature map to possess identical sizes as the input image. See Table 3.1 for two models with similar architecture but different depths.

Based on this architecture, I run two cascaded stages of segmentation prediction for lesion detection. As indicated in the LiTS challenge [15], it is a good practice to employ individual detectors that each focus on one task. As the input image size is typically around $512 \times 512 \times 90$, directly extracting lesions from the raw MRI/CT of the abdomen is infeasible due to hardware constraints. Instead, a first *liver detector* can focus on extracting a coarse liver bounding box out of the original image, using a class count $C = 2$ in Table 3.1 and treating lesion labels as part of the liver. Since the main objective of this stage is to roughly identify the location of the liver in the whole abdomen so that the next stage is not constrained by hardware restrictions, we do not require high resolution and hence I resample the image to a spatial resolution of $(4 \text{ mm})^3$.

Layer	**Output size**	**32-layer (ResNet18)**	**48-layer (ResNet34)**
conv0	$224 \times 224 \times 32$	$3 \times 3 \times 3$, 16, stride 1	
conv1*	$112 \times 112 \times 16$	$\begin{bmatrix} 3 \times 3 \times 3, 32 \\ 3 \times 3 \times 3, 32 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 32 \\ 3 \times 3 \times 3, 32 \end{bmatrix} \times 3$
conv2*	$56 \times 56 \times 8$	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 4$
conv3*	$28 \times 28 \times 4$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 6$
conv4*	$14 \times 14 \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 3$
conv5*	$7 \times 7 \times 1$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$	
upconv5*	$14 \times 14 \times 2$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 1$	
upconv4*	$28 \times 28 \times 4$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 1$	
upconv3*	$56 \times 56 \times 8$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 1$	
upconv2*	$112 \times 112 \times 16$	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 1$	
upconv1*	$224 \times 224 \times 32$	$\begin{bmatrix} 3 \times 3 \times 3, 32 \\ 3 \times 3 \times 3, 32 \end{bmatrix} \times 1$	
upconv0	$224 \times 224 \times C$	$3 \times 3 \times 3, C$	

Table 3.1: **3D U-Net with Residual blocks.** Residual building blocks are shown in brackets following the ResNet paper. Down/up-sampling is performed on the first residual block in *conv/upconv* layers with an asterisk. Output sizes are shown assuming the input spatial dimension to be $224 \times 224 \times 32$.

After the liver has been identified, I extract the cuboid bounding the liver as the input for the second stage. This stage can then focus on detecting the lesions, using a class count $C = 3$ in Table 3.1, and classifying each voxel as belonging to background, liver, or lesion. As pointed out in previous works [82], simultaneously predicting both the liver and lesion in a model gives better performance than simply predicting lesions since the liver segmentation also provides a boundary within which lesions can appear. In the second stage, I use images with the original resolution to avoid artifacts introduced by the resampling operation [61]. As a final step, the output segmentation of the second stage is padded to label as background the rest of the abdomen.

Crop Size

While the formulation of fully-convolutional networks (FCN) [38] is able to handle input of variable sizes, in modern neural network training, for the sake of efficiency and stability, we typically use *mini-batches* of examples as the input, which would then require the input images to be of the same shape. Hence the variable-sized input images have to be transformed into a fixed shape to be batched. Here we adopt random cropping that randomizes the location at which the *image patches* are being extracted. Typically in 3D medical imaging, patches that are larger in the x and y axes but smaller in the z-axis, or, *image slabs*, are preferred [61] due to several reasons: (1) scans are conducted in sequence along the z-axis slices, and naturally the voxels along x- and y-directions have the most consistency and coherence; (2) the spacing along the z-axis is typically larger than that of the other two axes, if not equal; and (3) from a computational perspective, the memory and computational constraint is harsher for 3D networks than 2D, and reduction of input sizes is very much desired. Practically, in the first stage, I use a crop size of $64 \times 64 \times 32$, which covers a volume of 25.6 cm $\times$ 25.6 cm $\times$ 12.8 cm; in the second, $224 \times 224 \times 32$, covering a variable-sized grid depending on the resolution.

Loss for Imbalanced Labels

The vast majority of volume in the abdomen in the liver detection task belongs to the background. Incidentally, in the lesion detection task, most of the voxels belong to either the normal liver or the background. This creates very imbalanced counts of labels, and simply enforcing *cross-entropy loss* on the resulting predictions would implicitly bias the model towards dominant classes. We hence introduce loss formulations that mitigate this problem.

Let R be the ground truth segmentation with $r_{nc} \in \{0, 1\}$ being the one-hot encoding of voxel n and class c, P be the probabilistic segmentation map with $p_{nc} \in [0, 1]$, N be the total number of voxels, and C be the number of classes. *Weighted Cross-Entropy* (WCE) can be defined as

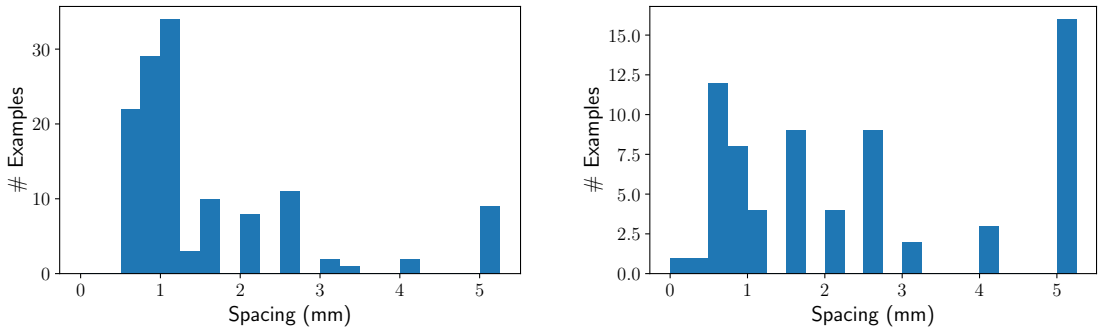
$$\text{WCE} = -\frac{1}{N} \sum_{n} \sum_{c} w_c r_{nc} \log \left(p_{nc} \right),$$

where w_c is a class-wise weight inversely proportional to the number of voxels with that label $w_c = (N/C) \, / \left(\sum_n \sum_c r_{nc} \right)$. This is a trivial extension from [1].

Another loss proposed to address the class imbalance problem is *Generalized Dice Loss* [83, 84]. It extends the Dice overlap measure to a differential form which is able to be optimized via back-propagation. It is defined as

$$\text{GDL} = 1 - 2 \frac{\sum_c w_c \sum_n r_{nc} p_{nc}}{\sum_c w_c \sum_n r_{nc} + p_{nc}},$$

where w_c is different from that in WCE. It is also used to provide invariance to different label set properties. The GDL_v version [83] uses $w_c = 1 / \left(\sum_n r_{nc} \right)^2$, aiming at correcting the contribution of each label by the inverse of its volume.



(a) LiTS training set axial slice spacing

(b) LiTS test set axial slice spacing

Figure 3-4: **Inter-slice Spacing in the LiTS Dataset.** The physical gap between the voxels presented in the volumes along the axial direction can vary dramatically, and thus denser slices can then be subsampled to looser ones as an augmentation.

Input Augmentation

Image augmentation cannot be overemphasized in the context of deep learning, and it is especially true in the case of segmentation networks where the output space is large, easily leading to overfitting. On top of it, the difficulty of collecting large segmentation datasets on medical images is expensive both monetarily and in terms of time. Several augmentation techniques have been explored extensively comparing their effectiveness [85, 86], and the augmentations used in my training pipeline are listed below.

- *Random Slicing*: The inter-slice spacing along the axial (z) direction is greatly inhomogeneous in the LiTS dataset. See Figure 3-4. I take an axial slice every kth slice, where $k_{\max} \geq k \geq 1$ and $k_{\max}$ allows the inter-slice distance to be still smaller than 5 mm. As far as I am aware, I have not seen this type of augmentation in the literature, and yet it is very intuitive in the case of 3-dimensional convolutional neural networks. In the 2-dimensional counterpart, since they only use single slices as the input, there is no need to augment along the z direction.

- *Random Rotation*: In clinical practice, it is common for the inspected patients to have a slight positional difference, and most specifically rotational. After

consulting with a board-certified radiologist, a random rotation along the axial direction of at most 15° is considered.

- *Random Zooming*: The physical size of the abdomen can vary from patient to patient, and zooming in or out randomly by at most 20% is considered. Note that some past works refer to it as *scaling* which I avoid using, as scaling can also ambiguously indicate multiplying the voxel values by a constant factor.

- *Random Contrast*: To make the model more robust against the input value range, I add at most 25% contrast changes to the images. This is also called *jittering*.

- *Random Blurring*: The radiologist I consulted recommends a random Gaussian blur at most 3 mm in radius to be added.

- *Random Cropping*: One of the main techniques to augment the input example is in fact cropping the input image. This is significantly more important in 3-dimensional networks since the parameter space of possible cropping positions is proportional to the cube of input image size, allowing more augmentation outcomes to be made.

Other types of image augmentation techniques such as elastic deformation [87] have been explored, but for real-time training, it is relatively time-consuming [88] compared to other augmentations, and hence I opt not to include it. A summary of the augmentations is provided in Table 3.2.

3.2.3 Post-processing

Although the lesion detection problem is formulated as a segmentation problem in which the labels of individual voxels are independently determined by the neural network, there is an inherent structure: we are predicting several embedded objects in a continuous volume. A very common phenomenon we observe with the predictions is occurrence of straying volumes outside the main liver segment. See Figure 3-5. We propose the following steps to remove the extraneous segments:

Augmentation	**Operation Direction**	**Parameter Range**
Random slicing	Along z axis	Uniform(1, $k_{\max}$), where $k_{\max}$ yields an inter-slice distance of 5 mm
Random rotation	Around z axis	Uniform($-15°, +15°$)
Random zooming	Along x and y axes	Uniform($-20\%, +20\%$)
Random contrast	Voxel-wise	Uniform($-25\%, +25\%$)
Random blurring	Along x and y axes	Uniform(0 mm, 3 mm)
Random cropping	Along all x, y, and z axes	Uniform across all possible cropping locations

Table 3.2: **A Summary of Augmentation Operations.**

1. Since lesions are in fact located within the liver, it should be temporarily considered as the liver while distinguishing the liver against the background. For each of the voxels, we sum of probabilities of it being either liver or lesion as p_1, and assign it a temporary (liver + lesion) label if $p_1 > p_{\text{liver}}$, where p_{liver} is a predetermined hyperparameter.

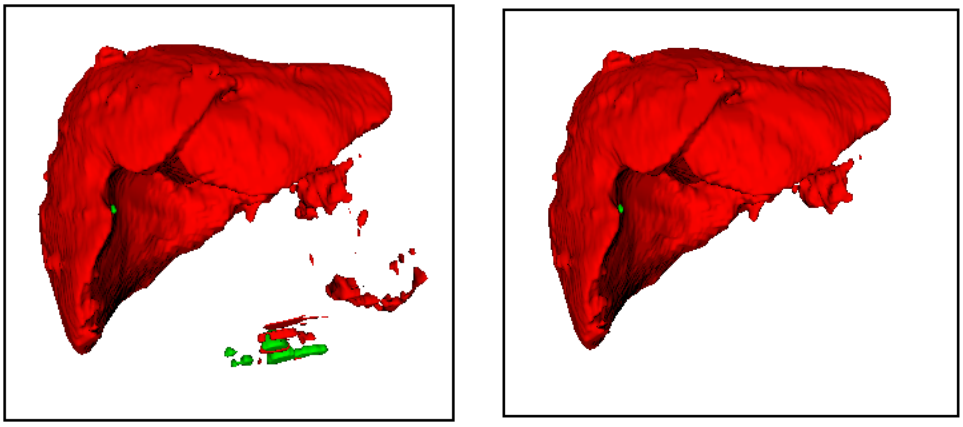
2. Identify the largest connected volume with voxels assigned to (liver + lesion) label in the previous step.

3. For each voxel within the largest connected volume of (liver + lesion), assign it the lesion label if the probability being the lesion $p_2 > p_{\text{lesion}}$; otherwise, assign it as the liver. p_{lesion} is as well a predetermined hyperparameter.

4. For all other voxels outside the largest connected volume, assign them as the background.

A good set of threshold values, as introduced later, is $p_{\text{liver}} = p_{\text{lesion}} = 0.8$ if using weighted cross-entropy for loss and $p_{\text{liver}} = p_{\text{lesion}} = 0.5$ for generalized Dice loss.

This post-processing is applied to both the cascaded stages, with the first stage model ignoring step 3 as the main mission is to only identify the raw location of the liver.



(a) Without post-processing (b) With post-processing

Figure 3-5: **3D Rendering of Predicted Segmentations**. Red indicates liver and green indicates lesion. Without the proposed post-processing there will be extra bits of liver in the surrounding area, which lowers the performance.

3.2.4 Training Details

Model training with 3-dimensional U-Nets typically takes a few days on a fairly powerful workstation (each job taking up an nVidia GeForce RTX 2080 Ti on an Intel Xeon W-2195 with 36 threads). It is apparent that with the required run time, running hyperparameter tuning is infeasible. Moreover, due to the memory requirements of the networks, most experiments do not even support a training batch size larger than 4.

In the training phase, images are fed through the augmentation pipeline, cropped to identical input sizes, and concatenated along a new axis, forming a 5-dimensional tensor (3-dimensional image, the channel size, and the batch size). In the evaluation and test phase, I adopt a fully-convolutional network [38] that feeds the entire input image without cropping. It is easy for images to overwhelm a single GPU and hence we carry out all operations on CPU only in these phases.

As for the hyperparameters used in the experiments, I determine them empirically with a few pilot tests on the LiTS dataset. See Table 3.3. I proceed to use this setting for later experiments on the longitudinal MRI dataset.

Hyperparameter	Value	Criteria
Batch size	As large as possible (typically 4)	As many examples as a GPU can accomodate in the training loop
Initial learning rate	10^{-3} for generalized Dice loss, 10^{-4} for weighted cross entropy	The maximum learning rate for each loss type that allows convergence
Learning rate decay	0.5 over 65536 optimization steps	A reasonably large step count to allow models to come close to convergence
Weight decay	10^{-6}	A reasonable weight decay to prevent exploding weights

Table 3.3: **A Summary of Hyperparameters.** Since an extensive search of hyperparameter is overly time-consuming, I apply some simple heuristics to obtain a set of reasonable base hyperparameters used in all experiments including ablation studies. This set of hyperparameters is used in both LiTS dataset and the longitudinal MRI dataset unless otherwise stated.

3.3 Longitudinal Lesion Correspondence

The previous section lays out the methods that identify, in a scanned image, whether individual voxels are lesions, ordinary liver tissues, or irrelevant background. A very important problem in the clinics is whether the disease has progressed in between two imaging studies, and one of the easiest ways of assessing this is to look at the overall volume of the lesions, also called the *tumor burden*. However, such a metric would easily overlook the development of individual lesions as the disease might be partially *progressive* (i.e., growing) while being *responsive* (i.e., shrinking) to the treatment elsewhere. On top of the lack of granularity in characterizing the disease, the actual volumes of the lesions are hard to obtain due to the intensive requirement of expert annotations, and thus there are proxy standards such as the WHO tumor response criteria [89, 5] and the RECIST criteria [8]. They first identify lesions presented in two different studies, obtain simple measurements from them, and quantify the tumor changes with metrics and categorizations in a standardized way.

Other than just acquiring the measurements, the radiologists who inspected the studies are in fact implicitly trying to pair up lesions from both of the studies by reading two studies side by side. While automated methods are excellent in delivering

results for well-defined problems, this task of finding correspondences in two studies is non-trivial. Hence, in this section, I focus on comparing two images that are already segmented. Rather than comparing voxel-wise, a connected volume consisting of voxels of an identical label are grouped into *lesion objects*, and the analysis in this section is primarily done on these objects.

3.3.1 Lesion Co-registration

As the main goal is to compare two imaging studies, the *baseline* study and the *follow-up* study, done at different points in time, usually months apart, the lesions identified are inherently from different patient positions and different spatial coordinate systems. One immediate task is to align two images in a lesion-aware manner as opposed to only maximizing the image correlation. Solely applying the image registration techniques described in 3.2.1 would not necessarily transform lesions correctly onto another study as the previous co-registration technique focuses more on the organ boundaries and silhouettes.

In that regard, there is a very similar problem in the context of computer graphics. Registration of two different 3-dimensional free-form shapes, specifically represented by a set of points, has been widely investigated. *Iterative Closest Point* (ICP) [90, 91] takes two sets of points without any correspondence and iteratively figures out (1) the one-to-one correspondence relationship between the two groups and (2) the transformation from one space to another which minimizes the distance of the paired points. See Algorithm 1 for the complete algorithm.

Taking that into the context of lesion registration, we can utilize the ICP algorithm in the following manner to obtain a rigid transform from two segmentation maps:

1. Find all connected lesion volumes in both the baseline and follow-up segmentation maps.

2. Calculate the centers of mass for each of the lesion objects.

3. Apply ICP to two sets of centers of mass.

Algorithm 1: Iterative Closest Point for Lesion Co-registration

Input : Two sets of center of mass $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{X}|}$ from *baseline* and $\mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^{|\mathcal{X}|}$ from *follow-up*.

Output: Rigid transformation $\mathcal{T} : \mathcal{Y} \to \mathcal{X}$, $\mathcal{T}(\mathbf{y}) = \mathbf{R}\mathbf{y} + \mathbf{t}$.

$\mathbf{R} \leftarrow \mathbf{I}_3$, $\mathbf{t} \leftarrow \mathbf{0}$;

$\bar{\mathbf{x}} \leftarrow \frac{1}{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \mathbf{x}_i$, $\bar{\mathbf{y}} \leftarrow \frac{1}{|\mathcal{Y}|} \sum_{j=1}^{|\mathcal{Y}|} \mathbf{y}_j$;

while *not converged* **do**

`// Find correspondence`

$c_i \leftarrow \arg\min_c \|(\mathbf{x}_i - \bar{\mathbf{x}}) - \mathcal{T}(\mathbf{y}_c - \bar{\mathbf{y}})\|$;

$\bar{\mathbf{x}} \leftarrow \frac{1}{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \mathbf{x}_i$, $\bar{\mathbf{y}} \leftarrow \frac{1}{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \mathbf{y}_{c_i}$;

`// Obtain optimal transform`

$\mathbf{W} \leftarrow \sum_{i=1}^{|\mathcal{X}|} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{y}_{c_i} - \bar{\mathbf{y}})^\top$;

$\left(\mathbf{U}, \mathbf{S}, \mathbf{V}^\top\right) = \mathtt{SVD}(\mathbf{W})$;

$\mathbf{R} \leftarrow \mathbf{U}\mathbf{V}^\top, \mathbf{t} \leftarrow \bar{\mathbf{x}} - \mathbf{R}\bar{\mathbf{y}}$

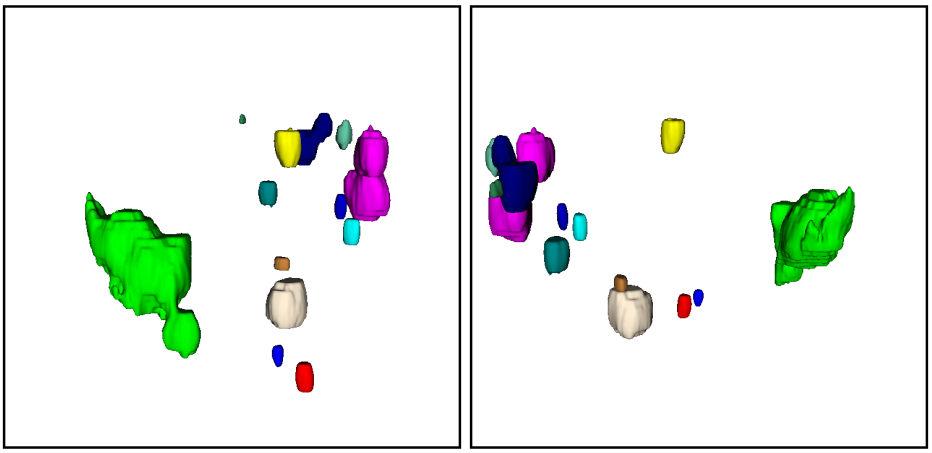
end

An example of the problem at hand is shown in Figure 3-6, which compares two studies of the same patient in time. The main subjects of investigation in this section are the individual lesions as labeled in different colors. Note that typically radiologists do not look at direct 3-dimensional rendering but scroll through a sequence of 2-dimensional slices of the image.

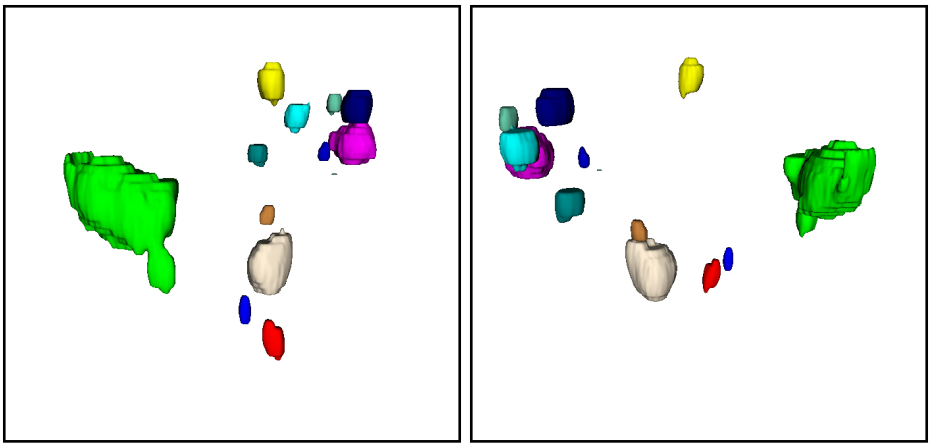
3.3.2 Correspondence Refinement

The ICP described in the previous section, while being able to roughly align the spatial transformation of the lesions in the baseline and follow-up studies, fails to capture some realistic constraints in real-world scenarios. Since the two studies in consideration look at snapshots at different points in time, we can expect lesions to more or less have a one-to-one relationship, except in the case of *progressive disease* where new lesions are formed or *complete response* where the lesions completely vanish. On top of that, the relative sizes of the lesions are likely to stay similar: larger lesions will remain the dominant lesions in a follow-up examination in most of the cases.

With the above observation, I hence propose to apply some heuristics and rules to refine the lesion correspondences so that they end up being one-to-one. The following



(a) The baseline study lesions



(b) The follow-up study lesions

Figure 3-6: **3D Rendering of Individual Lesions for Two Associated Studies.** Different color indicates different lesion objects. Two different viewpoints are shown.

rules define a *reward* between a pair of lesions (i, j) from the baseline and follow-up studies, respectively, and the objective function will be finding a mapping to allow maximization of the corresponding rewards.

Affinity Reward

While ICP should achieve most of the lesion alignment, it is not perfect. I allow lesions in the vicinity of another to be matched with it, preferring closer ones, thus I define an *affinity reward*

$$R_{\text{affinity}}\left(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j\right) = \exp\left(-\frac{1}{2}\frac{\left|\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j\right|^2}{\sigma_a^2}\right), \tag{3.1}$$

where σ_a is a hyperparameter which I set to be 5 mm empirically.

Volumetric Similarity Reward

As laid out before, across different studies of the same patient, we expect the lesions to be stable in terms of size, except for the case where there are *partial response* or *progression* and we expect lesions to somewhat grow/shrink at a similar rate. I encourage lesions of similar volumes to be matched, thus I define the *volumetric similarity reward* which describes a similarity between lesion volume pairs to be

$$R_{\text{volume}}(V_i, V_k) = \exp\left(-\frac{1}{2}\frac{\left(\log\left(V_i/V_j\right)\right)^2}{\sigma_v^2}\right), \tag{3.2}$$

where σ_v is a hyperparameter controlling how much logarithmic difference is tolerable and is set to be 0.5 empirically.

The rewards defined here are applied to a pair of detected lesions in a baseline i and a follow-up study j. Overall, there will be $|\mathcal{X}| \times |\mathcal{Y}|$ reward values $R(i,j) = R_{\text{affinity}}(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) + R_{\text{volume}}(V_i, V_j)$, and the goal is to find a one-to-one relationship $i \leftrightarrow j$ so that $\sum_{(i,j)} R(i,j)$ is minimized, i.e., the similarity is *maximized*.

To achieve this, we feed the kernel matrix into the Kuhn-Munkres algorithm [92], also known as the Hungarian algorithm, which looks for a minimization/maximization of objectives between two sets of items while deriving one-to-one correspondences between the two. After finding the pairs, any unmatched lesion, depending on whether it belongs to the baseline or follow-up study, is categorized as *resolved* or *newly developed*.

3.4 Evaluation

3.4.1 Accuracy

In evaluating multi-class classification, there are several categories of classification results depending on what the ground truth and the predictions are. We calculate

		Ground truth		
		Positive	Negative	
Prediction	Positive	True positive (TP)	False positive (FP)	Precision = $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$ (PPV)
	Negative	False negative (FN)	True negative (TN)	
		Recall = $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$ (Sensitivity)	Specificity = $\frac{\mathrm{TN}}{\mathrm{FP+TN}}$	Acc = $\frac{\mathrm{TP+TN}}{\mathrm{TP+FN+FP+TN}}$ Dice = $\frac{2\times\mathrm{TP}}{2\times\mathrm{TP+FP+FN}}$

Table 3.4: **Metrics Used in Classification Evaluation.** PPV stands for positive predictive value.

true positive (TP), false negative (FN), false positive (FP), and true negative (TN) counts. There are also several derivative metrics from these raw counts listed in Table 3.4. These metrics are used in two different contexts for evaluation: voxel-wise and lesion-wise.

Voxel-wise Accuracy

The output of the segmentation module can be easily evaluated on a voxel-wise basis. Since there are two classes other than the background, namely the liver and the lesion, the following metrics are evaluated over all voxels in the image of interest for both of the classes: precision, recall (sensitivity), specificity, accuracy, and Dice coefficient, treating other classes as negative.

The main issue here is that most evaluation units here belong to TN, which would make metrics such as accuracy and specificity indistinguishably close to one regardless of classifier quality. Hence it is more desired to focus on other metrics.

In the case of Dice coefficient, since it is evaluated based on the overlap of ground truth and predicted outcome, the resulting number can be noisy for cases with only a small number of voxels labeled as positive. To mitigate having highly variable statistics, I evaluate Dice with two versions.

- **Per-study Dice (macro Dice)**: This is the most naïve way of calculating a single Dice score for multiple studies, derived as $\mathrm{Dice}^{\mathrm{macro}} = \frac{1}{N}\sum_n \frac{2\times\mathrm{TP}_n}{2\times\mathrm{TP}_n+\mathrm{FP}_n+\mathrm{FN}_n}$ where N is the total number of studies. The metric is simply averaged across

all studies.

- **Global Dice (micro Dice)**: Instead of averaging after the division in evaluating per-study Dice score, we sum up the raw counts by calculating $\text{Dice}^{\text{micro}} = \frac{\sum_n 2\times\text{TP}_n}{\sum_n 2\times\text{TP}_n+\text{FP}_n+\text{FN}_n}$.

It is easy to tell that the difference between the per-study and global metrics is whether the averaging happen pre- or post- Dice calculation. The same computation applies to other metrics.

Lesion-wise Accuracy

The evaluation with individual voxels, while depicting the overlap fraction between predictions and radiologist labelings, fails to capture a critical problem in a clinical setting: *how many lesions are there and how many have been captured by the algorithm?* In reaction to that, a metric on the lesion level is needed to describe this correctly.

I define the lesion-wise accuracy by first looking for segmentation regions, labeling them into distinct *objects*[1], and only operating on top of these disjoint objects in the calculation of accuracy. To be specific, any predicted lesion is labeled TP as long as the predicted lesion overlaps with a ground truth lesion. The ground truth lesion then becomes *claimed* by a prediction and any subsequent matches to it will be invalidated. I repeat this matching pursuit until every predicted lesion has been addressed, and any predicted lesion that has not been matched to a ground truth is considered FP. Aside from unmatched predictions, unmatched ground truth lesions are labeled FN.

Varying Overlap Due to the nature of the lesion matching process being fairly similar to *object detection* in computer vision [93], we consider altering the threshold above which the prediction is considered a match with the ground truth. The main idea behind this is that by raising the intersection-over-union (IoU), we raise our bars for a detection to be considered successful, and thus create a harsher evaluation

[1] One way of doing it is via `scipy.ndimage.label`.

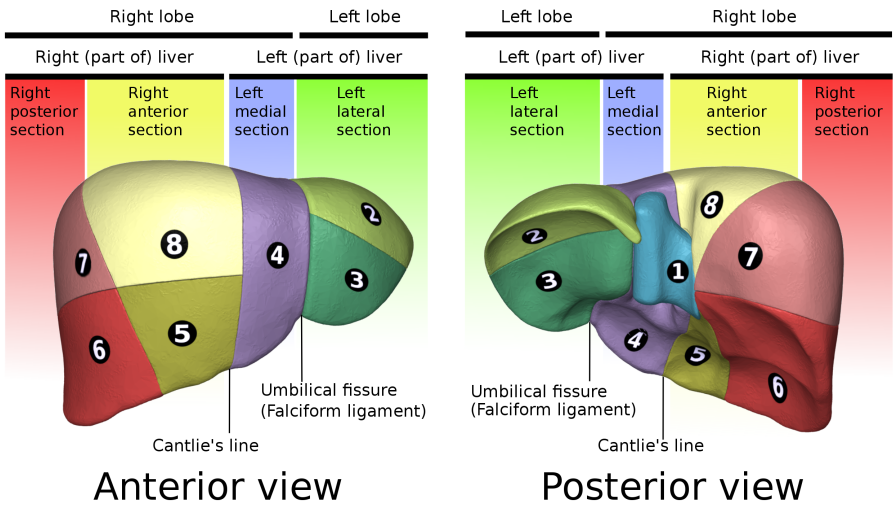


Figure 3-7: An illustration of the Couinaud liver segment system[2]. Note that while only eight segments are shown, we use an alteration that subdivides segment 4 into sub-segments, forming a total of nine.

criterion. By inspecting the performance change with respect to varying overlap levels, we obtain a sense of how spatially accurate the detections are – a more spatially accurate prediction is more resistant to high overlap thresholds.

The matching process as described earlier would then correspond to an IoU of 0, as all matches are accepted as long as the overlap is nonzero. A typical set of values used as the threshold is $\{0, 0.25, 0.50\}$.

Performance dependence on lesion size Besides an overview of all lesions, we are also interested in how well lesions can be discovered depending on their sizes. In response to that, we bin lesions into three categories by their diameter similar to [94]: (1) $<$ 10 mm, (2) 10 – 20 mm, and (3) $>$ 20 mm. Aside from global metrics that cover all lesions, the metrics are then reported additionally in these categories.

The definition of the accuracy is not as straightforward as the conventional binary classification and the sum of three count bins (TP/FN/FP) is not constant, but the intention of defining these baskets is that we are able to define *precision* and *recall* (*sensitivity*). This family of metrics, as informed by a radiologist, suffices to provide basic clinical understanding about the performance of the system.

[2]Generated by Database Center for Life Science (DBCLS) - Polygon data is from BodyParts3D,

Couinaud Segment Accuracy

One major drawback of the voxel-wise evaluation as mentioned above is the extreme imbalance incurred by the majority of the background voxels. On top of the imbalance, Dice score is typically used in the context of computer scientific evaluation while it lacks a direct clinical meaning. The evaluation by lesion as proposed above partially mitigates this problem by only looking at lesions as a whole and not by its voxels, and yet it is still missing a critical clinical interpretation of true negatives (TN): how non-diseased patients can be accurately diagnosed as disease-free. The lack of the notion TN also disables one of clinicians' favorite tools – *specificity*, and thus there is a need to push for analyses at a disease level.

It is possible to define a cohort of both diseased patients with lesions in the liver and another group of disease-free patients, followed by checking, in a per-patient manner, if any lesion is detected correctly by the system in a patient in which lesions should be present. However, the cohort defined in the data I investigate only includes patients that have metastatic neuroendocrine tumors, rendering the per-patient TN evaluation impossible. Hence, I turn to using an evaluation based on the 9-segment Couinaud system for liver [95] as shown in Figure 3-7. I provide a diagnostic characteristics comparison between the algorithm and the ground-truth as manually identified. This approach would allow a definition of TP, FN, TP, and TN in a more balanced manner on the liver segments[3], thus deriving both sensitivity and specificity on a predefined clinically meaningful liver segment.

In the Couinaud system, a segment is deemed positive if at least one lesion is found within the said segment, and this rule applies to both the ground truth and the predictions. Should any lesion be found on the border, it is assigned to the Couinaud segment that overlaps with it the most. The absence of lesions in the segment is naturally negative. Across all patients, the classification buckets (TP/FN/TP/TN) are then calculated and metrics such as *sensitivity* and *specificity* are derived.

[3]Note the Couinaud liver segments here differ from the task of *segmentation*, which refers to the classification of image elements.

We used the variance estimator derived in [3] to estimate the 95% confidence interval, as this accounts for the interdependence of the data on the Couinaud segment level.

3.4.2 Interval Change Assessment

To assess the lesion detection and correspondences output by Section 3.3.2, two radiologists were asked to use RECIST 1.1 [8] restricted to the liver to annotate the diagnosis. The process was carried out in the following manner.

- Two dominant lesions are identified as *target lesions* in the baseline study, determined by their size and suitability to be repeatedly measured accurately. The diameters of these target lesions are measured and classified into categories below.
  - *Complete Response*: If all target lesions disappear.
  - *Partial Response*: If the sum of the longest diameters decreases by over 30% or more.
  - *Progressive Disease*: If the sum of the longest diameters increases by at least 20% and at least 5 mm.
  - *Stable Disease*: Otherwise.
- Other lesions with a diameter greater than 10 mm are considered *non-target lesions* whose diameters are not recorded, yet their presence and disappearance are noted across the baseline and follow-up studies and classified as follows.
  - *Complete Response*: If all non-target lesions disappear.
  - *Non-Complete Response/Non-Progressive Disease*: If persistence of one or more non-target lesions is observed but no new lesions.
  - *Progresive Disease*: If appearance of one or more new non-target lesions is observed.

Target Lesions	**Non-Target Lesions**	**New Lesions**	**Overall Response**
Complete Response	Complete Response	No	Complete Response
Complete Response	Non-Complete Response Non-Progressive Disease	No	Partial Response
Partial Response	Non-Progressive Disease	No	Partial Response
Stable Disease	Non-Progressive Disease	No	Stable Disease
Progressive Disease	Any	Any	Progressive Disease
Any	Progressive Disease	Any	Progressive Disease
Any	Any	Yes	Progressive Disease

Table 3.5: **RECIST 1.1 Criteria.** The overall response is determined based on three aspects of evaluation on *target lesions*, *non-target lesions*, and *new lesions*.

- Note if there is appearance of new lesions.
- Combine the observations above and refer to Table 3.5 for the overall evaluation of response.

The evaluation of the automatic methods is compared against the overall RECIST criteria as annotated by two radiologists on the hold-out set of 18 longitudinal cases. Kohen's kappa is used to determine the concordance between the two.

Chapter 4

Results and Discussion

In this section, I present the results for experimentation on the *segmentation module* including different modeling choices and losses laid out in Section 3.2 and the *longitudinal correspondence module* as described in Section 3.3.

4.1 Segmentation Model Design

The end goal of this work is to build a full framework capable of detecting lesion changes in time, and the segmentation model that generates a single point estimate is, without a doubt, a critical piece in the pipeline. It not only has to perform reasonably well in terms of concordance with the radiologists but be consistent and stable enough. There are a few model choices described in Section 3.2.2 where it is not immediately clear what performs better. I hereby present some experimentation on the publicly available dataset LiTS, comparing the designs in Table 4.1, and the results and discussion are elaborated below. Note the model here is not state-of-the-art if compared on the LiTS challenge website[1] where several works such as 3D AH-Net [96], H-DenseUNet [61], and V-Net [97] top the leaderboard. However, the main objective here is not achieving the highest scores but to perform a comparative study on which components of a neural network design help the most in the learning process.

[1]LiTS Challenge Leaderboard: `https://competitions.codalab.org/competitions/17094`.

Model Arch	Aug	Crop Size	Loss Func	Steps	Loss	Acc	Dice
Augmentation							
ResNet18	–	224 × 224 × 32	GDL	200k	0.385	0.979	0.601
ResNet18	Slicing	224 × 224 × 32	GDL	200k	0.545	0.977	0.599
ResNet18	Rotation	224 × 224 × 32	GDL	200k	0.458	**0.981**	0.643
ResNet18	Zooming	224 × 224 × 32	GDL	200k	0.343	0.980	0.652
ResNet18	Contrast	224 × 224 × 32	GDL	200k	0.417	0.979	0.586
ResNet18	Blur	224 × 224 × 32	GDL	200k	0.338	0.979	0.640
ResNet18	All	224 × 224 × 32	GDL	200k	**0.291**	**0.981**	**0.684**
Crop Size							
ResNet18	All	224 × 224 × 32	GDL	100k	**0.328**	**0.979**	**0.652**
ResNet18	All	112 × 112 × 64	GDL	100k	0.339	0.976	0.630
ResNet18	All	112 × 112 × 32	GDL	100k	0.425	0.977	0.619
Model Architecture							
ResNet18	All	112 × 112 × 64	GDL	100k	**0.338**	0.976	0.630
ResNet34	All	112 × 112 × 64	GDL	100k	0.339	0.974	**0.631**
ResNet50	All	112 × 112 × 64	GDL	100k	0.345	**0.977**	0.624
Loss Function							
ResNet18	All	112 × 112 × 64	GDL	200k	–	0.979	**0.664**
ResNet18	All	112 × 112 × 64	WCE	200k	–	0.979	0.631

Table 4.1: **LiTS Model Design Experiments.** Multiple sets of component choices are presented here, with each varying one parameter. The *Model Arch* column represents the down-sampling branch; in *Aug* I compare augmentations, *all* denotes that all augmentations are used; *Crop Size* is the random cropping output size in the training phase; GDL stands for generalized Dice loss and WCE means weighted cross-entropy. *Steps* are the number of optimization steps. All numeric values are reported on the LiTS test set. Dice score is the per-study version.

As found in a few pilot tests, the first stage of the pipeline, which is responsible for locating the liver, proves to be rather insensitive to parameter choices and consistently yields a Dice score above 0.99. I do not evaluate this stage and focus on the second that detects lesions.

Augmentation

I begin by using each augmentation technique on the baseline model. In Table 4.1 under the *augmentation* section, random cropping is always applied since it produces identical shapes for images in the same batch. With the addition of different augmen-

tation, we find that random slicing and random contrast do not contribute in terms of Dice score, which is the most important indicator for efficacy. It is also interesting to note that the test loss and test accuracy are not too indicative as the loss spans a wide range for different experiments, and accuracy is almost always close to one.

The most useful augmentations to add are, ordered by Dice increment, random zooming, random rotation, and then random blurring. They respectively increase the Dice score by 0.051, 0.042, and 0.039 points. Practically, combining all would give the best Dice score (0.684) as shown in the last row in the group, in exchange for some input pipeline slowdowns.

Crop size

The crop size controls how much memory is used in training, thus allowing a different number of images in a mini-batch. Note that in the test phase I use a fully-convolutional network (FCN), thus waiving the need to crop images for input.

I try three random crop sizes, with the largest being $224 \times 224 \times 32$ and the smallest being $112 \times 112 \times 32$. Note that due to the discrepancy in memory consumption, $112 \times 112 \times 32$ is, in fact, able to facilitate a batch size of 16 as opposed to 4 on other runs, and hence we allow the model to be trained with that batch size. The reason for allowing a bigger batch size is that the smaller input crop size is already a disadvantage from a field-of-view perspective (i.e., it is only observing a smaller region of the image), we compensate by offering it more examples per batch to learn from in order for the version to be compared fairly.

Even with the ability to see more examples, $112 \times 112 \times 32$ still fails to outperform other crop sizes. $224 \times 224 \times 32$ performs the best with a significant gap ahead of others. Due to memory constraints, I have not tested larger crop sizes. In subsequent experiments in the next section, I will use $224 \times 224 \times 32$ by default.

Model Architecture

I compare three different depths of ResNet [2] on their performance used as the down-sampling branch in Figure 3-3. Two (ResNet18 and ResNet34) of the three are laid

out in detail in Table 3.1 and ResNet50 is constructed similarly to the original ResNet paper, except that it is in 3D.

Refer to the *architecture* section in Table 4.1. ResNet34 yields just marginally better performance on Dice score compared to ResNet18, while ResNet50 seems to be overkill. Based on the observations above, using ResNet18 is typically sufficient.

Loss Function

Comparing two loss functions from Section 3.2.2, we have the results in Table 4.1 in the last group, *loss function*. The loss values are not listed as the two have different ranges, but it is clear that GDL is slightly better. In fact, there is another advantage to GDL, which will be further discussed in depth below.

4.2 Segmentation Accuracy on MRI

Now that I have established the basic reasoning of parameter choices in the neural network, I turn to concentrate my focus on the longitudinal lesion dataset in the modality of MRI.

The immediate difference as we can observe here is the change in the number of channels. In LiTS data, we have only one CT channel, which becomes four in the dynamic contrast MR acquisition. First, I am interested to see if performance changes at all between the two datasets. I train a base model on 105 training images and evaluate on 38 hold-out images. Note that out of the 105 training images, 17 are unpaired (i.e., studies without prior or later studies from the same patient) and that there are no patient overlaps between the two splits.

On the test images, I compute the metrics including their per-study and global versions as described in 3.4.1. As there are two important parameters, namely the thresholds in the post-processing p_{liver} and p_{lesion}, the metrics are provided on a grid of these parameters. As the first stage of the model, which locates the liver within the abdomen, is found to be very insensitive to the configuration (Dice score is consistently above 0.99), we do not explicitly evaluate this stage and only focus on the lesion

detection model.

The base model uses a ResNet18 backbone, a crop size of $224 \times 224 \times 32$, an augmentation pipeline consisting of random rotation, random zooming and random contrast, a batch size of two[2], and an initial learning rate of 10^{-3} using the Adam optimizer. The learning rate is decayed by 0.5 every 2^{16} steps until 4 decays, when I stop training.

4.2.1 More on Loss Functions

I conduct experiments with two loss functions again: weighted cross-entropy (WCE) and generalized Dice loss (GDL) as defined in Section 3.2.2. The per-study test set metrics are provided in Figure 4-1. The most critical metric to which to pay attention here is the Dice score of lesion detection, which is presented in the second row in each of the sub-figures: while WCE achieves a higher Dice score of 0.646 at its plateau, better than the 0.637 for GDL, it is easily seen that the score is extremely sensitive to the choice of parameters.

To arrive at a good set of hyperparameters, one can look at both rows of contour plots, using the top row to select p_{liver} and the bottom to select p_{lesion}. An optimal choice of for WCE is $p_{\text{lesion}} \sim 0.8$ and $p_{\text{liver}} \sim 0.8$ where the Dice score plateaus. As for GDL, selecting $p_{\text{liver}} = p_{\text{lesion}} = 0.5$ is a fair balance among Dice, precision, and recall.

Owing to its relative stability, I favor GDL models and use them in subsequent analysis and longitudinal lesion studies.

4.2.2 Global and Per-study Voxel-wise Accuracy

Aside from the per-study metrics, if we also compare the global metrics in Figures 4-1b and 4-2c, we can identify the scores in the global case to be generally higher (e.g., for lesion Dice, 0.726 in global vs. 0.636 in per-study). This is because cases with minor lesions, which are in fact harder to detect, are down-weighted in the global

[2]The most images I can fit in an nVIDIA GTX 2080 Ti with this input size.

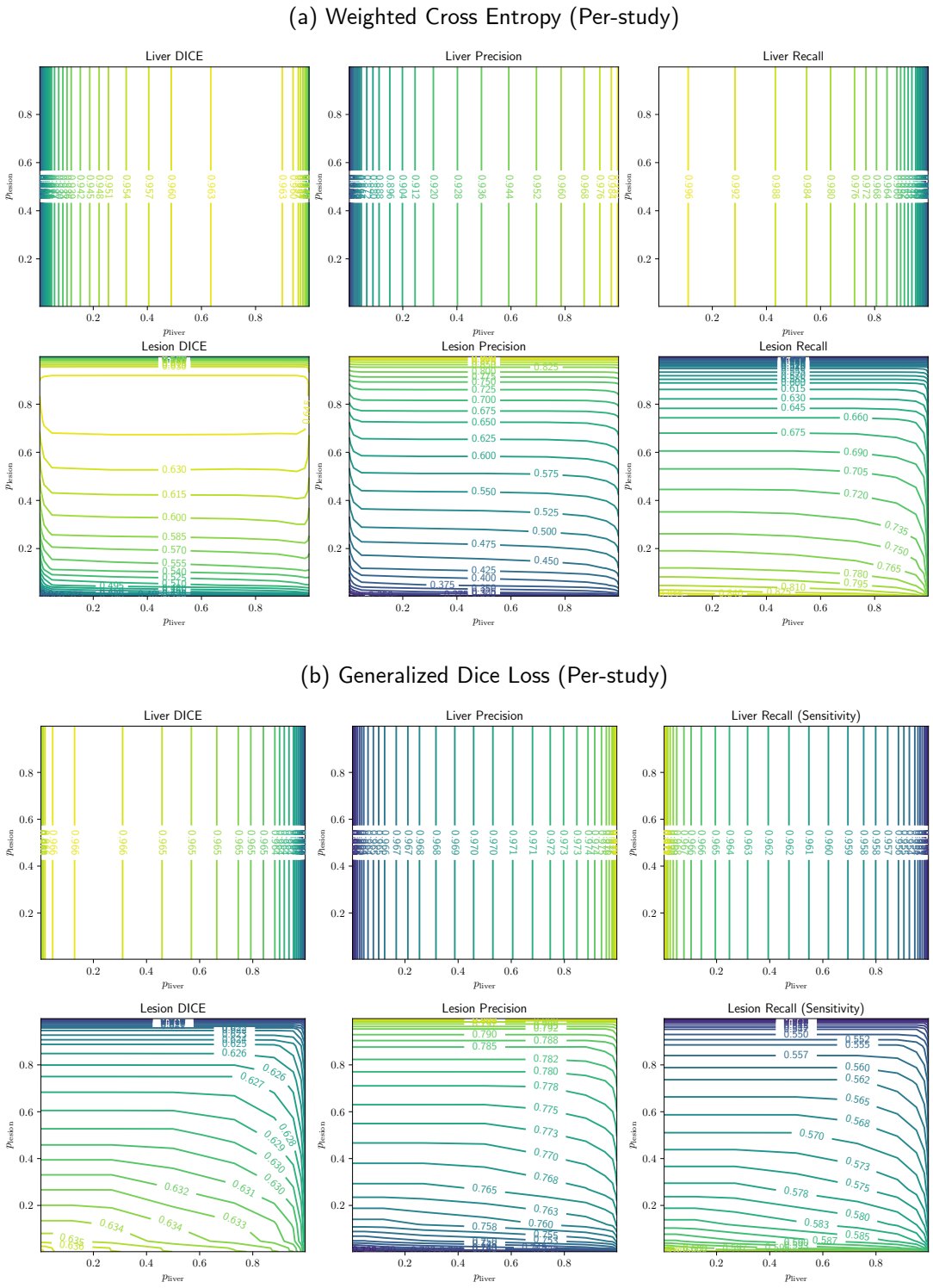


Figure 4-1: **Per-study Segmentation Performance for Different Loss Functions.** In each plot grid, the top row shows the metrics for liver detection and the bottom shows lesion detection. Metrics are shown for hyperparameters p_{liver} and p_{lesion} which determine the post-processing probability thresholds.

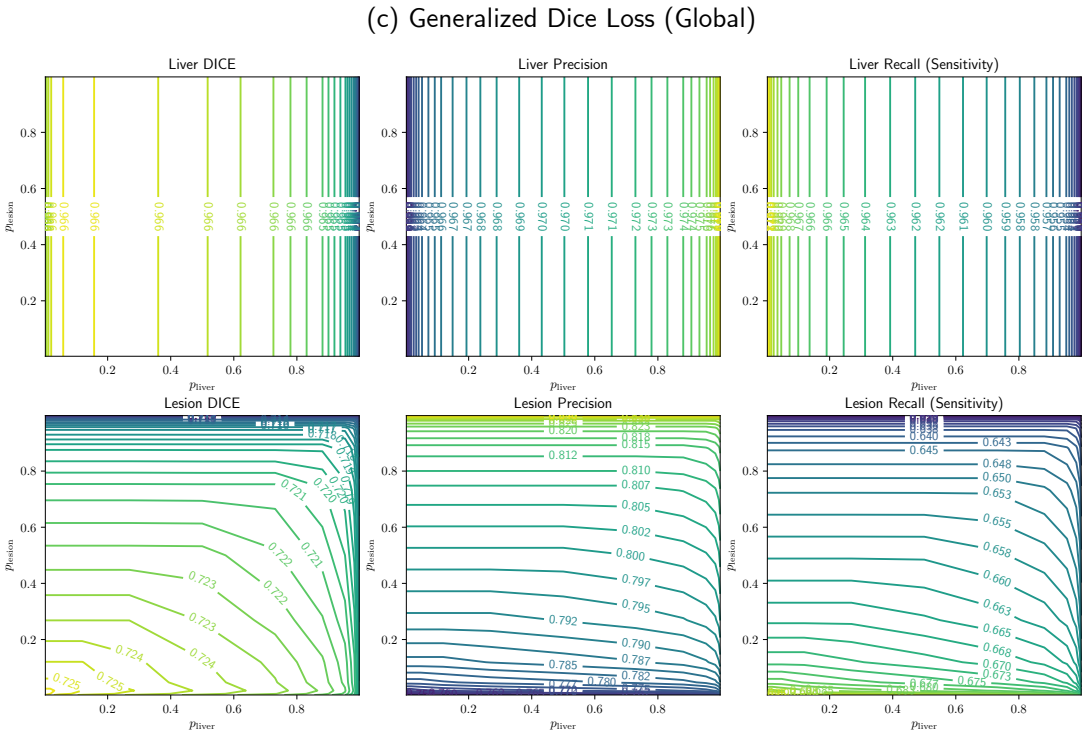


Figure 4-2: **Global Segmentation Performance Contour Plot on MRI Dataset.** Continued from Figure 4-1. Global means that the TP/FN/FP/TN counts are first aggregated over the entire dataset and then the metrics are calculated accordingly. Larger lesions will be up-weighted in this manner.

case where the raw voxel counts are aggregated globally.

4.2.3 Detection of Liver Lesions

To evaluate the efficacy of the segmentation algorithm, I follow the descriptions in Section 3.4.1 and derive detection results for lesion objects. In Table 4.2, I present the sensitivity, the positive predictive value (PPV), and the Dice score. In total, there are 618 lesions in the dataset, among which 437 are detected along with some false positives that are not present in the ground truth. This is a global sensitivity of 0.707 (95% CI: 0.617, 0.798) and a PPV of 0.857 (95% CI: 0.792, 0.922).

As the metric version changes from per-study, to per-patient, then to global, the metrics gradually increase due to the up-weighting of the larger and easier-to-detect lesions. Easier detection for larger lesions can be also observed in the global metric

Sensitivity

Overlap	TP	FN	FP	Per-study	Per-patient	Global			
						All	<10mm	10-20mm	>20mm
0.00	437	181	73	0.677 (0.585, 0.770)	0.688 (0.590, 0.785)	0.707 (0.617, 0.798)	0.487 (0.353, 0.621)	0.844 (0.804, 0.884)	0.938 (0.854, 1.021)
0.25	394	224	120	0.637 (0.544, 0.731)	0.652 (0.556, 0.747)	0.638 (0.560, 0.715)	0.416 (0.348, 0.485)	0.780 (0.709, 0.852)	0.769 (0.627, 0.912)
0.50	265	353	250	0.453 (0.349, 0.557)	0.467 (0.344, 0.591)	0.429 (0.301, 0.557)	0.181 (0.000, 0.386)	0.588 (0.501, 0.676)	0.651 (0.480, 0.822)

Positive Predictive Value

Overlap	TP	FN	FP	Per-study	Per-patient	Global			
						All	<10mm	10-20mm	>20mm
0.00	437	181	73	0.842 (0.785, 0.899)	0.845 (0.759, 0.931)	0.857 (0.792, 0.922)	0.670 (0.521, 0.819)	0.964 (0.932, 0.996)	1.000 (1.000, 1.000)
0.25	394	224	120	0.783 (0.722, 0.844)	0.799 (0.726, 0.871)	0.767 (0.694, 0.839)	0.531 (0.369, 0.692)	0.907 (0.857, 0.958)	1.000 (1.000, 1.000)
0.50	265	353	250	0.547 (0.450, 0.643)	0.578 (0.453, 0.704)	0.515 (0.436, 0.593)	0.222 (0.059, 0.385)	0.687 (0.589, 0.784)	0.933 (0.833, 1.000)

Dice Score

Overlap	TP	FN	FP	Per-study	Per-patient	Global			
						All	<10mm	10-20mm	>20mm
0.00	213480	108034	47015	0.711 (0.679, 0.743)	0.715 (0.674, 0.756)	0.734 (0.687, 0.780)	0.655 (0.617, 0.694)	0.724 (0.691, 0.756)	0.749 (0.689, 0.810)
0.25	207101	77683	43042	0.739 (0.709, 0.770)	0.740 (0.700, 0.780)	0.774 (0.751, 0.798)	0.685 (0.656, 0.713)	0.755 (0.735, 0.775)	0.798 (0.775, 0.822)
0.50	186801	52094	37261	0.794 (0.781, 0.808)	0.792 (0.778, 0.806)	0.807 (0.786, 0.828)	0.766 (0.746, 0.785)	0.785 (0.768, 0.802)	0.824 (0.801, 0.847)

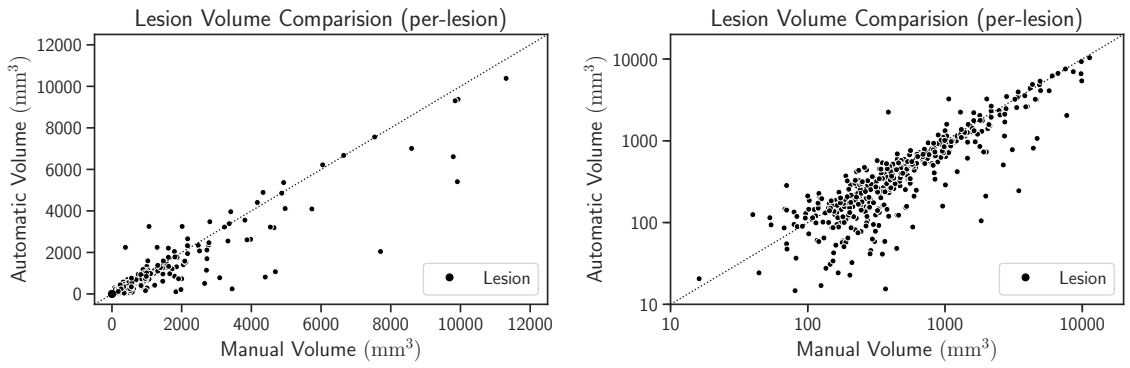
Table 4.2: **Detection Result on Liver Lesions.** Sensitivity and PPV are calculated based on lesion *objects* which consist of a connected volume of voxels. Dice score table shows the voxel counts for these respective objects. Different overlap values represent the minimum intersection-over-union (IoU) threshold for detections so that larger IoUs are stricter criteria. TP/FN/FP are true positive, false negative, and false positive. Note there are no true negatives (TNs). In per-study metrics, the metrics are first calculated per-study then averaged across all images; in per-patient, the TP/FN/FP counts are aggregated per-patient, and then the metrics are calculated before being averaged across. Finally, in the global case, the counts are aggregated globally and a single metric value is then calculated. 95% confidence interval is marked in parentheses.

Sensitivity					
Overlap	**TP**	**FN**	**FP**	**TN**	**Per-segment**
0.00	160	28	13	141	0.851 (0.773, 0.930)
0.25	139	49	13	141	0.739 (0.654, 0.825)
0.50	97	91	13	141	0.516 (0.344, 0.688)
Positive Predictive Value					
Overlap	**TP**	**FN**	**FP**	**TN**	**Per-segment**
0.00	160	28	13	141	0.925 (0.870, 0.980)
0.25	139	49	13	141	0.914 (0.850, 0.979)
0.50	97	91	13	141	0.882 (0.790, 0.974)
Specificity					
Overlap	**TP**	**FN**	**FP**	**TN**	**Per-segment**
0.00	160	28	13	141	0.916 (0.871, 0.961)
0.25	139	49	13	141	0.916 (0.870, 0.961)
0.50	97	91	13	141	0.916 (0.869, 0.962)

Table 4.3: **Detection Result on Liver Lesions (Cuoinaud Segment).** All metrics are derived *per-segment*: a Couinaud segment with at least one lesion is categorized as positive. See Table 4.2 for definition for overlap. 95% confidence interval is calculated according to [3] and is marked in parentheses.

columns. For example, sensitivity is only 0.487 for lesions less than 10 mm in diameter but rises to 0.938 for lesions larger than 20 mm in diameter. This phenomenon is even clearer with PPV where it becomes 1.00 for large lesions. Since overlap controls the level above which we recognize a prediction to be a valid detection (i.e., it is sufficiently close to the actual lesion labeled by the radiologist), the larger it is the tougher the detection problem is and hence sensitivity and PPV are expected to be lower with increasing overlap requirements. Dice score in the context of detection is different from that of Section 4.2 since only voxels counted towards a valid detection. Hence Dice score in fact increases when increasing the overlap threshold. With that said, it is still apparent that larger lesions have higher Dice scores in general.

I then attribute the lesions to their Couinaud liver segment and the results are shown in Table 4.3. In this experiment, there are 188 positive and 154 negative liver segments in the ground truth, among which the model detects 160 to be true positives.



(a) Linear Scale

(b) Log Scale

Figure 4-3: **Lesion Volume Comparison per Lesion.** Volumes for individual lesions are plotted, detector-generated against radiologist-annotated. Two scales are shown here due to the wide span of lesion volumes. Intraclass correlation coefficient is 0.958 (95% CI: 0.950, 0.965).

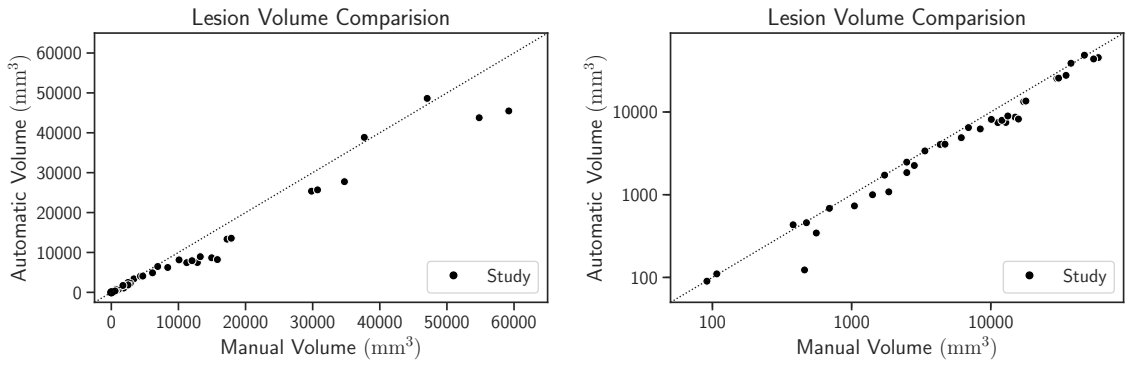
This is a sensitivity of 0.851 (95% CI: 0.773, 0.930), a PPV of 0.925 (95% CI: 0.870, 0.980), and a specificity of 0.916 (95% CI: 0.871, 0.961).

4.3 Interval Change Assessment

For evaluation of the overall lesion detection pipeline including the segmentation model and the longitudinal lesion correspondence model, I continue to use the longitudinal MRI dataset as it provides paired studies that enable comparison across time.

4.3.1 Liver Tumor Burden

To quantify changes of lesion volumes across different studies of a single patient in time, I am first interested in how the predicted volumes correlate with the annotated volumes. See Figures 4-3 and 4-4, where each point represents a lesion and a study respectively. The volumes from the detector show high concordance with the ground truth, yielding an intraclass correlation coefficient (ICC) [98] of 0.958 (95% CI: 0.950, 0.965) in the per-lesion case, and 0.962 (95% CI: 0.927, 0.980) when we sum lesion volumes per study. Note that since there lacks a correspondence between the ground



(a) Linear Scale

(b) Log Scale

Figure 4-4: **Lesion Volume Comparison per Study.** Volumes for lesions in 36 individual studies are plotted, detector-generated against radiologist-annotated. Two scales are shown here due to the wide span of lesion volumes. Intraclass correlation coefficient is 0.962 (95% CI: 0.927, 0.980)

truth lesions and the predictions, we use the same scheme as described in Section 3.4.1 to find pseudo-correspondences.

4.3.2 Longitudinal Assessment

Following techniques described in Section 3.3, we obtain correspondences of lesion across different studies in time.

There are 19 patients in total in the original test set, but after a closer inspection from the radiologist, one patient went through liver resection and is not suitable for evaluation because resection cases do not exist in training data. See Figure 4-5. In comparing the assessment of tumor burden changes across consecutive studies, 17 out of 18 total examined cases show agreement, corresponding to a Kohen's kappa of 0.909 (95% CI: 0.736, 1.000).

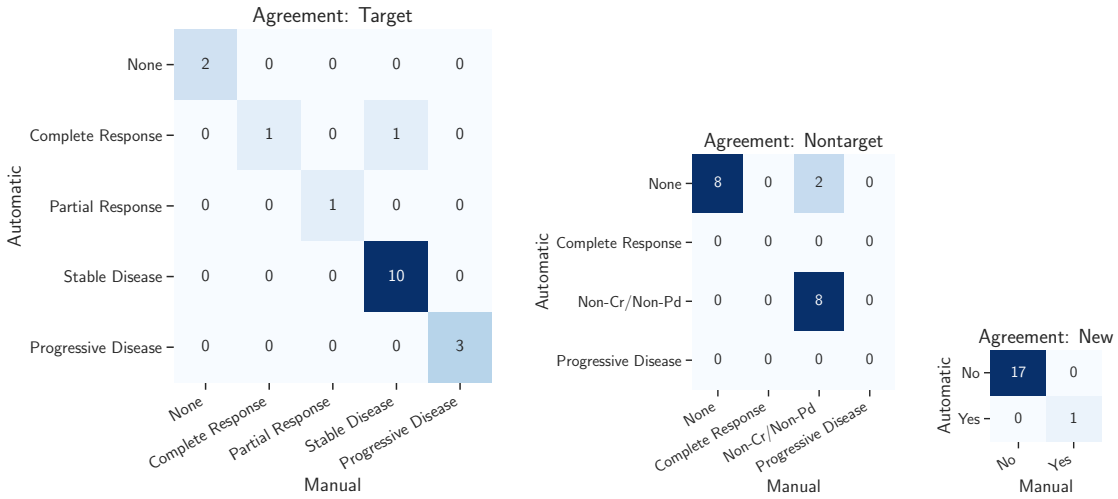
Furthermore, besides a standard evaluation using RECIST criteria, which is designed for efficient clinical evaluation with reasonable fidelity, now that we have an automated system capable of giving volumetric assessment within minutes, I am also interested in how volume-based evaluation compares.

From Figure 4-6, I plot the fractional volumetric changes in the lesions from baseline studies to the follow-up against radiologist-annotated RECIST. The more to the

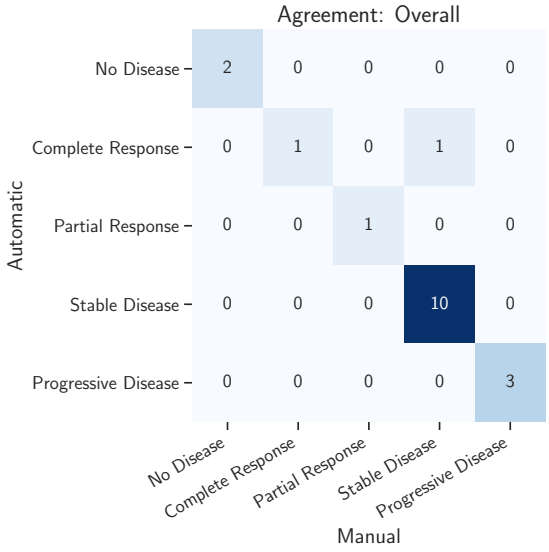
right, the more clinically problematic the cases are. There is a clear trend that with an increase in clinical severity of the interval changes, the higher the fractional changes are in the lesion volumes, regardless of whether we focus on Couinaud segments or entire studies. An ANOVA (analysis of variance) analysis shows $p = 0.04$ in the per-segment case and $p = 0.01$ in the per-study case, which suggests a fair amount of significant difference between the categories.

However, as we can also identify in the *Stable Disease* and *Progressive Disease* columns, there is not a hard cutoff distinguishing the two clinically different categories. The evidence suggests that I also look at another dimension that is also factored in while deriving the RECIST result: the absolute sizes of lesions.

In Figure 4-7, the y-axis remains similar and the RECIST responses are still color-coded. The x-axis now shows the absolute values of the volumes inferred by the automated system. It is interesting to see that it is required not only to have an increase of tumor burden over a certain threshold fraction but to have larger lesions to begin with, in order to be diagnosed as progressive. Though I do not plan to concretize the thresholds on volumetric analysis, yet the chart shows a remote connection with the actual RECIST criteria, which is in fact simply based on the 2-dimensional analysis. More thorough research effort will have to be put into annotating and relating the RECIST criteria and volumetric studies to accurately define the borders according to which we can be confident to assert diagnosis.

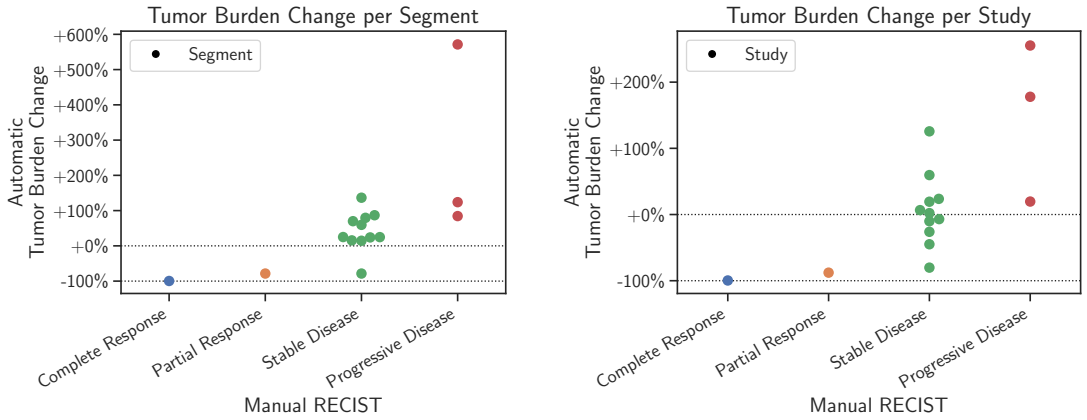


(a) Target Lesions (b) Non-target Lesions (c) New



(d) Overall Evaluation

Figure 4-5: **RECIST Evaluation against Radiologists.** Top row shows three aspects of how RECIST criteria is evaluated on 18 test cases for both radiologists and the automated method.



(a) Per-segment (b) Per-study

Figure 4-6: **Tumor Burden Change vs. RECIST.** On the y-axis is the volume change in tumor burden as evaluated by the automatic method while on the x-axis is the RECIST evaluated by radiologists. In (a), each dot is a Couinaud segment in the baseline studies that contain lesions; in (b), each dot is for a study. Dashed lines indicate key levels.

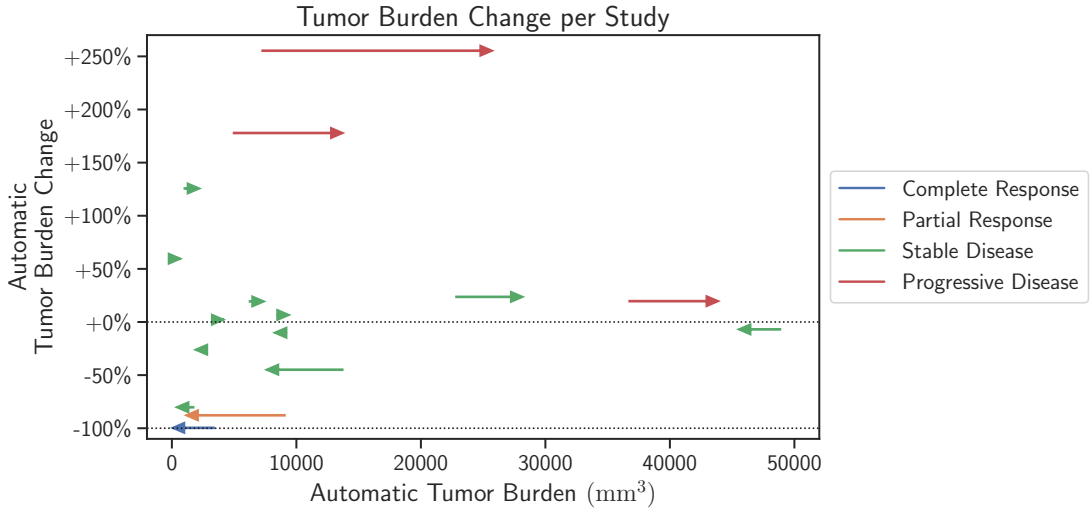


Figure 4-7: **Tumor Burden Change. Fractional vs. Absolute.** Each arrow denotes a case, pointing from the volume of the baseline study to that of the follow-up on the x-axis. Along the y-axis, the fractional change of the tumor burden is shown. Radiologist-annotated RECIST evaluation is color-coded.

Chapter 5

Conclusion

Longitudinal assessment of tumor burden from the liver is a clinically important task but suffers from both inconsistency and time consumption. Guidelines such as RECIST aim to relieve these problems for human evaluators, but these need not apply for machine-based methods as they can reproducibly produce tedious assessments much faster than human experts. Hence I have set the goals in this work to verify the feasibility of such a system, compare how concordant the system is with radiologists on existing evaluation guidelines, and what additional value we can obtain from it.

In all, I present an abdominal MRI dataset, on which I develop a workflow that detects liver lesions and subsequently compares lesions from two studies of the same patient in time to determine the tumor responses to the treatment during this period. As far as I know, this is the first work to develop an algorithm to perform the task in an automated way.

In the benchmarking process of the framework, I have some key findings regarding both the model selection, how the outcomes of this system correlate with existing guidelines, and how it provides insights beyond the guideline. I summarize these as follows:

- Do perform an extensive search over the parameters used in the CNN model. Factors like augmentation, complexity of the network, and loss function can have an apparent impact on the model performance. For example, in lesion

detection when the lesions are typically sparse, use generalized Dice loss, which proves to be stable.

- The lesion detection results, on a per-study basis, match closely to the annotation of experts on a cohort of 18 test subjects.
- On the ability to follow the guidelines that define tumor responses, the system agrees with expert annotation on 17 out of 18 patients, showing the potential to augment clinician decisions routinely.
- The system is able to produce volumetric tumor burden assessment that is not feasible to perform manually outside of small exploratory research trials.

There are very promising signs from this work on the possibility of automatic systems aiding radiologists. With that said, the sample size of this work can be a shortcoming, but I am fairly certain that researchers in this field will come up with larger datasets and clinical trials of greater scale that further verify the efficacy and usability of such automatic systems, both in terms of guidelines that are currently in use and more complex volumetric criteria that would eventually benefit clinicians in their everyday assessments.

Bibliography

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Keith Rust. Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1(4):381, 1985.

[4] Robert H El-Maraghi and Elizabeth A Eisenhauer. Review of phase ii trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase iii. *Journal of clinical oncology*, 26(8):1346–1354, 2008.

[5] AB Miller, BFAU Hoogstraten, MFAU Staquet, and A Winkler. Reporting results of cancer treatment. *cancer*, 47(1):207–214, 1981.

[6] Joseph Baar and Ian Tannock. Analyzing the same data in two ways: a demonstration model to illustrate the reporting and misreporting of clinical trials. *Journal of Clinical Oncology*, 7(7):969–978, 1989.

[7] Yoshiaki Tsuchida and Patrick Therasse. Response evaluation criteria in solid tumors (recist): new guidelines. *Medical and Pediatric Oncology: The Official Journal of SIOP—International Society of Pediatric Oncology (Societé Internationale d'Oncologie Pédiatrique*, 37(1):1–3, 2001.

[8] Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, D Sargent, Robert Ford, Janet Dancey, S Arbuck, Steve Gwyther, Margaret Mooney, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

[9] Julius Chapiro, MingDe Lin, Rafael Duran, Rüdiger E Schernthaner, and Jean-François Geschwind. Assessing tumor response after loco-regional liver cancer therapies: the role of 3d mri. *Expert review of anticancer therapy*, 15(2):199–205, 2015.

[10] Julius Chapiro, Rafael Duran, MingDe Lin, Rüdiger E Schernthaner, Zhijun Wang, Boris Gorodetski, and Jean-François Geschwind. Identifying staging markers for hepatocellular carcinoma before transarterial chemoembolization: comparison of three-dimensional quantitative versus non–three-dimensional imaging markers. *Radiology*, 275(2):438–447, 2015.

[11] E Eisenhauer, P Therasse, J Bogaerts, L Schwartz, D Sargent, R Ford, J Dancey, S Arbuck, S Gwyther, M Mooney, et al. 32 invited new response evaluation criteria in solid tumors: revised recist guideline version 1.1. *Ejc Supplements*, 12(6):13, 2008.

[12] Richard G Abramson, Carrie R McGhee, Nikita Lakomkin, and Carlos L Arteaga. Pitfalls in recist data extraction for clinical trials: beyond the basics. *Academic radiology*, 22(6):779–786, 2015.

[13] Mehrdad Moghbel, Syamsiah Mashohor, Rozi Mahmud, and M Iqbal Bin Saripan. Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography. *Artificial Intelligence Review*, 50(4):497–537, 2018.

[14] Gabriel Chartrand, Phillip M Cheng, Eugene Vorontsov, Michal Drozdzal, Simon Turcotte, Christopher J Pal, Samuel Kadoury, and An Tang. Deep learning: a primer for radiologists. *Radiographics*, 37(7):2113–2131, 2017.

[15] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

[16] Luc Soler, Hervé Delingette, Grégoire Malandain, Johan Montagnat, Nicholas Ayache, Christophe Koehl, Olivier Dourthe, Benoit Malassagne, Michelle Smith, Didier Mutter, et al. Fully automatic anatomical, pathological, and functional segmentation from ct scans for hepatic surgery. *Computer Aided Surgery*, 6(3):131–142, 2001.

[17] Kyung-Sik Seo and Jong-An Park. Improved automatic liver segmentation of a contrast enhanced ct image. In *Pacific-Rim Conference on Multimedia*, pages 899–909. Springer, 2005.

[18] Seung-Jin Park, Kyung-Sik Seo, and Jong-An Park. Automatic hepatic tumor segmentation using statistical optimal threshold. In *International Conference on Computational Science*, pages 934–940. Springer, 2005.

[19] Hanung Adi Nugroho, Dani Ihtatho, and Hermawan Nugroho. Contrast enhancement for liver tumor identification. In *MICCAI workshop*, volume 41, page 201, 2008.

[20] Anirudh Choudhary, Nicola Moretto, Francesca Pizzorni Ferrarese, and Giulia A Zamboni. An entropy based multi-thresholding method for semi-automatic segmentation of liver tumors. In *MICCAI workshop*, volume 41, pages 43–49, 2008.

[21] Marcin Ciecholewski and Marek R Ogiela. Automatic segmentation of single and multiple neoplastic hepatic lesions in ct images. In *International work-conference on the interplay between natural and artificial computation*, pages 63–71. Springer, 2007.

[22] Nader H Abdel-massieh, Mohiy M Hadhoud, and Khalid M Amin. Fully automatic liver tumor segmentation from abdominal ct scans. In *The 2010 International Conference on Computer Engineering & Systems*, pages 197–202. IEEE, 2010.

[23] Yingyi Qi, Wei Xiong, Wee Keng Leow, Qi Tian, Jiayin Zhou, Jiang Liu, Thazin Han, Sudhakar K Venkatesh, and Shih-chang Wang. Semi-automatic segmentation of liver tumors from ct scans using bayesian rule-based 3d region growing. In *MICCAI workshop*, volume 41, page 201, 2008.

[24] Itay Ben-Dan and Elior Shenhav. Liver tumor segmentation in ct images using probabilistic methods. In *MICCAI Workshop*, volume 41, page 43, 2008.

[25] Laura Fernández-de Manuel, José L Rubio, Maria J Ledesma-Carbayo, Javier Pascau, Jose M Tellado, Enrique Ramón, Manuel Desco, and Andrés Santos. 3d liver segmentation in preoperative ct images using a levelsets active surface method. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3625–3628. IEEE, 2009.

[26] Damon Wong, Jiang Liu, Yin Fengshou, Qi Tian, Wei Xiong, Jiayin Zhou, Yingyi Qi, Thazin Han, S Venkatesh, and Shih-chang Wang. A semi-automated method for liver tumor segmentation based on 2d region growing with knowledge-based constraints. In *MICCAI workshop*, volume 41, page 159, 2008.

[27] Marius George Linguraru, William J Richbourg, Jeremy M Watt, Vivek Pamulapati, and Ronald M Summers. Liver and tumor segmentation and analysis from ct of diseased patients via a generic affine invariant shape parameterization and graph cuts. In *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*, pages 198–206. Springer, 2011.

[28] Yoav Taieb, Ofer Eliassaf, Moti Freiman, Leo Joskowicz, and Jacob Sosna. An iterative bayesian approach for liver analysis: tumors validation study. In *MICCAI workshop*, volume 41, page 43, 2008.

[29] Laurent Massoptier and Sergio Casciaro. A new fully automatic and robust algorithm for fast segmentation of liver tissue and tumors from ct scans. *European radiology*, 18(8):1658, 2008.

[30] Yrjö Häme. Liver tumor segmentation using implicit surface evolution. *The Midas Journal*, pages 1–10, 2008.

[31] Yrjö Häme and Mika Pollari. Semi-automatic liver tumor segmentation with hidden markov measure field model and non-parametric distribution estimation. *Medical image analysis*, 16(1):140–149, 2012.

[32] Jiayin Zhou, Wei Xiong, Qi Tian, Yingyi Qi, Jiang Liu, Wee Keng Leow, Thazin Han, Sudhakar K Venkatesh, and Shih-chang Wang. Semi-automatic segmentation of 3d liver tumors from ct scans using voxel classification and propagational learning. In *MICCAI workshop*, volume 41, page 43, 2008.

[33] Tarak Ben Saïd, O Azaiz, Faten Chaieb, Slim M'hiri, Faouzi Ghorbel, and Olfa Azaiz. Segmentation of liver tumor using hmrf-em algorithm with bootstrap resampling. In *2010 5th International Symposium On I/V Communications and Mobile Network*, pages 1–4. IEEE, 2010.

[34] Xing Zhang, Jie Tian, Dehui Xiang, Xiuli Li, and Kexin Deng. Interactive liver tumor segmentation from ct scans using support vector classification with watershed. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6005–6008. IEEE, 2011.

[35] Yuanzhong Li, Shoji Hara, and Kazuo Shimura. A machine learning approach for locating boundaries of liver tumors in ct images. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 400–403. IEEE, 2006.

[36] Jingran Wen, Xiaoyan Zhang, Ye Xu, Zuofeng Li, and Lei Liu. Comparison of adaboost and logistic regression for detecting colorectal cancer patients with synchronous liver metastasis. In *2009 International Conference on Biomedical and Pharmaceutical Engineering*, pages 1–6. IEEE, 2009.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[39] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2016.

[40] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[41] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[42] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[43] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[44] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[45] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[46] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[47] Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Deepmedic for brain tumor segmentation. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pages 138–149. Springer, 2016.

[48] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[49] Wei Chen, Boqiang Liu, Suting Peng, Jiawei Sun, and Xu Qiao. S3d-unet: separable 3d u-net for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 358–368. Springer, 2018.

[50] Baris Kayalibay, Grady Jensen, and Patrick van der Smagt. Cnn-based segmentation of medical imaging data. *arXiv preprint arXiv:1701.03056*, 2017.

[51] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In *International MICCAI Brainlesion Workshop*, pages 287–297. Springer, 2017.

[52] Suting Peng, Wei Chen, Jiawei Sun, and Boqiang Liu. Multi-scale 3d u-nets: An approach to automatic segmentation of brain tumor. *International Journal of Imaging Systems and Technology*, 2019.

[53] Qi Dou, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin, and Pheng-Ann Heng. 3d deeply supervised network for automated segmentation of volumetric medical images. *Medical image analysis*, 41:40–54, 2017.

[54] Zhuotun Zhu, Yingda Xia, Wei Shen, Elliot K Fishman, and Alan L Yuille. A 3d coarse-to-fine framework for automatic pancreas segmentation. *arXiv preprint arXiv:1712.00201*, 2, 2017.

[55] Chunfeng Lian, Mingxia Liu, Jun Zhang, Xiaopeng Zong, Weili Lin, and Dinggang Shen. Automatic segmentation of 3d perivascular spaces in 7t mr images using multi-channel fully convolutional network. In *Proceedings of the International Society for Magnetic Resonance in Medicine... Scientific Meeting and Exhibition. International Society for Magnetic Resonance in Medicine. Scientific Meeting and Exhibition*, volume 2018. NIH Public Access, 2018.

[56] Holger R Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. An application of cascaded 3d fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99, 2018.

[57] Xin Yang, Cheng Bian, Lequan Yu, Dong Ni, and Pheng-Ann Heng. Hybrid loss guided convolutional networks for whole heart parsing. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 215–223. Springer, 2017.

[58] Lequan Yu, Xin Yang, Jing Qin, and Pheng-Ann Heng. 3d fractalnet: dense volumetric segmentation for cardiovascular mri volumes. In *Reconstruction, segmentation, and analysis of medical images*, pages 103–110. Springer, 2016.

[59] Lei Bi, Jinman Kim, Ashnil Kumar, and Dagan Feng. Automatic liver lesion detection using cascaded deep residual networks. *arXiv preprint arXiv:1704.02703*, 2017.

[60] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016.

[61] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[62] Bram Van Ginneken, Tobias Heimann, and Martin Styner. 3d segmentation in the clinic: A grand challenge. In *MICCAI Workshop on 3D Segmentation in the Clinic: A Grand Challenge*, volume 1, pages 7–15, 2007.

[63] IRCAD. 3d image reconstruction for comparison of algorithm database.

[64] TCIA. The cancer genome atlas liver hepatocellular carcinoma.

[65] Ivana Despotović, Bart Goossens, and Wilfried Philips. Mri segmentation of the human brain: challenges, methods, and applications. *Computational and mathematical methods in medicine*, 2015, 2015.

[66] Christopher M Collins, Wanzhan Liu, Weston Schreiber, Qing X Yang, and Michael B Smith. Central brightening due to constructive interference with, without, and despite dielectric resonance. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 21(2):192–196, 2005.

[67] John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.

[68] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310, 2010.

[69] Denis P Shamonin, Esther E Bron, Boudewijn PF Lelieveldt, Marion Smits, Stefan Klein, and Marius Staring. Fast parallel image registration on cpu and gpu for diagnostic classification of alzheimer's disease. *Frontiers in neuroinformatics*, 7:50, 2014.

[70] Thomas W Sederberg and Scott R Parry. Free-form deformation of solid geometric models. *ACM SIGGRAPH computer graphics*, 20(4):151–160, 1986.

[71] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.

[72] Dinggang Shen and Christos Davatzikos. Hammer: hierarchical attribute matching mechanism for elastic registration. *IEEE transactions on medical imaging*, 21(11):1421–1439, 2002.

[73] Carlos Oscar Sánchez Sorzano, Philippe Thévenaz, and Michael Unser. Elastic registration of biological images using vector-spline regularization. *IEEE Transactions on Biomedical Engineering*, 52(4):652–663, 2005.

[74] Andrew Simper. Correcting general band-to-band misregistrations. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 2, pages 597–600. IEEE, 1996.

[75] Rikard Berthilsson. Affine correlation. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, volume 2, pages 1458–1460. IEEE, 1998.

[76] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.

[77] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29(1):196–205, 2009.

[78] Jean-Marie Guyader, Livia Bernardin, Naomi HM Douglas, Dirk HJ Poot, Wiro J Niessen, and Stefan Klein. Influence of image registration on apparent diffusion coefficient images computed from free-breathing diffusion mr images of the abdomen. *Journal of Magnetic Resonance Imaging*, 42(2):315–330, 2015.

[79] Kamal Sahi, Stuart Jackson, Edward Wiebe, Gavin Armstrong, Sean Winters, Ronald Moore, and Gavin Low. The value of "liver windows" settings in the detection of small renal cell carcinomas on unenhanced computed tomography. *Canadian Association of Radiologists Journal*, 65(1):71–76, 2014.

[80] Xiaofei Sun, Lin Shi, Yishan Luo, Wei Yang, Hongpeng Li, Peipeng Liang, Kuncheng Li, Vincent CT Mok, Winnie CW Chu, and Defeng Wang. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *Biomedical engineering online*, 14(1):73, 2015.

[81] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[82] Eugene Vorontsov, An Tang, Chris Pal, and Samuel Kadoury. Liver lesion segmentation informed by joint liver segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1332–1335. IEEE, 2018.

[83] William R Crum, Oscar Camara, and Derek LG Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging*, 25(11):1451–1461, 2006.

[84] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

[85] Zeshan Hussain, Francisco Gimenez, Darvin Yi, and Daniel Rubin. Differential data augmentation techniques for medical imaging classification tasks. In *AMIA Annual Symposium Proceedings*, volume 2017, page 979. American Medical Informatics Association, 2017.

[86] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[87] Eduardo Castro, Jaime S Cardoso, and Jose Costa Pereira. Elastic deformations for data augmentation in breast cancer mass detection. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 230–234. IEEE, 2018.

[88] Chao Huang, Hu Han, Qingsong Yao, Shankuan Zhu, and S Kevin Zhou. 3d u2-net: A 3d universal u-net for multi-domain medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 291–299. Springer, 2019.

[89] World Health Organization et al. Who handbook for reporting results of cancer treatment. 1979.

[90] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992.

[91] Zhengyou Zhang. Iterative point matching for registration of free-form curves and surfaces. *International journal of computer vision*, 13(2):119–152, 1994.

[92] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

[93] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[94] Eugene Vorontsov, Milena Cerny, Philippe Régnier, Lisa Di Jorio, Christopher J Pal, Réal Lapointe, Franck Vandenbroucke-Menu, Simon Turcotte, Samuel Kadoury, and An Tang. Deep learning for automated segmentation of liver lesions at ct in patients with colorectal cancer liver metastases. *Radiology: Artificial Intelligence*, 1(2):180014, 2019.

[95] Michael Schünke, Erik Schulte, Udo Schumacher, Lawrence M Ross, and Edward D Lamperti. *Thieme atlas of anatomy: Neck and internal organs*. Thieme, 2006.

[96] Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu.

3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 851–858. Springer, 2018.

[97] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[98] John J Bartko. The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1):3–11, 1966.