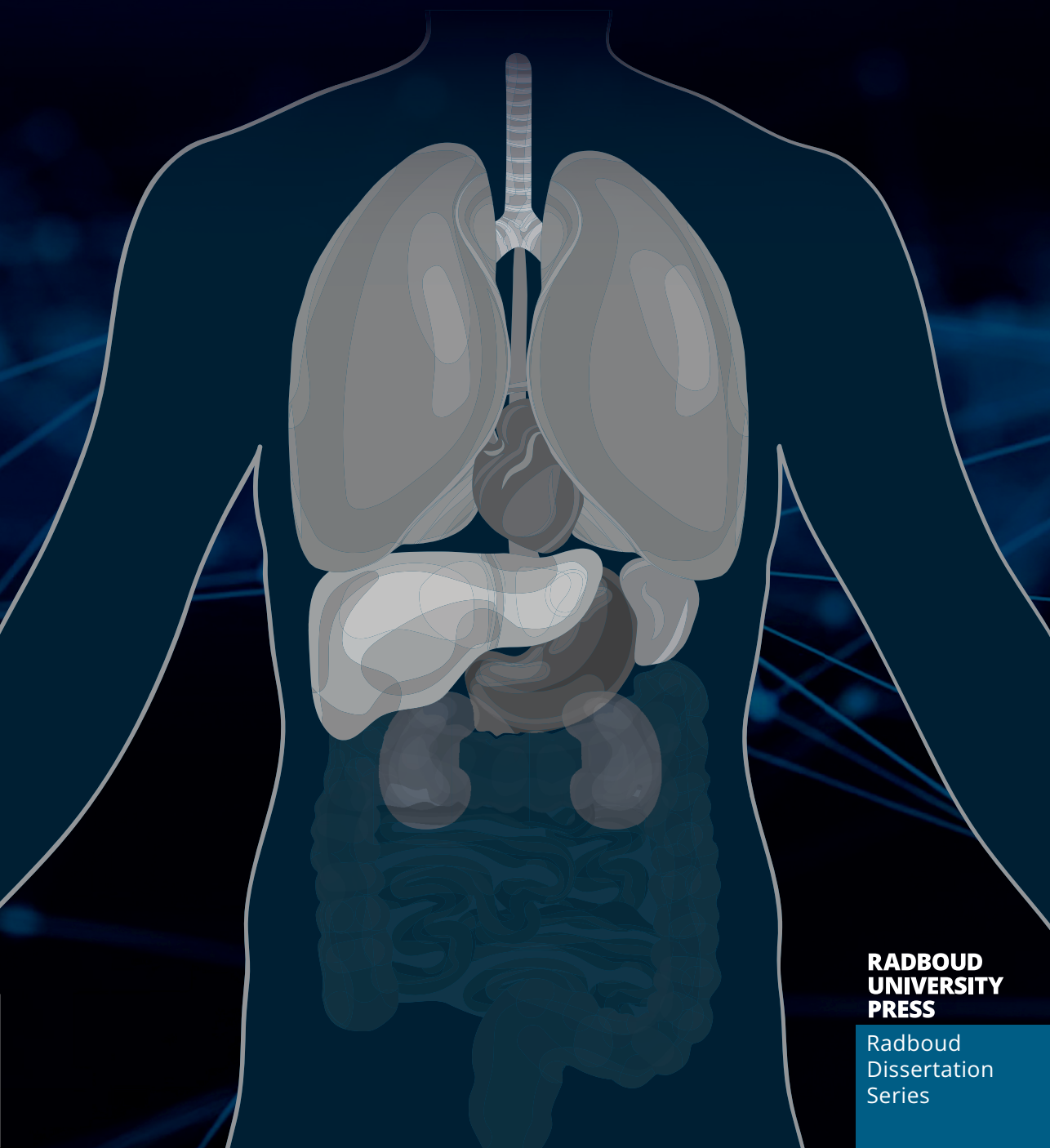
Deep Learning for Localization and Segmentation in Thorax Abdomen CT

Gabriel Efrain Humpire Mamani



RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Deep Learning for Localization and Segmentation in Thorax Abdomen CT

Gabriel Efrain Humpire Mamani

This book was typeset by the author using LaTeX2$_\varepsilon$.

Author : Gabriel Efrain Humpire Mamani
Title : Deep Learning for Localization and Segmentation in Thorax Abdomen CT

Radboud Dissertations Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Cover : Proefschrift AIO | Guntra Laivacuma
Printing : DPN Rikken/Pumbo
ISBN : 9789493296596
DOI : 10.54195/9789493296596

Free download at: www.boekenbestellen.nl/radboud-university-press/dissertations

RADBOUD
UNIVERSITY
PRESS

Deep Learning for Localization and Segmentation in Thorax Abdomen CT

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 1 juli 2024
om 10.30 uur precies

door

Gabriel Efrain Humpire Mamani

geboren op 3 september 1987
te Arequipa (Peru)

Promotoren:	**Prof. dr. B. van Ginneken**
	Prof. dr. W.M. Prokop
Copromotoren:	**Dr. ir. C. Jacobs**
	Dr. N. Lessmann
Manuscriptcommissie:	**Prof. dr. I.D. Nagtegaal**
	Prof. dr. A.L.A.J. Dekker (Maastricht University)
	Dr. T.C. Kwee (Universitair Medische Centrum Groningen)

The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Medical Center (Nijmegen, the Netherlands).
This work was funded by Fraunhofer ICON project Automation in Medical Imaging project.

TABLE OF CONTENTS

Introduction

1

Cancer is the second leading cause of death worldwide, after heart disease. Every year, ten million people die of cancer[1]. Medical imaging, particularly computed tomography (CT) scans, is crucial in detecting tumors and determining how to best treat the disease. However, the large amount of CT scans can be overwhelming for radiologists to analyze and report on. In the Netherlands alone, two million CT scans were made in 2020, a doubling since 2010[2]. In this thesis, we aimed to develop and validate computer methods to assist radiologists in analyzing CT scans of cancer patients, with the hope of reducing the time it takes them to report on these scans and the potential to generate more precise and quantifiable measures of disease. The technique that is used throughout this thesis is deep learning. We focus on methods to locate and segment structures, both organs, such as the spleen and the kidneys, and abnormalities, such as tumors or cysts.

This introductory chapter will provide a brief background on cancer imaging, computed tomography, deep learning, and the evaluation metrics used throughout the thesis. The chapter ends with a short outline of the thesis.

1.1 Cancer

Cancer is characterized by the uncontrolled growth and proliferation of abnormal cells in the body. Normally, old cells die, yet in cancer, these cells persist, growing without control, and mutating into abnormal cells. These abnormal cells may form a mass of tissue known as a tumor. When detected early, there are better treatment options for most types of cancers, therefore early detection is crucial to reduce the mortality rate. Over a span of 25 years, from 1991 to 2016, a combination of cancer treatments and early detection efforts, supported by clinical trials providing evidence of treatment efficacy, managed to reduce cancer deaths by 27%, as shown by a study by Siegel et al. in 2019[3]. This means that the number of cancer-related deaths per 100,000 people per year dropped from 215 to 156[3].

1.1.1 Cancer treatments

The treatment of cancer can vary widely depending on the type, stage, and location of the cancer, as well as the individual patient's health and preferences. However, some of the most important and commonly used treatments for cancer include surgery, chemotherapy, radiation therapy and immunotherapy. Surgical removal of cancerous tissue is a primary treatment for many types of cancer, particularly if the tumor is localized and has not spread to other parts of the body. Chemotherapy involves the use of drugs to kill or inhibit the growth of cancer cells. It can be ad-

ministered orally or through intravenous injections and is often used when cancer has spread to other parts of the body. Radiation therapy is a treatment using high-energy radiation beams to target and destroy cancer cells or shrink tumors. It can be used alone or in combination with other treatments. Immunotherapy drugs help the immune system recognize and attack cancer cells. They have shown great promise in treating various types of cancer[4].

1.1.2 Imaging of cancer

X-rays, CT scans, and advanced imaging techniques such as MRI and PET scans play a crucial role in the ongoing monitoring of cancer progression in patients. Among these, CT scans are the most widely used. This prominence arises from their ability to offer greater precision when compared to conventional 2D radiographs, lower cost and better availability when compared with MRI, and their less invasive nature and lower costs in contrast to PET scans.

In the management of cancer, CT scans have many applications. Beyond their diagnostic utility, CT scans are instrumental in guiding medical decisions. They also aid physicians during the process of taking biopsies, where tissue samples are collected for further analysis. Additionally, CT scans are essential in the meticulous planning of chemotherapy and radiation therapy regimens, ensuring that treatment is precisely targeted to the affected area while sparing healthy tissue.

Moreover, information from CT scans allows medical professionals to quantitatively assess change in the volume of organs and tumors over time. This longitudinal data is of paramount importance, serving as a yardstick to determine the effectiveness of ongoing cancer treatments. It provides insights into whether the treatment is achieving the desired results, and based on information from the CT the medical team can decide to adjust to the therapeutic approach, if necessary. Furthermore, CT scans are essential in identifying scenarios where palliative care may be the most appropriate course of action, thus promoting a holistic approach to cancer care that prioritizes the patient's well-being and quality of life.

1.2 Computed Tomography

X-rays are electromagnetic radiation first discovered by Wilhelm Conrad Röntgen in 1895[5]. A modern CT scanner consists of a source emitting x-rays and, on the opposite side, an array of detectors measuring the radiation that has passed through the patient. Both source and detector rotate around the patient, while the patient is lying on a table that is moved through the rotating ring of source and detector. From the

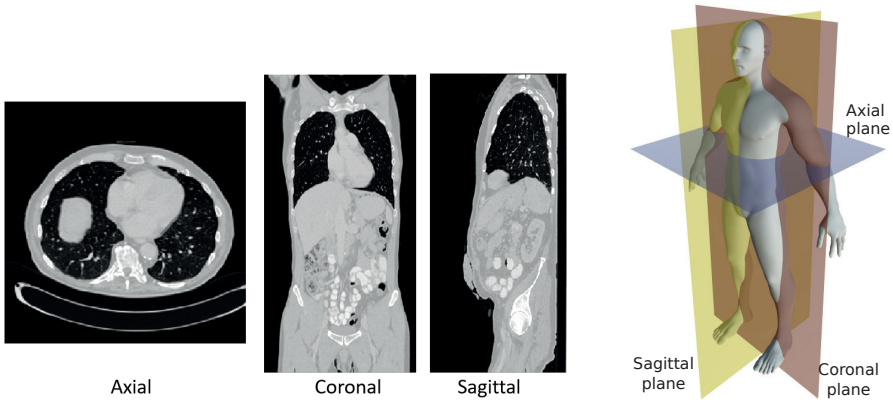


Figure 1.1: Orthogonal views

detector measurements it is possible to create 2D slices, cross-sectional images, of the patient with pixel values indicating how much of the x-ray radiation was absorbed by the tissue. These values are measured in Hounsfield Units (HU). Soft tissues, such as muscle, absorb fewer x-rays than hard tissues, such as bone. Common HU values are air -1000 HU, water 0 HU, fat -120 HU, muscles 40 HU, and bones 300 HU onward. As a result, radiologists may miss lesions if the lesions have a similar density as the directly surrounding tissue.

Intravenous (IV) and oral contrast agents can improve the visibility of certain structures within the body, but they can also cause allergic reactions. Intravascular contrast agents are often used during CT scans to help detect solid organ metastases, such as the liver, the adrenal glands, and the brain. However, the use of contrast media can also alter tissue attenuation and may cause artifacts. Artifacts can also be created by materials with a high atomic number, such as hip and vertebrae prostheses.

CT scans are composed of a sequence of 2D slices that are reconstructed into a 3D image, which can be viewed in multiple orthogonal views (axial, coronal, and sagittal, see Figure 1.1).

Modern CT scanners offer a higher resolution and are able to reconstruct thinner slices. It is now possible to make scans where each voxel has a size of $0.5 \times 0.5 \times 0.5$ mm, meaning each cubic centimeter of a patient already contains 8000 measurement values. These high-resolution CT scans further increase the amount of data that must be processed, add to the workload of radiology departments, and increase reading times for radiologists.

1.3 Medical Image Analysis and Deep Learning

Medical image analysis is the field that focuses on analyzing visual medical conditions using digital computers. This analysis has been traditionally performed using computer vision techniques, in which the feature extraction is handcrafted, meaning that human experts decide on which features to compute from the image or a local part of the image. These features were further processed by the computer.

However, Deep Learning (DL), a subfield of machine learning, has recently become a popular approach for medical image analysis[6,7]. DL allows for the direct learning of meaningful representations from data rather than requiring handcrafted feature extraction. DL models typically consist of multiple convolutional layers stacked on top of each other, which are applied to the input image to optimize the medical imaging task. Using multiple layers allows for the learning of different features at each layer, and the combination of convolutional layers, pooling, and non-linear operations make DL a powerful tool for medical image analysis. Optimization algorithms guide the neural network during training to update the weights of the model and regularization techniques are used to try to ensure that the DL model generalizes well to unseen data. During training, the parameters in the network, the weights, are continuously slightly adjusted to produce more correct output for a given input. The process by which these adjustments are computed for each weight is called backpropagation.

One challenge in using DL, for medical image analysis particularly, is the limited amount of annotated data available. DL typically requires large amounts of data to achieve high performance, and the lack of large, annotated datasets can make it difficult for DL models to perform or generalize well. While there may be plenty of medical imaging data to analyze, the amount of data annotated is typically small, because annotating (indicating which part of an image or even which voxel has which label) is expensive[7]. Data augmentation, a technique that slightly modifies the training data to generate new training data samples, can help the model to generalize better.

Another challenge in DL is that typically for every new task, a new network has to be trained. However, it is likely that a network that has already been optimized for some task, for example, segmentation of the right kidney, could be partly 'reused' when a network is needed for a slightly different task, for example, segmentation of the left kidney. Normally, the weights of the network are randomly initialized. In the example above we could initialize the networks for segmenting the left kidney with the weights of the already trained network for segmenting the right kidney. This process is called transfer learning. There are many ways in which this can be done

and this is the topic of research in Chapter 5.

In this thesis, we address the tasks of localization of structures and of segmentation. We now define each of these tasks.

1.3.1 Localization

Given an input image, a localization algorithm identifies and locates specific regions of interest, such as the spleen in Figure 1.2b. These algorithms are often used as a pre-processing step for more complex tasks such as 3D segmentation. In 3D images such as CT scans, a localization algorithm may analyze each 2D slice in the CT scan to classify the presence or absence of a particular organ. The combination of these classifications across all the slices in an orthogonal view creates a 1D bounding box, which can be performed in the other two orthogonal views to create a 3D bounding box for the organ. This can be particularly useful in the early stages of cancer detection and treatment planning, as it allows radiologists to more efficiently and accurately measure the size and location of tumors and other abnormalities. Previous research has generally focused on creating separate networks for each organ; in Chapter 2, we propose a multi-label method that can localize eleven structures using a single network.

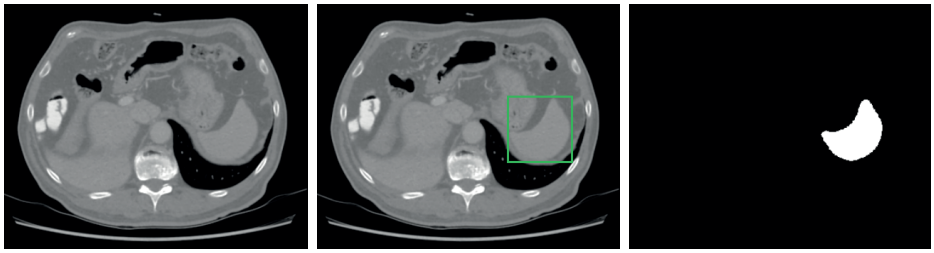
1.3.2 Segmentation

Given an input image, a segmentation algorithm automatically identifies and outlines organs of interest, such as the spleen in Figure 1.2c. Traditionally, semi-automatic and automatic methods based on computer vision techniques such as thresholding, watershed, random-walk, and level-sets were used for organ segmentation. However, the development of DL models such as 2D U-Net[8] and 3D U-Net[9], which utilize an encoder-decoder architecture, have significantly improved the accuracy and efficiency of organ segmentation.

In this thesis, we demonstrate the use of these techniques to segment various structures, including the spleen, kidney, and kidney abnormalities.

1.4 Evaluation metrics

This section outlines the metrics used throughout this thesis to evaluate the performance of the proposed methods. In the equations below, X represents the post-processed output of the network, and Y represents the reference standard. The function $SurfDist(A, B)$ measures the minimum distance from a voxel on surface A to a



(a) (b) (c)

Figure 1.2: (a) Example of input slice. (b) Shows the output of the spleen localization, (c) shows the output of the spleen segmentation.

voxel on surface B.

1.4.1 Wall distance

This metric measures the distance between two 2D slices that are in the same orthogonal direction. The organ localization method proposed in Chapter 2 uses this metric to measure the distance in millimeters between the reference standard and the predicted bounding boxes.

1.4.2 Dice score

The Dice score measures the overlap between two objects in proportion to the sum of their volumes. It is commonly used for segmentation tasks to compare the performance of different methods. High Dice scores (close to one) indicate a high level of overlap between the two objects. All the proposed methods in this thesis used the Dice score to report the results and compare performance.

$$\text{Dice score} = \frac{\text{Area of overlap}}{\text{Total area}} = \frac{2 * volume(X \cap Y)}{volume(X) + volume(Y)} = \frac{2*}{+} \tag{1.1}$$

1.4.3 Jaccard coefficient

Similar to the Dice score, the Jaccard coefficient measures the proportionality between the overlap and union of two objects. High Jaccard coefficients (close to one) indicate a high overlap between the two objects. This metric is also known as intersection over union (IoU) or Jaccard index. Chapter 2 of this thesis uses this metric to report results.

$$\text{Jaccard coefficient} = \frac{\text{Area of overlap}}{\text{Area of union}} = \frac{volume(X \cap Y)}{volume(X \cup Y)} = \quad (1.2)$$

1.4.4 Maximum Hausdorff distance

This metric returns the maximum surface distance between two objects. Equation 1.3 shows that the maximum Hausdorff distance represents the maximum value among the minimum distances from a voxel in Y to X. For this metric, values close to 0 are preferred. This metric was reported in Chapter 3.

$$\text{Max. Hausdorff dist.} = max(max(SurfDist(X,Y)), max(SurfDist(Y,X))) \quad (1.3)$$

1.4.5 95 percentile Hausdorff distance

This is a derived metric of the maximum Hausdorff distance. While the maximum Hausdorff distance measures the maximum minimum distance between two objects, the 95 percentile Hausdorff distance reports the 95 percentile of these minimum surface distances. Many recent segmentation papers report this metric. Similar to the maximum Hausdorff distance, values close to 0 are preferred. This metric was reported in Chapter 3.

$$\text{95\% Hausdorff dist.} = max(Percentile_{95}(SurfDist(X,Y)), Percentile_{95}(SurfDist(Y,X)) \quad (1.4)$$

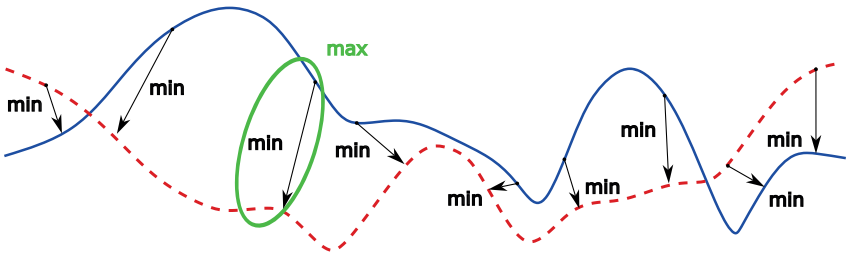


Figure 1.3: Representation of Hausdorff distance between two contours. The function SurfDist(X,Y) is represented by black arrows and returns the minimum distance from point X to point Y. Note that SurfDist(Y,X) may be different than SurfDist(X,Y). Maximum Hausdorff distance takes the maximum distance of the maximum distances between SurfDist(X,Y) and SurfDist(Y,X). Whereas 95% Hausdorff distance takes the maximum of the 95 percentile of the SurfDist(X,Y) and the 95 percentile of SurfDist(Y,X). Figure adapted from Kaspar [10].

1.5 Thesis outline

The work presented in this thesis is focused on using DL for detecting and segmenting structures in CT scans. In **Chapter 2**, we present a method for localizing organs in 2D orthogonal views; this method combines the outputs of each orthogonal view to compose a 3D bounding box per organ. In **Chapter 3**, we apply a state-of-the-art segmentation algorithm using Convolutional Neural Networks (CNN) to segment the spleen, achieving performance comparable to that of an independent observer. In an observer experiment, the radiologist rated the segmentation quality as 94% as ready for clinical use. Additionally, we performed an experiment to measure the splenic volume change over time. In **Chapter 4**, we segment the kidneys and kidney abnormalities, including cysts, lesions, masses, metastases, and tumors. We conducted an ablation study to analyze the performance of five components of the method. In **Chapter 5**, we explore the use of transfer learning to segment additional structures using a partially annotated dataset (a junction of publicly available datasets and data from public challenges). Finally, **Chapter 6**, provides the general discussion and summary of this thesis.

Multilabel structure localization

2

G.E. Humpire Mamani, A. Setio, B. van Ginneken and C. Jacobs

Original title: Efficient organ localization using multi-label ConvNets in thorax-abdomen CT scans

Published in: Physics in Medicine and Biology 63(8):085003, 2018

Abstract

Automatic localization of organs and other structures in medical images is an important preprocessing step that can improve and speed up other algorithms such as organ segmentation, lesion detection, and registration. This work presents an efficient method for simultaneous localization of multiple structures in 3D thorax-abdomen CT scans. Our approach predicts the location of multiple structures using a single multi-label convolutional neural network for each orthogonal view. Each network takes extra slices around the current slice as input to provide extra context. A sigmoid layer is used to perform multi-label classification. The output of the three networks is subsequently combined to compute a 3D bounding box for each structure. We used our approach to locate 11 structures of interest. The neural network was trained and evaluated on a large set of 1884 thorax-abdomen CT scans from patients undergoing oncological workup. Reference bounding boxes were annotated by human observers. The performance of our method was evaluated by computing the wall distance to the reference bounding boxes. The bounding boxes annotated by the first human observer were used as the reference standard for the test set.
Using the best configuration, we obtained an average wall distance of 3.20 ± 7.33mm in the test set. The second human observer achieved 1.23 ± 3.39mm. For all structures, the results were better than those reported in previously published studies. In conclusion, we proposed an efficient method for the accurate localization of multiple organs. Our method uses multiple slices as input to provide more context around the slice under analysis, and we have shown that this improves performance. This method can easily be adapted to handle more organs.

2.1 Introduction

An automatic organ localization is an important preprocessing step that can improve other medical image processing steps. A 3D organ localization method can be used to discard non-relevant areas of the scan for subsequent algorithms, e.g., full 3D organ segmentation. For instance, organ localization methods have been used with automatic organ segmentation methods[11–14] to improve the performance and to reduce the ratio of false positive segmentations.

Many algorithms have already been proposed for the localization of organs in CT scans[11–21]. We describe the most relevant literature, split into two parts: the first part summarizes classical machine learning approaches – involving hand-crafted features and a machine learning classifier – and the second part focuses on deep learning approaches.

Criminisi *et al* used mean intensities over displaced, asymmetric cuboidal regions of the volume combined with a random forest classifier to localize several anatomical structures in CT scans[17–19]. Landmarks and relative position context features were used to refine the final bounding boxes. Cuingnet et al.[11] used global contextual information to obtain the initial bounding boxes of the kidneys and refined these with regression forests. The organ localization was used as a pre-processing step for automatic kidney segmentation. In a later work, Gauriau et al.[22] used an extended cascade of random forest regressors. A confidence map of the organs was obtained by voting at a voxel level; the prediction was thresholded to get the final bounding box. Recently, Samarakoon et al.[23] introduced light random regression forests, which use less nodes than classical random regression forests but produced comparable results in the localization of organs in CT scans. Zhou *et al* used organ localization as a pre-processing step for full 3D segmentation of 18 organs[12,15,16]. Organs were localized per orthogonal view using template matching[24], hand-crafted features and local binary patterns[12,16]. A 3D bounding box per organ was determined by majority voting.

Deep Learning is an area in machine learning that has become popular in the last five years[6]. For image processing, Convolutional Neural Networks (ConvNets) are most used. ConvNets learn directly from the raw image data, reducing the semantic gap created by hand-crafted features and reducing the engineering time spent on designing features. A substantial number of works on organ detection already used neural networks. Shin et al.[26] used autoencoders to detect landmarks that can roughly indicate the location of organs. Small ConvNets were used to localize regions in CT scans[27,31]. Twelve regions of the body were detected in axial patches to approximate bounding boxes using a ConvNet of two convolutional layers[31]. In

Table 2.1: Overview of the structures that were detected by previously published organ localization methods. The number of CT scans used and whether the approach dealt with abnormalities is tabulated. Below the horizontal line, the methods that used Deep Learning (DL) are listed.

Method	Structures detected								CT scans	Abnormalities
	Lungs	Liver	Spleen	Kidneys	Gallbladder	Bladder	Sacrum	Femoral heads		
Zhan et al.[13]				■				■	40	No
Criminisi et al.[17]	■	■		■					39	Yes
Criminisi et al.[18]	■	■	■	■	■				100	Yes
Pauly et al.[20]	■	■							33	Yes
Cuingnet et al.[11]				■					223	Yes
Zhou et al.[15]		■	■	■					660	Yes
Criminisi et al.[19]	■	■	■	■	■			■	400	Yes
Gal et al.[25]	■								247	No
Shin et al.[26]		■	■	■					78	Yes
Zhou et al.[12]	■	■		■		■	■		1300	Yes
Zhou et al.[16]	■	■	■	■	■	■		■	300	Yes
Gauriau et al.[22]		■	■	■	■				130	No
Samarakoon et al.[23]		■	■	■				■	100	No
Roth et al.[27]	■	■							≈ 7	No
Humpire Mamani et al.[28]	■	■	■	■					553	Yes
de Vos et al.[29]		■							400	No
Hussain et al.[30]				■					100	Yes

a similar approach, Roth et al.[27] detected five regions of the body with a five-layer ConvNet.

ConvNets were also already used to obtain 3D bounding boxes around organs in CT scans[28–30,32]. A method to detect the heart, aortic arch, and descending aorta in the cardiac area was proposed by de Vos et al.[32] using AlexNet[33]. Different ConvNets were used for each organ and each orthogonal view. Each ConvNet returned 1D predictions, in which a static threshold was used to obtain binary predictions. Thereafter, the binary predictions were joined to obtain 3D bounding boxes. In a recent publication, de Vos et al.[29] proposed a single ConvNet able to detect six organs

in all the orthogonal views using Spatial Pyramidal Pooling to deal with the different input size. Hussain et al.[30] used single ConvNets per orthogonal view with slices of 256x256 as input. The predictions of each orthogonal view were concatenated into a fully connected layer to provide a voxel-wise prediction. This organ localization approach was used as a pre-processing step. In our previous work[28], we proposed a method to localize six organs in 3D thorax-Abdomen CT scans using a single multi-label ConvNet per orthogonal view. We localized six organs: the liver, spleen, and left and right lungs and kidneys. The input of each multi-label ConvNet contained three slices: the slice being analyzed and the slices located five slices above and below it. Feeding extra slices as context to predict the output of a certain slice was an important improvement compared to the work by de Vos et al.[29], where they only used the current slice under analysis. Each ConvNet returned 1D predictions, which were thresholded with an optimized threshold per organ and orthogonal view. The three thresholded predictions were joined to produce a 3D bounding box per organ.

In this study, we extended our previous work with several improvements. Firstly, we are using a deeper ConvNet architecture and extensively experimented with multiple configurations to optimize the number of slices in the input set and the spacing between the input slices. Secondly, we introduce more extensive data augmentation and train the model with a hard sample mining strategy which makes sure that difficult samples are presented more often to the network. Finally, we trained the method with a much larger number of scans and applied to detect more organs.

Table 2.1 summarizes the previous work and shows the organs targeted by each approach. The methods are compared by data set size and whether the method was applied to cases with abnormalities in the human body. Moreover, table 2.1 shows whether the method is based on deep learning.

2.2 Materials and methods

2.2.1 Patient data

The data used in this paper was collected from the Radboud University Medical Center, Nijmegen, the Netherlands. We collected CT data of patients who were referred from the oncology department to our department in 2015. In total, 1884 thorax-abdomen CT scans were collected from 921 patients. 443 patients had one scan, 198 patients had two, 145 patients had three, 134 had four to seven scans, and one patient had eight scans. Age of the patients ranged from 18 to 92 years with an average of 58 years. Table 2.2 shows the scanners and protocols used to acquire the thorax-abdomen CT scans. Since we collected CT data from patients who were

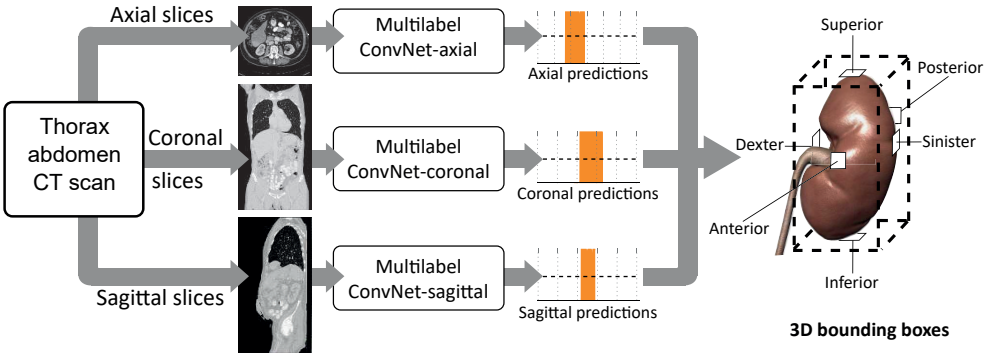


Figure 2.1: A diagram of the proposed approach to obtain 3D bounding boxes of multiple organs from a thorax-abdomen CT scan. From the 3D volume, orthogonal slices are extracted which are then fed into three separate multi-label convNets. The result of each convNet is a probability for each organ per orthogonal slice that the organ is present. Post-processing is applied to keep the largest component from the 1D predictions per orthogonal view. Combining the 1D bounding boxes leads to the final 3D bounding box. The final bounding box is composed by six wall sides: superior, inferior, anterior, posterior, dexter, and sinister.

undergoing oncological workup, many abnormalities are present. We intentionally collected a difficult set of scans which included many abnormalities and may have missing organs due to surgery.

Table 2.2: Summary of scanners and protocols used to acquire the 1884 CT scans in our data set.

Manufacturer	Scanner model	Recons. kernel	Slice thickness	Scans
Toshiba	Aquilion One	FC09	1 mm	410
Siemens	Somatom definition AS	I30f/3	1 mm	7
Siemens	Sensation 16	B30f	2 mm	982
Siemens	Sensation 64	B30f	1 mm	485
Total				1884

The data set was randomly split up at patient level into 60% for training, 20% for validation, and 20% for testing. The training set contained 1130 scans from 652 patients and was used to train the ConvNets. The validation set contained 377 scans from 120 patients and was used to find the best configuration of the ConvNet and to monitor the training process. The test set contained 377 scans from 149 patients and

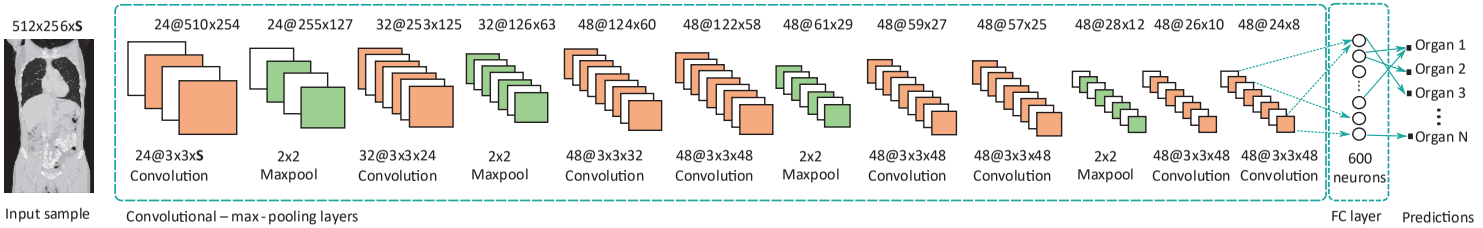


Figure 2.2: The architecture of the multi-label ConvNet consists of 8 convolutional layers, 4 max-pooling layers, and 1 fully-connected layer of 600 neurons to predict the presence of N organs. The filter size and feature map size are located at the bottom and top part, respectively. Note that the first convolution layer contains 24 kernels and $3 \times 3 \times S$ as filter size, where S represents the number of slices in the input set. The input size of the coronal and sagittal ConvNets is $512 \times 256 \times S$. For the axial ConvNet, the input is $256 \times 256 \times S$ and hence the sizes of the feature maps at the top of the figure are different for this ConvNet. The predictions of N organs for a single input are determined by the multilabel cost and sigmoid function.

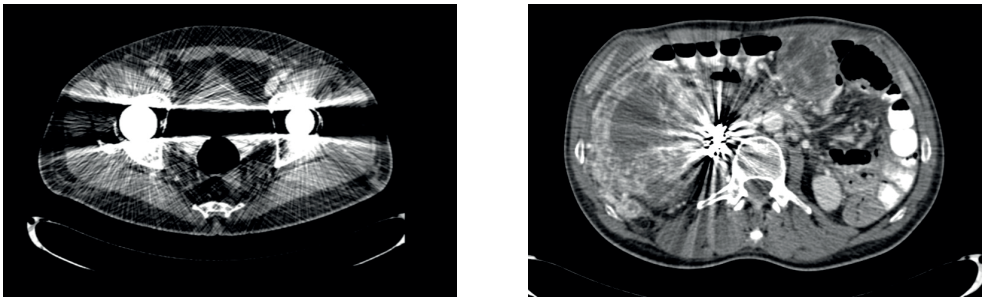
was used for evaluation.

2.2.2 Organ annotation

Five human observers annotated the location of 11 structures for this study: the liver, spleen, gallbladder, bladder, sacrum, and left and right lungs, kidneys, and femoral heads. All 1884 CT scans were initially divided among four human observers to be annotated. Then, an independent fifth human observer was asked to annotate the test set of this study, which left us with two sets of test set annotations. All human readers followed the same protocol to annotate the data in all three orthogonal views. They annotated the first and last slices with visible tissue of the organ as walls of the bounding box; this process was performed for every orthogonal view.

Common abnormalities found were tumors, cysts, fluid, and foreign objects such as clips and prostheses. Examples can be seen in figure 2.3, showing two slices with metallic objects inside the body that created streak artifacts during the CT image acquisition process, and figure 2.3b, showing several liver tumors and clips. Additionally, organs (e.g., gallbladder, kidneys, and spleen) can be absent in the human body due to surgery or anatomic variations. If a patient had a hip replacement (see figure 2.3a), the titanium prosthesis was not annotated as femoral head. Sometimes, organs can be very difficult to locate; for instance, the bladder and gallbladder are difficult to detect when the organ is empty, especially in slim patients.

Due to the large size of the dataset, we used an algorithm to speed up the annotation collection process of the lungs. An automatic lung segmentation tool was used to retrieve initial bounding boxes of the lungs in the training and validation set[34], which were subsequently checked for errors by our human annotators. The lungs in the test set were annotated completely manually.



(a) **(b)**

Figure 2.3: Examples of abnormalities in our data set: (a) prostheses replacing femoral heads, (b) surgical clips and tumors inside the liver.

2.2.3 Multi-label ConvNets for organ localization

We propose an automatic method to localize organs in 3D thorax-abdomen CT-scans based on the presence of the organs on 2D slices. The schematic overview is visualized in figure 2.1. The method consists of three multi-label ConvNets – one for each orthogonal view – that independently process the CT scan in three orthogonal views (axial, coronal, and sagittal). Each ConvNet returns predictions along the axis, indicating the presence of all organs. We thresholded the predictions for each organ at 0.5 and selected the largest connected component. We joined the largest 1D connected-components of each orthogonal view to create a 3D bounding box per organ, for instance around the left kidney as shown in figure 2.1. In the next subsections, we explain the method in more detail.

Preprocessing

The original CT scans were resampled from $512 \times 512 \times Z$, where Z represents the number of slices in the CT scan, to $256 \times 256 \times 512$ using cubic interpolation. Consequently, the slices that are extracted from a volume have a fixed input size: 256×256, 512×256, and 512×256 for axial, coronal, and sagittal views, respectively. The voxel intensities of the resampled CT volumes were subsequently rescaled from $[-1000, 400]$ Hounsfield Unit (HU) to $[0, 1]$; values outside this HU range were clipped.

ConvNet architecture

The proposed multi-label ConvNet is shown in figure 2.2. The architecture is an extended version of the architecture proposed in our previous work[28]. The first convolutional layer consists of 24 kernels of size $3 \times 3 \times S$, where S represents the number of slices in the input size. The second convolutional layer consists of 32 kernels of $3 \times 3 \times 24$. The third convolutional layer consists of 48 kernels of $3 \times 3 \times 32$. Layers four to the eight consist of 48 kernels of $3 \times 3 \times 48$. Max-pooling is applied after the first, second, fourth, and sixth convolutional layers in non-overlapping windows of 2×2. Every max-pooling reduces the size of each patch by half, for instance, from $24@510 \times 254$ to $24@255 \times 127$ after the first max-pooling layer of the coronal and sagittal networks. The last layer is a fully-connected layer with 600 neurons, where dropout[35] is applied with $p = 0.5$ to avoid overfitting. We used a sigmoid as activation function and the multi-label cost[36] to obtain the multi-label predictions. Rectified linear units (ReLU) are used in the convolutional and fully-connected layers.

Input settings

An important contribution of our study is that we input multiple slices. Therefore, we conducted experiments to show the benefit of this approach. We hypothesized that the performance of organ detection of a 2D slice can be improved by including neighboring slices. We defined S as the number of slices and Δ as the spacing between slices. The slice under analysis was always located in the middle of the set. We experimented with different configurations and evaluated its impact to the performance to find the best combination of S and Δ. In our experiments, $S = (1, 3, 5)$ and $\Delta = (1, 2, 3, 4)$ were evaluated. Note that 1S represents a configuration with only a single slice. As consequence, we had in total 9 experiments: Exp[1S, 3S1Δ, 3S2Δ, 3S3Δ, 3S4Δ, 5S1Δ, 5S2Δ, 5S3Δ, and 5S4Δ]. Each experiment contained three ConvNets, one per orthogonal view.

Based on the results of these experiments, we selected the configuration with the best performance and used that as the final configuration for our approach. We applied this configuration to the independent test set and evaluated the performance. The performance of the Exp1S configuration on the test set was assigned as the baseline system for comparison.

Training

The ConvNets were trained using the training and validation sets. We used a batch size of 80 slices for all experiments. In one epoch, all slices of the training set were shown to the ConvNet once. Glorot weight initialization was used to initialize the weights of the network. RMSprop[37] was used as gradient descent optimization with learning rate 0.001 and $\rho = 0.99$. Thereafter, the learning rate was reduced by $1/10$ every 5 epochs. Training stopped when the average area under the ROC curve (AUC) performance on the validation set stopped improving within the previous five epochs; the model with the highest AUC on the validation set was selected as the final model.

We applied data augmentation to reduce overfitting and improve generalizability of the ConvNet. Random rotation, translation, and scaling augmentations were used in the input set. The rotation augmentation was applied by randomly rotating the slice from -5 to 5 degrees using the slice center as the center of gravity with linear interpolation. The translation and scaling augmentations were applied with a random scale between [-0.1,0.1] to the size of the slice, where 0 represents no change. The same data augmentation was used for all the slices in the input. Furthermore, we used selective sampling[38]. Since ConvNets are typically trained with large datasets, the data was carefully prepared and organized to achieve the highest performance

possible. A dynamic approach called selective sampling was used to select specific samples according to the current ConvNet performance. By following this approach, difficult samples were shown more often than easy samples to the ConvNet during training. After training the ConvNet for one epoch, we applied selective sampling based on the loss value of the previous epoch for the samples. We defined the threshold as 20%; any sample with a loss less than that value was omitted during training in the current epoch.

Post-processing

A post-processing step was necessary to handle noisy predictions caused by, for example, abnormalities in the scan. We smoothed the raw 1D predictions obtained per orthogonal view using a 1D Gaussian filter with $\sigma = 10$ slices. Afterwards, a 0.5 threshold was applied to get a binary output. Connected-component analysis was performed, and each orthogonal view's largest 1D connected-component was kept. The three 1D binary outputs were joined to obtain a 3D bounding box per organ. If an organ was not detected on at least one of the binary predictions, the organ was assumed absent and no 3D bounding box output was produced.

Table 2.3: Performance (mean $\pm$ standard deviation) of the different ConvNet configurations on the validation set.

Experiment	Dice score	Jaccard coef.	Wall distance (mm)
Exp1S	0.82±0.22	0.74±0.24	5.70±11.65
Exp3S1Δ	0.87±0.21	0.81±0.23	3.94± 9.81
Exp3S2Δ	**0.88±0.21**	**0.83±0.23**	**3.32±9.28**
Exp3S3Δ	0.88±0.21	0.82±0.23	3.54± 9.40
Exp3S4Δ	0.87±0.21	0.81±0.24	3.91± 9.72
Exp5S1Δ	0.87±0.21	0.81±0.24	3.92±10.11
Exp5S2Δ	0.86±0.21	0.80±0.24	4.28±10.33
Exp5S3Δ	0.86±0.22	0.81±0.24	3.98± 9.78
Exp5S4Δ	0.87±0.21	0.82±0.23	3.78± 9.71

2.2.4 Evaluation

The 3D bounding boxes obtained by our method were compared to the ground truth using Dice score, Jaccard coefficient, and the wall distances to the reference bounding box. Dice score and Jaccard coefficient were defined as:

$$\text{Dice score} = \frac{2|X \cap Y|}{|X|+|Y|} \tag{2.1}$$

$$\text{Jaccard coefficient} = \frac{|X \cap Y|}{|X \cup Y|} \tag{2.2}$$

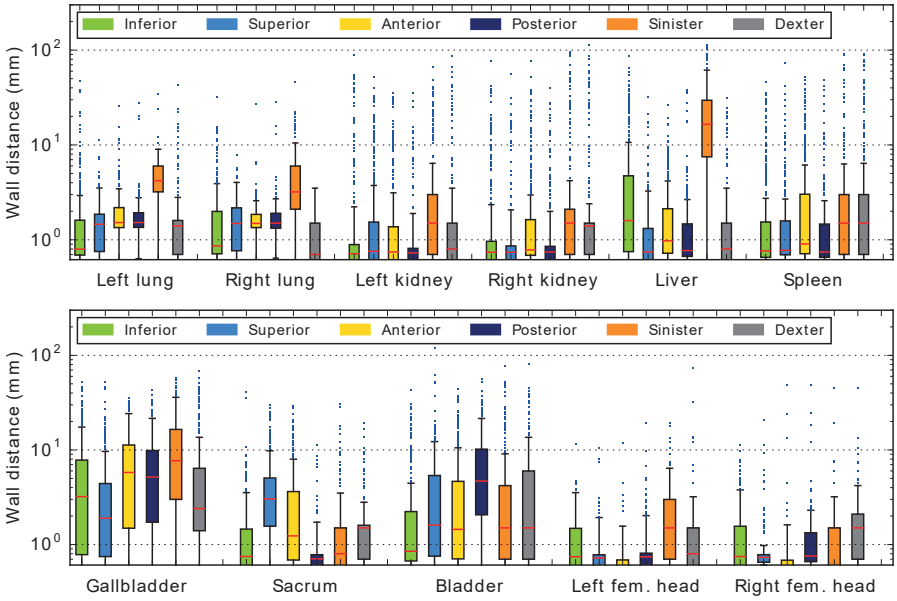
where X represents the predicted mask and Y the reference mask.

2.3 Results

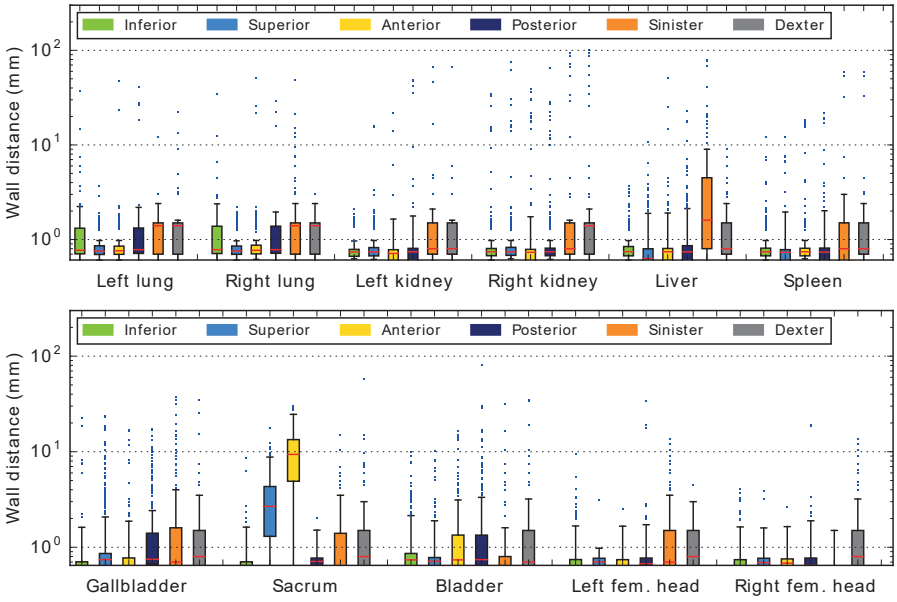
Table 2.3 shows the performances of the different configurations on the validation set. The results show a substantial performance difference between the single slice experiment (Exp1S) and the experiments that employ multiple input slices for all evaluation metrics; this finding confirms our hypothesis that using multiple input slices allows the network to perform a more accurate assessment.

We evaluated the influence of our post-processing steps on our results. We computed the accuracy, sensitivity, specificity, false positive, and false negative rate at the slice level for the three ConvNets with and without post-processing. We found that extracting the largest connected component reduced the amount of false positive slices and therefore slightly increased the overall accuracy at the slice level. Furthermore, we computed the mean wall distance error for Exp3S2Δ without smoothing in the post-processing step on the validation set. We found that the smoothing was beneficial for most organs and slightly decreased the mean wall distance error.

We selected the Exp3S2Δ as the final configuration for our approach because it had the lowest wall distance error. We also ran the Exp1S configuration on the test set as a baseline system for comparison. Table 2.4 shows the performances on the test set per organ. The average wall distance error for Exp1S, Exp3S2Δ, and second observer were 5.40$\pm$9.75mm, 3.20$\pm$7.33mm, and 1.23$\pm$3.39mm, respectively. Using the best configuration, we obtained an average wall distance of 3.20 ± 7.33mm. It failed to detect the bounding box in 43 cases for the gallbladder, 17 cases for the left kidney, 11 case for the right kidney, 6 cases for the spleen, 2 cases for the right femoral head, and 2 cases for the bladder. For the independent second human observer, wall distance was substantially smaller: 1.23 ± 3.39mm. Figure 2.4 shows box plots of wall distance per organ of the proposed approach (Exp3S2Δ) and second human observer, respectively.



(a) Box plots of the experiment Exp3S2Δ



(b) Box plots of the second observer

Figure 2.4: Comparison between Exp3S2Δ and the second observer. The box plots are grouped per wall side and organ, showing the absolute wall distance in mm on the test set. The red line represents the median. Note that the Y-axis is on a logarithmic scale.

2.4 Discussion

In this paper, we presented a method to localize organs in 3D thorax-abdomen CT scans using multi-label ConvNets. We use three ConvNets – one for each orthogonal view – to predict the presence of 8 organs and 3 bony structures. The 3D bounding boxes around each organ were obtained by combining the predictions of the three orthogonal views. Previous work used a single slice as input for the 2D ConvNets[29,32], but our proposed system uses a set of slices as input to provide more context of the information around the slice under analysis. Feeding multiple slices requires a small additional computation effort in the first convolutional layer, but gives a substantial performance boost. Our method already detects 11 structures simultaneously, but it can easily be adapted to handle more organs and structures. We expect that this approach will perform well in that scenario, because ConvNets have been already successfully applied to multi-class problems with up to 1000 classes[39].

To put our results into perspective, we compared our results with previous work using the wall distance as metric, as this is the most widely reported metric in prior research. Table 2.5 shows the results obtained by the experiments of this study, our and other previous work, and the second observer in this study. Note that it was only possible to make a direct comparison for all organs and structures between this study and our previous work[28]. To obtain results from the system of our previous work, we retrained that system using the same dataset, annotations, and organs described in this paper.

Table 2.5 shows that recent methods based on deep learning substantially reduced the mean wall distance error in comparison to traditional machine learning methods[28–30,32]. Moreover, our method improves upon previous work when we look at the reported metrics and results of these studies. Note that caution should be taken when comparing the performance of our method with performance from previous studies because different datasets were used (as shown in table 2.1); a comparative study on a single data set would provide a more objective comparison. Despite our good results, there is still room for further improvement, as is evident from the fact that the independent human observer still performs substantially better.

Table 2.5 shows that the liver wall distance of previous methods ranges from 8.87mm to 18.13mm, and our method obtained the lowest mean distance 5.84mm. The corresponding box plot (figure 2.4) shows relatively high distances for the sinister wall of the liver; a possible explanation is that in some cases there is only a small tip of the liver visible on the most sinister sagittal slices, making it difficult to determine where the liver ends. We assume this was also the reason for relatively poor performance of the independent human observer, and we saw a similar trend

Table 2.4: Performances of two of our configurations and the second observer on the test set per organ. The obtained Dice score, Jaccard coefficient, and the absolute wall distance are tabulated (mean ± standard deviation). The average wall distance error for Exp1S, Exp3S2Δ, and second observer were 5.40±9.75mm, 3.20±7.33mm, and 1.23±3.39mm, respectively.

Method	Organs										
	Left lung	Right lung	Left kidney	Right kidney	Liver	Spleen	Gallbladder	Sacrum	Bladder	Left femoral head	Right femoral head
Dice											
Exp1S	0.93±0.09	0.95±0.04	0.79±0.23	0.82±0.21	0.90±0.06	0.83±0.17	0.47±0.24	0.91±0.05	0.76±0.17	0.83±0.22	0.80±0.24
Exp3S2Δ	0.96±0.02	0.97±0.01	0.89±0.23	0.90±0.20	0.92±0.06	0.89±0.17	0.53±0.27	0.95±0.04	0.83±0.17	0.90±0.20	0.88±0.23
2nd Observer	0.98±0.02	0.98±0.01	0.96±0.06	0.95±0.12	0.98±0.02	0.96±0.09	0.83±0.30	0.93±0.03	0.95±0.11	0.92±0.20	0.91±0.22
Jaccard											
Exp1S	0.87±0.11	0.91±0.07	0.70±0.24	0.73±0.22	0.82±0.09	0.73±0.19	0.34±0.19	0.84±0.08	0.64±0.18	0.75±0.21	0.71±0.23
Exp3S2Δ	0.93±0.03	0.94±0.03	0.84±0.23	0.85±0.21	0.86±0.09	0.84±0.19	0.40±0.23	0.90±0.06	0.73±0.20	0.85±0.19	0.83±0.22
2nd Observer	0.97±0.03	0.97±0.02	0.93±0.08	0.92±0.13	0.96±0.03	0.94±0.09	0.78±0.29	0.86±0.06	0.92±0.12	0.89±0.19	0.88±0.22
Wall distance (mm)											
Exp1S	4.60±11.16	3.33±7.05	5.88±10.48	5.36±10.47	7.75±14.42	6.07±10.19	9.46±11.29	4.32±6.27	7.30±8.62	2.48±4.85	2.84±5.40
Exp3S2Δ	2.31± 3.05	1.99±2.64	2.67± 7.18	3.03± 9.30	5.84±12.69	3.37± 8.46	7.09± 8.91	2.13±3.54	4.70±7.94	1.04±2.32	1.02±2.51
2nd Observer	1.11± 1.99	1.15±2.17	0.98± 2.99	1.65± 6.54	1.46± 3.81	1.03± 2.88	1.24± 2.92	2.60±4.56	1.04±2.83	0.67±1.36	0.57±1.03

in previous work[28,29].

Table 2.5: Mean wall distance per organ obtained by previous work, including our preliminary work, our proposed method, and the second observer. Distances are in millimeters and methods are in chronological order. Note that the results of other algorithms are obtained on different data sets. The results tabulated here for our previous work (Humpire *et al* 2017) are obtained after retraining that system using the data from this study.

Method	Organs										
	Left lung	Right lung	Left kidney	Right kidney	Liver	Spleen	Gallbladder	Sacrum	Bladder	Left femoral head	Right femoral head
Zhan et al. [13]	-	-	8.97	8.97	-	-	-	-	-	4.47	4.47
Pauly et al. [20]	14.78	15.02	-	-	18.13	-	-	-	-	-	-
Cuingnet et al. [11]	-	-	7.00	7.00	-	-	-	-	-	-	-
Criminisi et al. [19]	12.90	10.10	13.60	16.10	15.70	15.50	18.00	-	-	10.60	11.0
Gauriau et al. [22]	-	-	5.50	5.60	10.70	7.90	9.50	-	-	-	-
de Vos et al. [32]	-	-	-	-	10.80	-	-	-	-	-	-
Humpire Mamani et al. [28]	2.87	2.60	5.68	5.82	8.19	7.17	11.59	3.61	8.67	1.75	1.91
de Vos et al. [29]	-	-	-	-	8.87	-	-	-	-	-	-
Hussain et al. [30]	-	-	6.19	5.86	-	-	-	-	-	-	-
Samarakoon et al. [23]	-	-	11.52	10.98	15.82	14.84	-	-	-	7.67	7.42
Proposed method	**2.31**	**1.99**	**2.67**	**3.03**	**5.84**	**3.37**	**7.09**	**2.13**	**4.70**	**1.04**	**1.02**
2nd Observer	1.11	1.15	0.98	1.65	1.46	1.03	1.24	2.60	1.04	0.67	0.57

Regarding the sacrum, interestingly, table 2.5 shows that the second observer had a greater mean wall distance than our method (2.60±4.6mm vs 2.13±3.5mm, respectively). This may be due to the anatomical variations in the sacral promontory (anterior wall), which may look very similar to the L5 vertebrate in the coronal view.

The localization of the left and right lungs is challenging in the sagittal view because both lungs look identical in this direction. The single slice experiment (Exp1S)

could not handle this problem properly due to a lack of context around the slice. This issue was largely overcome when using multiple slices, as shown by the left lung-dexter and right lung-sinister wall distances in the box plots.

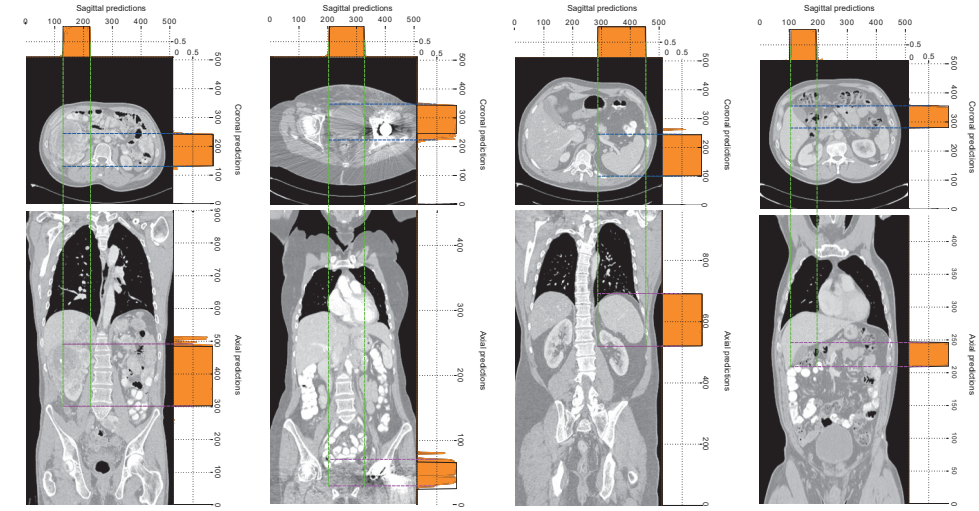
Locating the gallbladder is complicated due to its relatively small volume and sometimes highly irregular shape. Table 2.5 shows that it was the most difficult organ to localize using our method (7.09mm mean wall distance error).

Our data contained many anatomical abnormalities such as tumors, clips, cysts, and fluid. We note that several previous studies (see table 2.1) did not consider abnormal cases. Our results show that our method is able to handle these abnormalities generally well. The large training data set may be responsible for this effect. Our post-processing step is important to avoid irregular predictions provoked by abnormalities.

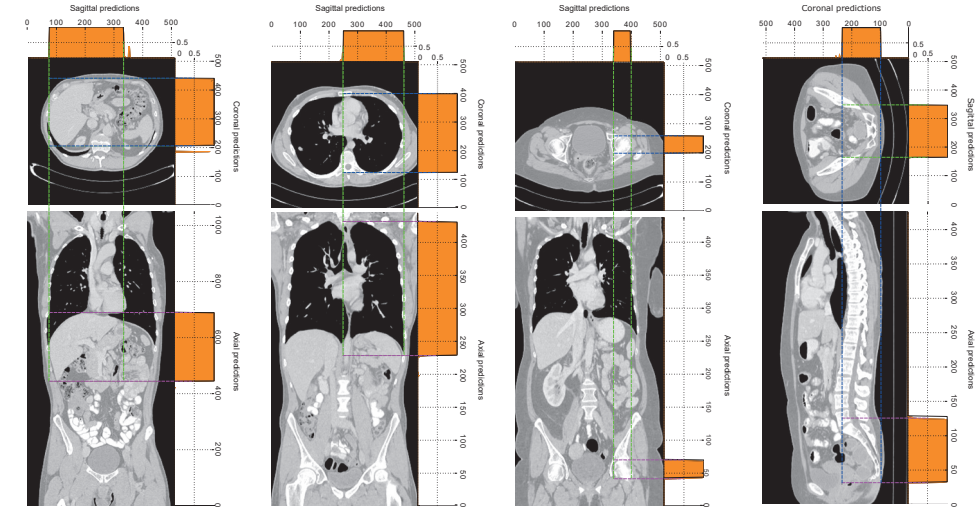
Figure 2.5 shows predictions from Exp3S2Δ and labels projected in axial and coronal slices obtained for the kidneys, bladder, and spleen. Figure 2.5(a) shows the predictions obtained for a complex right kidney surrounded by a large tumor. Our method was able to correctly localize the organ in all views. Moreover, despite that metallic materials can affect the surrounding organ intensities, figure 2.5(b) shows good bladder localization in the presence of a titanium prosthesis of the left femoral head.

As mentioned in the introduction, a possible area where the proposed approach can be used is automatic segmentation. Current popular deep learning approaches for semantic segmentation of structures include 2D U-Net, 3D U-Net and V-Net[8,9,40]. These approaches typically take a full slice or a large subvolume as input. This often leads to an unbalance in the number of positive and negative voxels during training which needs to be tackled by for example weight maps[8,9] or a dice loss function[40]. Focusing the network on a part of the image may speed up training and lead to faster inference time. Mask R-CNN is another popular approach in which regions are first extracted and segmentation is performed in a separate segmentation branch for the extracted regions of interest[41].

The ConvNets were implemented using Theano[42,43]. The experiments were executed using a single NVidia GeForce GTX 1080 on a high-end PC with at least 256 GB of RAM. Training time for a single ConvNet was in the order of 70 hours. Applying the Exp3S2Δ model to a single scan (with an average number of slices of 700) took approximately four seconds.



(a) Right kidney (b) Bladder (c) Spleen (d) Gallbladder



(e) Liver (f) Left lung (g) Left femoral head (h) Sacrum

Figure 2.5: ConvNet predictions projected along an axial and coronal slice. The blue, green, and pink lines represent the bounding box walls obtained by our proposed system, Exp3S2Δ; black lines indicate the reference bounding box. In the top part, sagittal predictions are plotted, coronal predictions are plotted next to the axial slices, and axial predictions are plotted next to the coronal slices. Note that (h) includes a sagittal slice for better visualization of the sacrum.

2.5 Conclusion

An efficient and robust automatic method for 3D localization of 11 structures in thorax-abdomen CT scans using a single ConvNet per orthogonal view is proposed. Multiple slices were used as input to provide more context around the slice under analysis and we have shown that this gave a substantial boost to the ConvNet performance. The proposed approach can localize organs even when abnormalities such as tumors, cysts, fluid, and metal artifacts are present. We compared our work to recent papers and have shown that our approach outperforms recent work on organ localization.

Spleen segmentation and Splenic Volume Change

3

G.E. Humpire Mamani, J. Bukala, E. Scholten, M. Prokop, B. van Ginneken and C. Jacobs

Original title: Fully Automatic Volume Measurement of the Spleen at CT Using Deep Learning

Published in: Radiology: Artificial Intelligence, 2(4):e190102, 2020

Abstract

Purpose: To develop a fully automated algorithm for spleen segmentation and to assess the performance of this algorithm in a large dataset.

Materials and Methods: In this retrospective study, a three-dimensional deep learning network was developed to segment the spleen on thorax-abdomen CT scans. Scans were extracted from patients undergoing oncologic treatment from 2014 to 2017. A total of 1100 scans from 1100 patients were used in this study, and 400 were selected for development of the algorithm. For testing, a dataset of 50 scans was annotated to assess the segmentation accuracy and was compared against the splenic index equation. In a qualitative observer experiment, an enriched set of 100 scan-pairs was used to evaluate whether the algorithm could aid a radiologist in assessing splenic volume change. The reference standard was set by the consensus of two other independent radiologists. A Mann-Whitney U test was conducted to test whether there was a performance difference between the algorithm and the independent observer.

Results: The algorithm and the independent observer obtained comparable Dice scores (P = .834) on the test set of 50 scans of 0.962 and 0.964, respectively. The radiologist had an agreement with the reference standard in 81% (81 of 100) of the cases after a visual classification of volume change, which increased to 92% (92 of 100) when aided by the algorithm.

Conclusion: A segmentation method based on deep learning can accurately segment the spleen on CT scans and may help radiologists to detect abnormal splenic volumes and splenic volume changes.

3.1 Introduction

Splenic volume change (SVC) can occur as a result of infection, lymphoma, injury, variations in splenic vascularization, and other reasons[44–51]. Full manual segmentation of the spleen in three dimensions is time-consuming and not feasible in clinical practice. Instead, visual estimation or an approximation equation is typically used by radiologists to assess the size of the spleen. To the best of our knowledge, there are no studies that have investigated whether substantial SVC goes undetected using these methods. During oncologic treatment, SVC can occur as an adverse effect of chemotherapy[52]. A precise SVC may help clinicians in their treatment choices. The first work in splenic volume approximation used the splenic index[53,54]. The splenic index is calculated using the equation $V = 30 + 0.58 \times D \times L \times H$, where depth ($D$), length ($L$), and height ($H$) are two-dimensional measurements of the spleen in the axial or coronal plane. Figure 3.1 shows these measurements in two-dimensional sections of a CT scan. A precise three-dimensional segmentation can achieve an accurate volumetric measurement of the spleen. Methods such as multiatlas[55–57], graph-cut[56–58], active shape models[59], active contours[60], level-sets[61], and random forest[62] have been extensively used to segment the spleen.

In recent years, Deep learning (DL) approachesconvolutional neural networks in particularhave achieved high performance in many areas of computer vision and have been successfully applied in medical imaging[7,63–67]. A sequence of convolutional layers is applied to the image to optimize segmentation tasks, every convolution can highlight different features, and combining these layers with pooling and nonlinear operations make these networks very powerful. For medical imaging, 2D U-Net[8], 3D U-Net[9], and variant architectures have been successfully used to segment structures and organs[68–73]. These architectures are based on a contracting path of convolutions followed by an expanding path of convolutions to produce voxel-wise predictions. The deepest convolutions learn global features, and the last convolutions obtain the fine segmentation prediction. In an end-user comparison[74], three commercial systems showed that the liver and spleen segmentation volumes were fast and accurate, but the initial fully automatic segmentation failed for some cases and differed by 0.4%9.8% from the final segmentation after correction for the remaining cases. The readers in this previous study took between 1 and 3 minutes on average to perform the corrections. The study showed that the performance of the fully automatic initial segmentation can be improved[74]. In this study, an automatic segmentation algorithm for the spleen was developed on a large dataset of thorax-abdomen CT scans from patients undergoing oncologic workup. Because these patients undergo various types of cancer treatment (eg, chemotherapy and/or

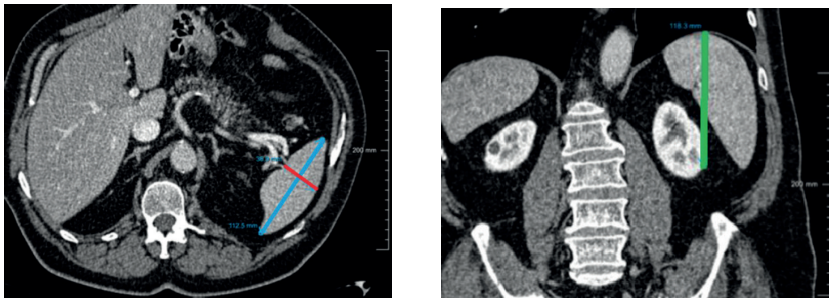


Figure 3.1: An example of the two-dimensional measurements needed to compute the splenic index: depth in blue, height in red, and length in green. Note that depth and height do not necessarily have to be measured on the same transversal section.

radiation therapy) and are at different stages of disease, the images contained both local and widespread abnormalities throughout the scan. Our system was developed using a dataset of 400 CT scans (selected from 1100 patients) and tested using a dataset of 50 scans. Finally, a qualitative observer experiment with an experienced radiologist was conducted to assess whether the algorithm can help radiologists in assessing SVC in 100 patients (selected from 500 patients).

3.2 Materials and Methods

3.2.1 Patient Data

The data in this retrospective study were collected from Radboud University Medical Center. The institutional review board granted a consent waiver for the clinical images used in this study. We retrieved all thorax-abdomen CT studies referred from the oncology department between January 2014 and December 2017. In total, 7415 studies from 2386 patients (mean age, 58 years; range, 1992 years; 54.7% women) were retrieved. Part of this dataset (918 CT scans from 918 patients) was previously used for a different study on developing an algorithm for organ localization[75]. We only included contrast materialenhanced CT scans in this study (n = 6972 studies). From the included data, we randomly selected 2150 CT scans from 1650 patients to obtain four datasets (A, B, C, and D) as depicted in Figure 3.2. As the patients in this dataset underwent an oncology workup, the scans typically presented multiple abnormalities, such as tumors, cysts, and lesions, which may alter the normal anatomy of the spleen. Additional information on CT imaging and datasets is described in Appendix 3.4.

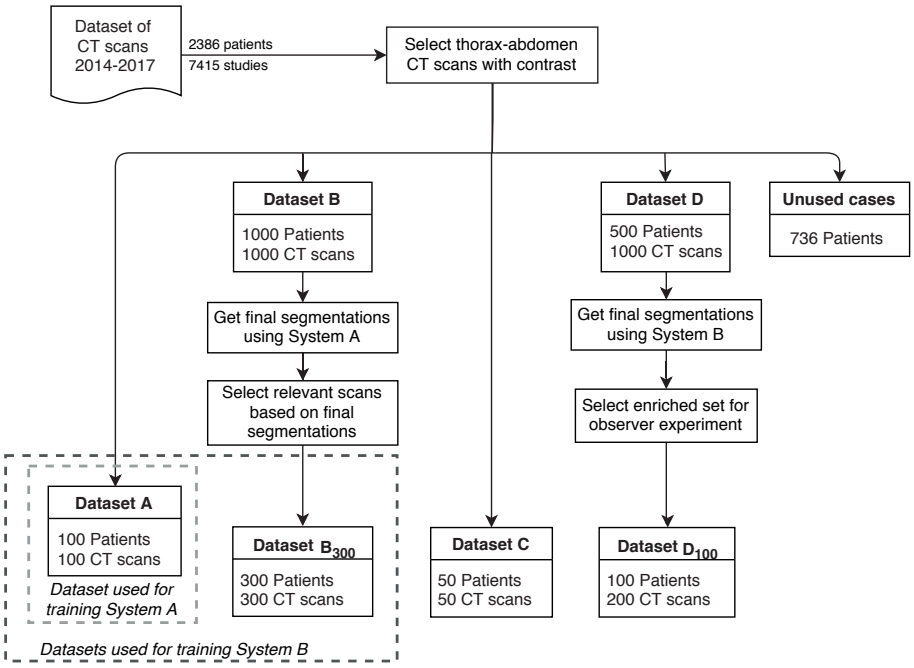


Figure 3.2: Flowchart shows the criteria to distribute the CT scans used in this study into datasets. Dataset A was used for training system A, and dataset A plus dataset B_{300} were used for training system B. Dataset C was used for testing systems A and B. Dataset D_{100} was used for the qualitative observer experiment. Note that dataset B_{300} and D_{100} are subsets of datasets B and D, respectively.

3.2.2 CT Imaging

CT scanners from two manufacturers were used to acquire the CT scans: Toshiba (Aquilion One) and Siemens (Sensation 16, Sensation 64, and Somatom Definition AS). The reconstruction kernels were FC09, FC09-H, B30f, B30fs, and I30f. The contrast agents used were iomeprol, iohexol, iobitridol, and iopromide (Imeron [Bracco Imaging], Omnipaque [GE Healthcare], Xenetix [Guerbet], and Ultravist [Bayer], respectively) with amounts varying between 15 and 140 mL. The section thickness ranged from 0.5 to 3 mm, with most (98.9%) having a section thickness of 1 or 2 mm.

3.2.3 Reference Standard Annotation

On all CT scans in the first training set (dataset A), the spleen was manually segmented by medical students using a tool developed in-house. Students were instructed to verify that the segmentation was correct on all transversal sections and

to peer-review each other. The annotations included the splenic hilum if it was surrounded by splenic parenchyma. Dataset A was used as training data for the first system. Subsequently, we used the first system to obtain the final segmentations of dataset B. These final segmentations were used for selecting 300 additional scans for training of the second system; that gave us dataset B_{300}. The final segmentations of the initial system on dataset B_{300} were manually corrected by the same medical students to train a second system. Later, dataset A plus dataset B_{300} were used for training a second system. For the test set of 50 CT scans (dataset C), the same procedure was used to annotate the spleen in all scans; this was then used as the reference standard for testing the system later on. In addition, one medical student (herein referred to as "independent observer") also annotated dataset C independently without consulting other students or the experienced radiologist. An experienced radiologist (E.T.S., > 30 years of experience in chest radiology) was consulted in difficult cases, performed a quality check, and adjusted (if necessary) the annotations of dataset A, B_{300}, and C (reference standard).

3.2.4 Preprocessing and DL Network Settings for Automatic Spleen Segmentation

Values outside of the attenuation range (-500 to +400 HU) on the CT scans were clipped to discard unnecessary data for this task. The scans and reference masks were resampled to $1 \times 1 \times 1$-mm resolution using cubic and nearest neighbor interpolation, respectively. We used the 3D U-Net network[9] as the architecture of our system because it uses three-dimensional context to predict the results. This segmentation network and its two-dimensional variant reached high performance in multiple applications[8,68,69,76–78]. Because of the large memory footprint of the 3D U-Net, each scan was divided into patches. At the edges of the CT scan, mirroring was used as border handling when the patch covered an area outside the scan. Additional details on the inputs can be found in Appendix 3.4.

3.2.5 Network Training

The network performance was evaluated after every epoch (in one epoch, every CT scan in the training set was used once) using the Dice score as the metric to select the optimal model. The training stopped when the mean Dice score stopped improving for 10 epochs. The optimal model of each network was used to evaluate the test set (dataset C). We trained our first network from scratch using dataset A. The network was evaluated after every epoch using 30% of the training scans. The training

stopped after 21 epochs to find the optimal model, which we referred to as system A. We used system A to process dataset B (n = 1000) to visually identify relevant cases. These relevant scans composed dataset B_{300} (see Appendix 3.4). We trained a new network from scratch using dataset B_{300} plus dataset A: segmentation system B. Thus, system B was trained using 400 scans. The network stopped training after 43 epochs. We evaluated both segmentation systems A and B on the test set (dataset C) of 50 scans.

3.2.6 Postprocessing for Automatic Spleen Segmentation

To produce the final segmentation results, each patch was processed separately, and the results were stitched together and thresholded at 0.5 to obtain binary results. Afterward, we applied connected components analysis and only retained the largest connected component. The output was then resampled back to the original scan resolution using nearest neighbor interpolation.

3.2.7 Qualitative Observer Experiment

To test the clinical usefulness (detection of growth or shrinkage of the spleen over time) of our segmentation system, we performed an observer study using an enriched set of cases. To define growth or shrinkage, we used a tolerance of $\pm$25% in the SVC in this study. Thus, SVC of less than -25% was classified as shrinkage, and SVC of greater than +25% was classified as growth. Values within -25% to +25% were considered normal SVC. We computed the SVC over time in dataset D (500 new patients) to obtain the enriched dataset D_{100} (100 patients). See Appendix 3.4 for more details. The scan-pairs in dataset D_{100} were presented in a random order to an experienced radiologist in a dedicated workstation. We considered three different reading modes for splenic volume change assessments (SVCa): visual SVCa, automatic SVCa, and assisted SVCa. See Appendix 3.4 for more details. A radiologist and a 4th-year resident defined the reference standard for dataset D_{100}. They classified the scan-pairs visually as is currently done in clinical practice. In case of disagreement, a consensus meeting was held. The consensus reference standard was used to compare against visual SVCa, assisted SVCa, and automatic SVCa.

3.2.8 Statistical Analysis and Evaluation

Dice scores, relative absolute volume difference, maximum Hausdorff distance, and average symmetric surface distance (ASSD) were used to measure the similarity between the predictions and the reference masks. Per metric, we reported the mean,

standard deviation, and two-sided 95% Confidence Intervals (CI) computed using 1000 random bootstraps. We computed the P values using the Mann-Whitney U test to test whether there was a statistical difference between the final system and the human observer (primary objective), and between the prototype and the final system (secondary objective). A P value less than .05 (two-tailed) was considered statistically significant. The metrics in this article can be found in Appendix 3.4. We used an in-house developed Python 3.6 (`https://www.python.org/`) script to perform the statistical analysis.

3.2.9 Algorithm Availability

The segmentation algorithm can be tested online at `https://grand-challenge.org/algorithms/spleen-segmentation/`, in which interested readers can register and upload anonymized thorax-abdomen CT scans; an online workstation showing the segmentation overlays in three dimensions will be output.

3.3 Results

3.3.1 Comparison of Segmentation Methods

First, we compared our automatic spleen segmentation methods on the test set (dataset C) including the independent observer (Table 3.1). System A obtained a Dice score of 0.950 $\pm$ 0.040 (95% CI: 0.938, 0.959), system B obtained 0.962 $\pm$ 0.016 (95% CI: 0.957, 0.966), and the independent observer obtained 0.964 $\pm$ 0.012 (95% CI: 0.961, 0.967). Figure 3.3 shows boxplots comparing the evaluation metrics presented in Appendix 3.4 among methods on the test set. The surface distancebased metrics (maximum Hausdorff, 95% Hausdorff, and ASSD) show that system B had fewer outliers than system A, whereas system B and the independent observer were comparable. Table 3.1 and Figure 3.3b, show that the splenic volumes computed by the splenic index were not reliable. A Mann-Whitney U test was performed to compare the Dice score performance between system A, system B, and the independent observer. The difference between system A and system B (P = .019), and between system A and the independent observer (P = .011) were statistically significant, but the difference between system B and the independent observer was not significant (P = .834). Table 3.2 compares the performance of previous segmentation work. Table 3.2 shows that not all methods can assess abnormalities in the spleen. The relative absolute volume difference shows that the splenic index, Gloger et al.[61], system B, and the independent observer obtained 16.56%, 6.30%, 4.39%, and 3.93%, respectively. In addition,

the relative absolute volume difference between system B and the independent observer (used as reference) was 2.37% (5.17 mL). On the basis of all the metrics, system B outperformed system A. Therefore, we considered system B as the automatic SVCa for the qualitative observer experiment.

Table 3.1: Comparison of Performance among the Experiments on the Test Set. Data shown for dataset C (50 CT scans). The bottom part of the table shows the 95% confidence intervals computed based on 1000 random resamplings (bootstraps). ASSD = average symmetric surface distance, NA = not applicable. † No standard deviation reported since this metrics is zero centered and we use the absolute value.

Method	Metrics				
	Dice score	Relative abs. vol. difference (%)†	Max. Hausdorff dist. (mm)	95% Hausdorff dist. (mm)	ASSD (mm)
			Mean ± SD		
Splenic index	N/A	16.562	N/A	N/A	N/A
System A	0.950 ± 0.040	5.988	8.463 ± 7.645	3.050 ± 5.209	0.825 ± 0.703
System B	0.962 ± 0.016	4.391	5.799 ± 5.819	2.052 ± 4.071	0.629 ± 0.313
Independent observer	0.964 ± 0.012	3.935	5.395 ± 2.438	1.447 ± 0.401	0.606 ± 0.168
			95% Confidence Intervals (CI)		
Splenic index	N/A	12.831-21.012	N/A	N/A	N/A
System A	0.938-0.959	3.938-8.894	6.527-10.785	1.892-4.671	0.660-1.047
System B	0.957-0.966	3.426-5.500	4.435-7.668	1.372-3.262	0.557-0.729
Independent observer	0.961-0.967	3.122-4.799	4.765-6.073	1.341-1.559	0.560-0.653

3.3.2 Results of the Qualitative Observer Experiment

Comparison of SVCa

For the qualitative observer experiment, the two readers who were selected to define the reference standard had disagreement in 13 scan-pairs, and a consensus meeting was held to define the final reference standard. In total, 59 cases were categorized as normal, 26 as growing, and 15 as shrinking in the reference standard. Table 3.3 compares the visual SVCa, automatic SVCa, and assisted SVCa assessments to the reference standard. The visual SVCa classified 81% (81 of 100) of the patients correctly. During the visual SVCa, the radiologist visually approximated the SVC classification in 80 of the 100 patients. In the remaining 20 of the 100 patients, the radiologist used the splenic index because the visual approximation was not evident. The automatic SVCa classified 89% (89 of 100) of the patients correctly. Finally, the

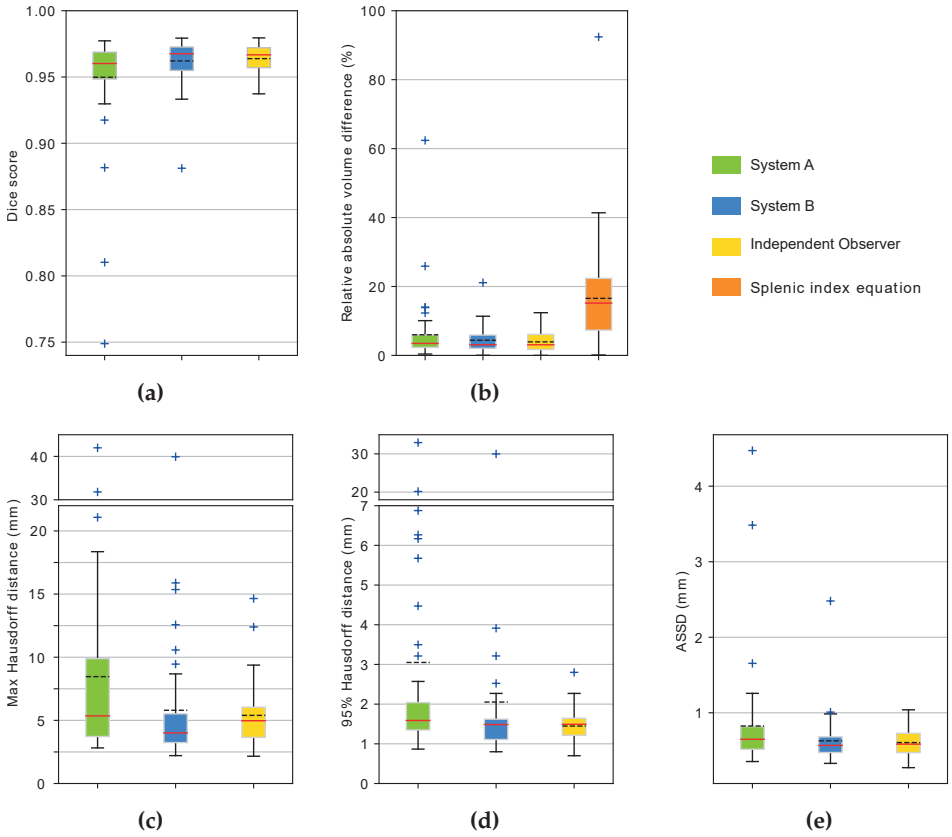


Figure 3.3: Boxplots show the performance of system A, system B, the independent observer, and the splenic index on the test set (dataset C). The methods are compared using, (a) Dice score, (b) relative absolute volume difference, (c) maximum Hausdorff distance, (d) 95% Hausdorff distance, and (e) average symmetric surface distance (ASSD). Mean and median values are depicted with black dashed and red lines, respectively. Note that C and D have two ranges for the y-axis to zoom-in to the boxplots body. Table 3.1 summarizes these results.

assisted SVCa classified 92% of the patients correctly. In total, when observing the three-dimensional automatic segmentations and their volumes (ie, when going from visual SVCa to assisted SVCa), the radiologist changed the classification in 15% (15 of 100) of the patients. In 11 of these patients (73%, 11 of 15), this change resulted in the correct category in the reference standard. For five of these 15 patients, the SVC values were close to the threshold of 25% defined in this study (25.16%, 25.79%, 27.23%, 27.24%, and 27.44%). Figure 3.4 shows examples of the SVC analysis. Figure 3.4A and 3.4B, shows patients with the minimum (-58%) and maximum (+140%)

SVC, respectively. Figure 3.4C, shows a patient where the radiologist changed their classification from normal to growth SVC after seeing our segmentations (assisted SVCa). The probable reason for this change is that the spleen grew proportionally in all the directions. Figure 3.4D, shows a patient with -10% SVC computed by our method (automatic SVCa). In the visual SVCa, the radiologist classified this patient as shrinkage SVC because it looks small in the sagittal plane. In the assisted SVCa, the radiologist changed his classification from shrinkage to normal SVC.

Spleen segmentation ratings

The independent radiologist visually rated the quality of the automatic spleen segmentations of 200 CT scans from 100 patients (dataset D_{100}). The radiologist rated 87% (174 of 200) of the segmentations as excellent, 7% (14 of 200) as good, 3.5% (seven of 200) as bad, and 2.5% (five of 200) as failure. The radiologist grouped 94% (87% excellent and 7% good) of the segmentations as reliable segmentations. Figure 3.5 shows examples of this classification using probability maps in which black contours highlight the final output of the algorithm after postprocessing. Figure 3.5a shows a patient with large tumors in the liver and left kidney; the radiologist rated this segmentation as excellent. Figure 3.5b depicts a patient with a beavertail liver (enlarged liver attached to the spleen); this segmentation was rated as good because of a small error (<5 mm). Figure 3.5c shows a bad segmentation in which the algorithm did not perform well in the upper region of the spleen; this may be the result of the low contrast enhancement on this scan. Figure 3.5d depicts a segmentation failure caused by a dilated stomach.

3.4 Discussion

In this article, we developed an algorithm to segment the spleen using DL on three-dimensional thorax-abdomen CT scans from patients undergoing oncologic workup. The final system (system B, 0.962 Dice) and the independent observer (0.964 Dice) obtained comparable results with no significant (P = .834) difference. In the qualitative observer experiment, we showed that a radiologist improved the performance when assisted by the algorithm to assess SVC.

For the development of the algorithm, an initial dataset of 100 random scans was annotated (dataset A) to train the first system (system A). System A obtained a mean Dice score of 0.950 (95% CI: 0.938, 0.959) on the test set (see Table 3.1). After adding the 300 relevant cases (dataset B_{300}) from dataset B to the training set, a second system (system B) was trained, and this system reached a mean Dice score of 0.962 (95%

Table 3.2: Comparison between our best performing system and previous work. The methods are compared using Dice score, relative absolute volume difference, maximum Hausdorff distance, 95% Hausdorff distance, and average symmetric surface distance (ASSD). The methods from Zhou et al.[64], Roth et al.[65], Gibson et al.[66] use deep learning to segment the spleen. Abd = abdominal, Th-abd = thorax-abdominal. Modality is CT unless otherwise mentioned.

Method	Dice score	Relative abs. vol. difference (%)	Max. Hausdorff dist. (mm)	95% Hausdorff dist. (mm)	ASSD (mm)	Contains abnormalities	Modality
Gauriau et al.[62]	0.870±0.150	-	-	-	2.6±3.0	No	Abd
Wood et al.[60]	0.873	-	-	-	-	Yes	Abd
Gloger et al.[61]	0.906±0.037	6.30±5.40	-	-	1.73±0.68	No	MRI
Zhou et al.[64]	0.920	-	-	-	-	No	Th-abd
Wolz et al.[57]	0.920±0.092	-	-	-	2.27±3.03	Yes	Abd
Tong et al.[56]	0.925±0.065	-	-	-	-	Yes	Abd
Huo et al.[63]	0.926	-	-	-	-	Yes	MRI
Roth et al.[65]	0.928±0.080	-	-	-	-	No	Th-abd
Landman et al.[67]	0.930	-	-	-	-	No	MRI
Okada et al.[58]	0.932±0.052	-	-	-	1.26±2.43	No	Abd
Gibson et al.[66]	0.950	-	-	2.40	0.80	No	Abd
Linguraru et al.[55]	0.952±0.014	-	-	-	0.70±0.10	Yes	Abd
System B	0.962±0.016	4.391±3.790	5.799±5.819	2.052±4.071	0.629±0.313	Yes	Th-abd
2nd observer	0.964±0.012	3.935±3.101	5.395±2.438	1.447±0.401	0.606±0.168	Yes	Th-abd

CI: 0.957, 0.966) on the test set. The independent observer obtained a comparable mean Dice score of 0.964 (95% CI: 0.961, 0.967).

Figure 3.3 shows that system B had a better and more robust performance with fewer outliers than system A. The most challenging case on the test set had a beavertail liver on the CT scan, obtaining a 0.88 Dice score for system B. In the same case, the independent observer obtained a Dice of 0.94, showing that it was also difficult for the independent observer. Table 3.1 and Figure 3.3 show that our algorithm approximated the independent observers performance for all metrics. Our selection process of relevant cases boosted the performance from 0.950 Dice (system A trained with initial dataset A) to 0.962 Dice (system B trained with datasets A and B_{300}).

Based on the visual ratings of the segmentation quality, our method could reliably handle difficult cases. Figures 3.5c and 3.5d show that an abnormal anatomy can lead to less accurate spleen segmentation.

In the SVC analysis, the readers had to come to a consensus in 13 cases to define the reference standard on dataset D_{100} . In the observer experiment, the radiologist changed the classification of 15% (15 of 100) of the patients when going from the visual SVCa to the assisted SVCa. This resulted in a more reliable SVC classification because the SVC is now computed based on precise segmentations and not based on an approximation as the volume obtained by the splenic index. Figure 3.4D, shows

Table 3.3: Comparison of the Visual SVCa, Assisted SVCa, and Automatic SVCa versus the Consensus-Based Reference Standard (dataset D_{100}) in the Qualitative Observer Experiment. The visual SVCa obtained 19 mistakes, assisted SVCa eight mistakes, and automatic SVCa 11 mistakes. Note that the consensus-based reference standard followed the same protocol as the visual SVCa. SVC = splenic volume change, SVCa = splenic volume change assessment.

	Reference Standard			
Visual SVCa	Shrinkage SVC	Normal SVC	Growth SVC	Total
Shrinkage SVC	**9**	2	0	11
Normal SVC	6	**54**	8	68
Growth SVC	0	3	**18**	21
Total	15	59	26	**100**
Assisted SVCa				
Shrinkage SVC	**13**	0	0	13
Normal SVC	2	**54**	1	57
Growth SVC	0	5	**25**	30
Total	15	59	26	**100**
Automatic SVCa				
Shrinkage SVC	**14**	3	0	17
Normal SVC	1	**49**	0	50
Growth SVC	0	7	**26**	33
Total	15	59	26	**100**

a scan-pair in which the radiologist was likely misled. The stomach of the patient is full in the baseline scan, which pushes the spleen toward the ribs. On the follow-up scan, the stomach is empty, giving more space to the spleen to expand. Although the volume changed -10% over time, it was within the range of normal SVC defined by us. This indicated that our method can help radiologists to reduce bias when measuring SVC. In this work, the threshold to classify shrinkage, normal (no substantial), and growth SVC was defined as +25%. Three cases obtained automatic SVCa values around this fixed threshold. Future investigations will be useful to define better thresholds for clinical practice to classify SVC. Note that our qualitative observer experiment resulted in percentages that were not representative for a large random set because we were using the enriched dataset D_{100}. This subset was created by selecting 50 random patients classified as having substantial (either growth or shrinkage) SVC and 50 random patients classified as normal (no substantial) SVC

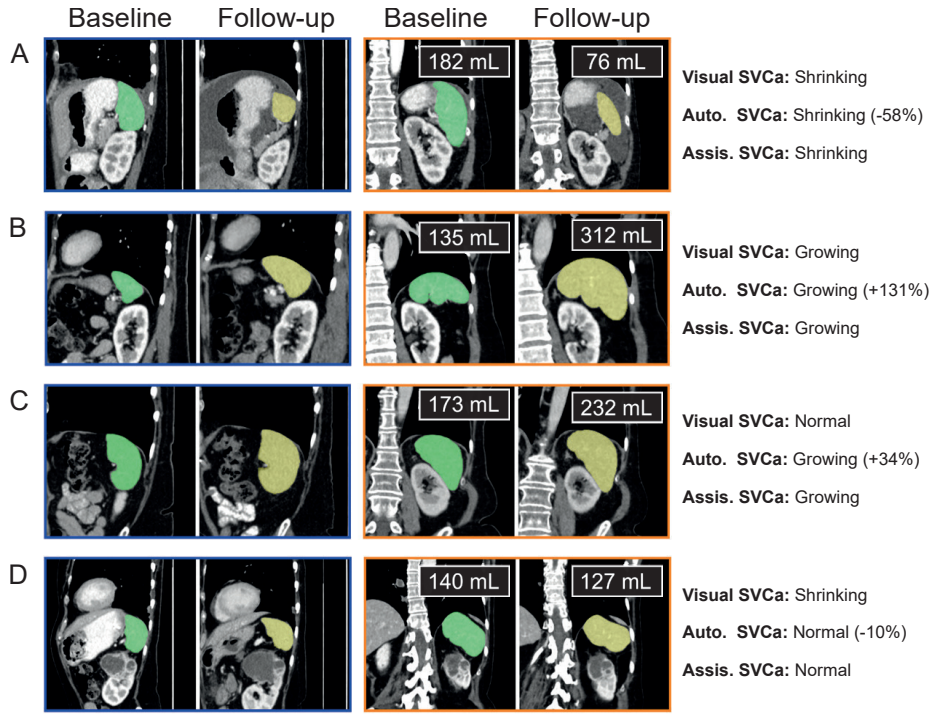


Figure 3.4: Examples of splenic volume change (SVC) classification of scan-pairs from dataset D_{100} (200 CT scans from 100 patients) in the qualitative observer experiment. Sections surrounded by blue and orange rectangles show the automatic segmentations in the sagittal and coronal orthogonal views, respectively. A, B, Scan-pairs in which the visual SVCa and automatic SVCa classification match. A, The scan-pair with the largest negative SVC. B, The scan-pair with the largest positive SVC in the dataset. C, D, Scan-pairs in which the visual SVCa and automatic SVCa classification differ. C, The radiologist classified the scan-pair as normal SVC in the visual SVCa but changed it to growth SVC in the assisted SVCa after seeing the segmentations produced by automatic SVCa (system B). D, Similarly, the radiologist classified the scan-pair as shrinking SVC in the visual SVCa but changed it to normal SVC in the assisted SVCa. All the sections show 230 $\times$ 230 mm and have a window center of 60 HU and a window width of 360 HU. SVCa = splenic volume change assessment.

after automatic SVCa. A fully random selection from dataset D to obtain dataset D_{100} would have resulted in a higher number of normal cases, which would have been less interesting for the observer experiment of our study.

Previous work is summarized in Table 3.2. The methods that used DL[64–66] were methods for multiorgan segmentation. None of the mentioned articles selected relevant cases from a large set of scans as we did in this study. Gibson et al.[66] obtained

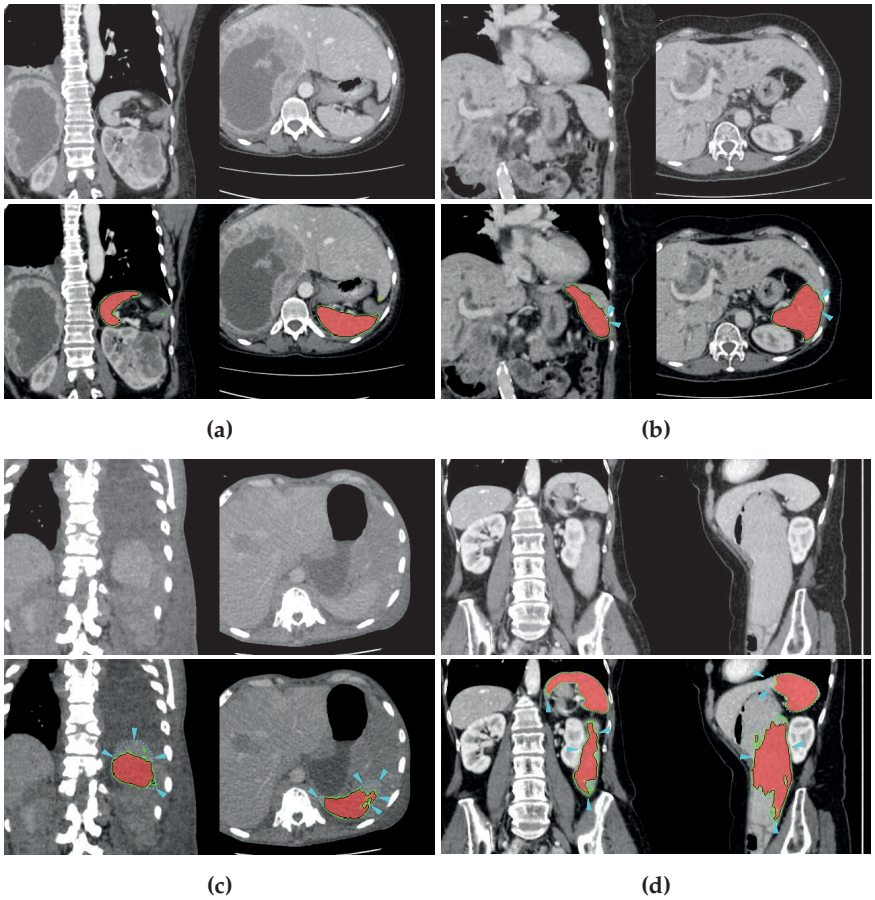


Figure 3.5: Examples of (a) excellent, (b) good, (c) bad, and (d) failed segmentation of the qualitative classification performed by the radiologist on dataset D_{100} (200 CT scans from 100 patients, spleen masks are unavailable) as part of the qualitative observer experiment. The figures show raw probabilities (before postprocessing) obtained by system B on dataset B_{300}. Red regions represent high probabilities ($P \geq 50\%$) of spleen presence. Green to transparent gradient regions represent low probabilities ($P < 50\%$) of spleen presence. The black contour around the raw probabilities represents the final output after postprocessing that is used to compute the splenic volume. The cyan dashed lines and triangles point to mistakes. Coronal and axial planes are shown, but (d) shows coronal and sagittal planes for better visualization. All images have a window center of 60 HU and a window width of 360 HU.

0.950 Dice score after applying transfer learning to improve their results. Linguraru et al.[55] used probabilistic atlas and registration to segment the liver and spleen (0.952 Dice). These methods were trained with less data (from 90 to 331 scans) than the data used for our best performing system (450 scans).

Our method showed reliable results; however, it had some limitations. For instance, patients with severe distortions in the body may obtain irregular automatic segmentations. Similarly, when the spleen is absent (splenectomy), this algorithm may segment a small false-positive region in the region where the spleen is usually located. These erroneous segmentations can be prevented by discarding candidates under a certain threshold of splenic volume. Although we qualitatively measured the performance of our algorithm in a large dataset, a quantitative measurement on a large set would help to validate our algorithm more thoroughly but requires a substantial annotation effort. Another limitation of this study was related to the selection of the 300 relevant scans from dataset B to obtain dataset B_{300} because this selection was performed by a single observer and another observer may have selected different cases. This may have introduced bias, but only affected the training of our algorithm and we expect this effect to be small. A trained medical student was used as the human independent observer for the quantitative validation, and an experienced radiologist may have performed slightly better. Finally, this algorithm was trained and evaluated using data from a single hospital. Future studies should focus on training and validation using multicenter data to increase the robustness of the algorithm.

In conclusion, fully automated spleen segmentation is feasible in complex scenarios such as oncologic follow-up. The performance of the DL algorithm was comparable to that of an independent observer on the test set. This method showed potential to help radiologists in classifying SVC accurately. Future studies are needed to investigate how this algorithm can affect the workflow of a radiologist and what effect it has on the overall scan interpretation. Future validation studies should include multicenter data and should be performed prospectively to test whether this algorithm can be safely and reliably used in clinical practice.

Appendix

Description of A, B, C, and D Datasets

Since the developed algorithm had not been tested before, we were not able to assume a performance level and as a result, we did not perform power calculations for this study. Datasets A, B, and C contain 100, 1000, and 50 CT scans respectively,

one CT scan per patient; these scans were randomly selected from the initial set of 2150 CT scans. Dataset B_{300} is a subset of dataset B and contains 300 relevant (50% challenging and 50% normal cases) CT scans. Datasets A and B300 were used for the development of the systems, while dataset C was used for testing purposes and performance measurements. Datasets A, B_{300}, and C were fully annotated. Dataset D was created by extracting 500 scan-pairs (follow-up scans) from 500 patients, from the initial set of 2150 CT scans, making a dataset of 1000 CT scans. This dataset was used for selecting 100 patients for dataset D_{100}, which will be used to perform the qualitative observer experiment. We made sure that each patient could only be part of one of our main four datasets (A, B, C, and D). Figure 3.2 shows the scan selection flowchart.

Network Inputs

During training, the network received the input set composed by $[I, R, \omega]$ and returned the probability map P, where I represents a patch from the CT scan, R a binary patch from the reference mask, and ω the weight-map. Every patch $(I, R, P,$ and $\omega)$ in the input set corresponds to the same center of gravity. The weight-map ω is computed from R to compensate for the frequency between background and foreground. During training, the network learns to reduce the difference between P and R using Softmax with weighted (ω) cross-entropy as the cost function. For every segmentation experiment, the dimension of I was $156 \times 156 \times 156$ while R, P and ω were $68 \times 68 \times 68$. For training, the input sets were selected based on the size of P, the stride among input sets was 34mm in all the orthogonal directions. We trained the network with positive (the spleen is present in R) and negative (the spleen is absent in R) input sets. We selected all the positive input sets and randomly selected (same number as the positive input sets) non-overlapping negative input sets. The final segmentation of a CT scan is obtained by providing input sets that consist of Is. Subsequently, the model returns Ps; the stride among input sets was 68 mm (no overlap). Ps were stitched together to produce the final segmentation output.

Metrics for Measuring Similarity between Predictions and Reference Masks

The following metrics were used to measure the similarity between the predictions and the reference masks.

$$\text{Dice score} = \frac{2 * volume(X \cap Y)}{volume(X) + volume(Y)}, \tag{3.1}$$

$$\text{Relative abs. vol. diff.} = 100 * absolute\left(1 - \frac{volume(X)}{volume(Y)}\right), \tag{3.2}$$

$$\text{Max. Hausdorff dist.} = max(max(SurfDist(X,Y)), max(SurfDist(Y,X))), \tag{3.3}$$

$$\text{95\% Hausdorff dist.} = max(Percentile_{95}(SurfDist(X,Y)), Percentile_{95}(SurfDist(Y,X)), \tag{3.4}$$

$$\text{Avg. symmetric surf. dist.} = \frac{mean(SurfDist(X,Y)) + mean(SurfDist(Y,X))}{2} \tag{3.5}$$

where X represents the post-processed output (final segmentation) of the network (Section 3.2.6) and Y represents the reference mask. The function $SurfDist(A, B)$ measures the minimum distance from a voxel of surface A to a voxel in surface B.

In practice, high Dice scores (close to one) represent a high overlap between X and Y. Metrics based on surface distances (maximum Hausdorff, 95% Hausdorff, and ASSD) measure the (max, 95% percentile, or average) distance in millimeters from Y to X; values close to zero are better. For clinical precise applications, such as guided surgery, metrics based on surface distances are more relevant than overlapping measurements.

Selection of relevant scans for automatic segmentation (Dataset B_{300})

System A was used to process all scans in dataset B (n = 1000) to find relevant cases. Since manual annotation of scans is time-consuming, we aimed to identify cases where the segmentation algorithm failed and add these scans as training data. A researcher visually classified the quality of the resulting segmentations on dataset B as good (small errors included), bad (failures included), and splenectomy. Predictions containing small errors of up to 10 mm from the boundaries of the spleen were classified as good, the remaining predictions were classified as bad or splenectomy. In total, 818 scans were classified as good, 150 scans as bad, and 32 scans as splenectomy. The 150 bad cases and 150 randomly selected good cases were selected as additional training data (see Select relevant scans based on final segmentations box in Figure 3.2). The main errors were mainly undersegmentation caused by abnormal spleen shape, beavertail liver, tumors in the surrounding structures, and large organs. We instructed our medical students to correct (if needed) the 300 predictions in a similar procedure as dataset A was annotated. These 300 cases composed dataset B_{300}.

Enriched dataset for the observer experiment (Dataset D_{100})

To create the enriched dataset for the qualitative observer experiment, we computed the SVC over time in dataset D (500 new patients) and selected 50 patients from

dataset D in which our system measured substantial SVC over time (either growth or shrinkage) and 50 random patients from dataset D with no substantial SVC (normal). These cases created dataset D_{100} (subset of dataset D, see Select enriched set for observer experiment in Figure 3.2), this dataset contains 100 patients, every patient has two scans for SVC analysis, having a total of 200 CT scans.

Modes of Splenic Volume Change Assessment (SVCa)

For visual SVC, the radiologist visually classified the spleen on the second scan as growing, shrinking, or normal (no substantial) SVC. Manual measurements and the splenic index equation were allowed when the SVC was not visually evident. For automatic SCVa, the measurements were fully automatic. For assisted SVCa, the radiologist was aided by the algorithm. We showed the 3D segmentations as an overlay, the automatically calculated volumes, and the growth percentage computed by our method to the radiologist. Based on this information, the radiologist classified the SVC again as growing, normal, or shrinking. If the segmentation of our algorithm was suboptimal, the radiologist would see this in the overlay and take this into account for their assessment. To get insight into the quality of the segmentations, we asked the radiologist to visually rate the quality of each segmentation as excellent (no oversegmentation neither undersegmentation), good (a minor error up to 5 mm), bad (oversegmentation or undersegmentation over 5mm), or failure (segmentation out of the spleen).

Implementation of the convolutional neural networks

The networks were implemented using Keras and Tensorflow in Python 3.6. The segmentation experiments were executed on a cluster environment with PCs equipped with NVIDIA GTX 1080 and 1080ti graphics cards and 256 GB of RAM. An epoch to train system A took 70 minutes while it took 6 hours for system B because of the larger training set. A system requires from 2 to 3 minutes, depending on the size of the thorax-abdomen CT scan, to process a full scan and get the final segmentation.

Kidney abnormality segmentation in thorax-abdomen CT scans

4

G.E. Humpire Mamani, N. Lessmann, E.Th. Scholten, M. Prokop, C. Jacobs, B. van Ginneken

Original title: Kidney abnormality segmentation in thorax-abdomen CT scans

Published in: arXiv:2309.03383, 2023

Abstract

In this study, we introduce a deep learning approach for segmenting kidney parenchyma and kidney abnormalities to support clinicians in identifying and quantifying renal abnormalities such as cysts, lesions, masses, metastases, and primary tumors. Our end-to-end segmentation method was trained on 215 contrast-enhanced thoracic-abdominal CT scans, with half of these scans containing one or more abnormalities.

We began by implementing our own version of the original 3D U-Net network and incorporated four additional components: an end-to-end multi-resolution approach, a set of task-specific data augmentations, a modified loss function using top-k, and spatial dropout. Furthermore, we devised a tailored post-processing strategy. Ablation studies demonstrated that each of the four modifications enhanced kidney abnormality segmentation performance, while three out of four improved kidney parenchyma segmentation. Subsequently, we trained the nnUNet framework on our dataset. By ensembling the optimized 3D U-Net and the nnUNet with our specialized post-processing, we achieved marginally superior results.

Our best-performing model attained Dice scores of 0.965 and 0.947 for segmenting kidney parenchyma in two test sets (20 scans without abnormalities and 30 with abnormalities), outperforming an independent human observer who scored 0.944 and 0.925, respectively. In segmenting kidney abnormalities within the 30 test scans containing them, the top-performing method achieved a Dice score of 0.585, while an independent second human observer reached a score of 0.664, suggesting potential for further improvement in computerized methods.

All training data is available to the research community under a CC-BY 4.0 license on `https://doi.org/10.5281/zenodo.8014289`.

4.1 Introduction

Kidney cancer is a significant global health issue, ranking as the 12^{th} most deadly cancer in the world, with an estimated 14,700 deaths in 2019 and approximately 73,820 new cases of kidney & renal pelvis cancer worldwide[3]. With the increasing number of cases, automated tools are needed to assist clinicians in managing this burden. For instance, by following nephrometry scoring systems[79], automatic kidney tumor segmentation methods may help specialists to detect and get reliable measurements of kidney tumors.

Previous research on kidney segmentation has employed a variety of conventional methods such as region growing[80,81], active shape models[82], active contours[83,84], graph cut[85,86], level-sets[87,88], snakes[89], random forest[90], and watersheds[91]. However, to the best of our knowledge, there are only a few methods that focus on segmenting kidney tumors or cysts in the literature. Linguraru et al.[83] proposed a semi-automatic method that combines fast marching and active geodesic contours to segment renal tumors. Kim and Park[81] used thresholds and histograms to segment the kidneys and applied texture analysis to the kidney parenchyma to find seeds for a region-growing algorithm to perform kidney tumor segmentation. Chen et al.[92] proposed a method to predict kidney tumor growth in mm^2/day, manually segmenting the kidney tumors and using a reaction-diffusion model to predict their growth. Kaur et al.[93] proposed an iterative segmentation method for renal lesions, which uses spatial image details and distance regularization.

In recent years, CNNs have shown to be more effective than traditional methods based on classical computer vision techniques and machine learning. Their ability to learn directly from raw data has led to their widespread use in segmenting organs and structures in different modalities. For instance, Zheng et al.[94] used an AlexNet-based method to localize the kidneys to define a seed for an active shape model algorithm to segment the kidneys in patients with either abdominal surgery or kidney tumors. Sharma et al.[95] used a network that takes the first 10 layers of the VGG-16 network and upsampled them in a decoder fashion to segment the kidneys of patients with renal insufficiency. Encoder-decoder networks such us 2D U-Net[8] and 3D U-Net[9] proved to be robust to tackle medical segmentation tasks in multiple medical imaging segmentation challenges[76,96]. Variants of these models have been extensively proposed and applied to a wide variety of tasks, including kidney segmentation. For instance, Taha et al.[97] segmented the artery, vein, and ureter around the kidneys using a 2D U-Net-like network that allows the deeper layers to influence more to the final prediction. Jackson et al.[98] used a 3D U-Net-like network to segment the kidneys. Moreover, several methods used deep learning to segment

kidney tumors[99,100]. Yu et al.[99] proposed Crossbar-Net, a network that segments kidney and kidney tumors and uses horizontal and vertical patches instead of traditional squared patches. The network is divided into sets of sub-networks; a set consists of a sub-network for vertical and another for horizontal patches. Yang et al.[100] proposed a 3D CNN using a pyramid pooling module to segment the kidneys and kidney tumors in abdominal CT angiographic scans.

The top competitors of the Medical Decathlon[76] and LiTS challenge[68,101] have achieved the highest performance using cascaded networks. These networks divide the tasks into sub-tasks, with one network per sub-task. These networks have different fields of view and thus complement each other, resulting in higher performance. For instance, a first network may segment the liver and the liver tumor as a single structure, aiming to determine the region of interest for the second network; the second network then aims to segment the liver tumor class only. Similarly, Blau et al.[102] used cascade networks to segment the kidney and kidney cyst in CT scans using a 2D U-Net. Their method used heuristics such as a distance transform and HU thresholding to select cyst candidates within the kidney region. A second (shallow) network classified whether a candidate represented a kidney cyst. Additionally, Haghighi et al.[103] used a localization network for pre-processing, which cropped the input for 3D U-Net to segment MRI images of the kidneys. In a recent challenge on segmentation of the kidney and kidney tumors on CT[104], nnUNet[76] was the best performing method. This method automatically adapts its hyperparameters based on a fingerprint of the data, resulting in optimal performance. Furthermore, it uses 5-fold cross-validation to obtain the final prediction.

In this study, we propose an automatic method for segmenting the kidney parenchyma and kidney abnormalities in thorax-abdomen CT scans and compare it with the nnUNet. We trained our method on 215 thorax-abdomen CT scans and tested on additional 50 scans; the dataset consisted of scans from patients undergoing oncological workup. The dataset contains patients at different stages of disease and therefore abnormalities can be present in multiple body regions.

4.2 Materials and Methods

4.2.1 Patient Data

The dataset used in this study was collected from the Radboud University Medical Center, Nijmegen, the Netherlands. We randomly retrieved 1905 studies from 929 patients referred by the oncology department in a 12 month period. These patients did not opt-out for use of their data for research, Protected health information

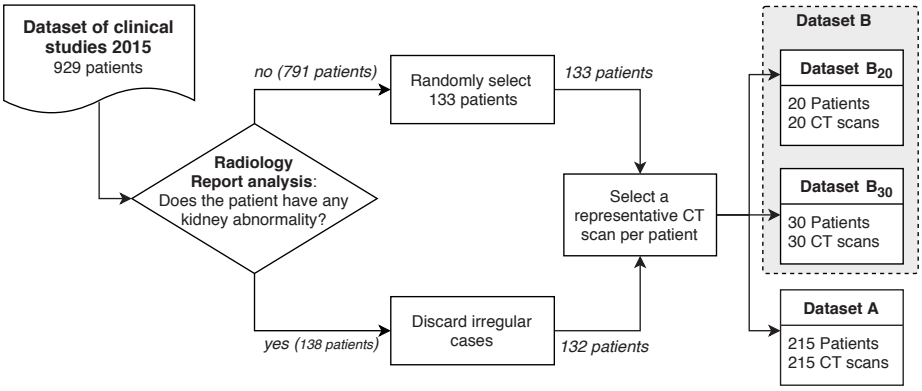
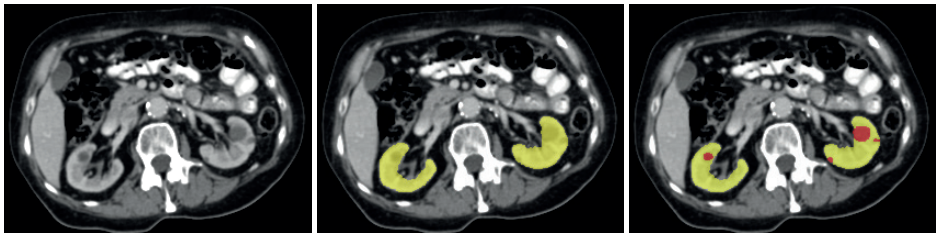


Figure 4.1: Diagram of the CT scans selection criteria for this study, with dataset A for training and datasets B_{30} and B_{20} for testing (with and without kidney abnormalities respectively).

was removed from the DICOM data. This retrospective study was approved by the medical-ethical review board of the hospital. CT scanners from two manufacturers were used to acquire the CT scans: Toshiba (Aquilion One) and Siemens (Sensation 16, Sensation 64, and Somatom Definition AS). The reconstruction kernels were FC09, FC09-H, B30f, B30fs, and I30f. The slice thickness ranged from 0.5 to 3 millimeters, 90% of them between 1 and 2 mm. Severe abnormalities throughout the body are present in this dataset resulting from disseminated disease, surgery, chemotherapy, radiotherapy, etc.

We selected a subset to perform our experiments; the procedure is summarized in Figure 4.1. We analyzed the radiology reports per study to intentionally select potential cases that contain kidney abnormalities such as cysts, lesions, masses, metastases, and tumors. In Dutch: *((‘cyste’ OR ‘cysten’), (‘laesie’ OR ‘lesies’), ‘massa’, (‘metastase’ OR ‘metastasen’), and ‘tumor’)*. Our selection criteria selected studies where the radiology report mentioned in the same sentence the kidneys *(’nier’ OR ’nieren’ NO ’bijnier’)* and any kidney abnormalities. Furthermore, only one clinical study per patient was selected to get a large variety of anatomies for the segmentation task. In case multiple studies for the same patient were found, we selected the study with the earliest acquisition date.

We employed a radiology report analysis to curate a dataset of 138 clinical studies from 138 patients with kidney abnormalities, including cysts, lesions, masses, metastases, or tumors. We excluded six patients with unusual anatomy, three patients who had received kidney transplants, two patients with kidneys of irregular size, and one patient with a horseshoe kidney. The inclusion and exclusion criteria



(a) Input CT scan **(b)** Annotations format 1 **(c)** Annotations format 2

Figure 4.2: Example illustrating the different annotation formats. Each subfigure shows the same axial section, with overlays depicting the annotations: (a) shows the axial CT section. (b) shows the annotations in format 1: parenchyma and kidney abnormalities as a single structure (yellow overlay). (c) shows the annotations in format 2: parenchyma (yellow overlay) and kidney abnormalities (red overlay) as different structures. All images have a window center of 60 HU and a window width of 360 HU.

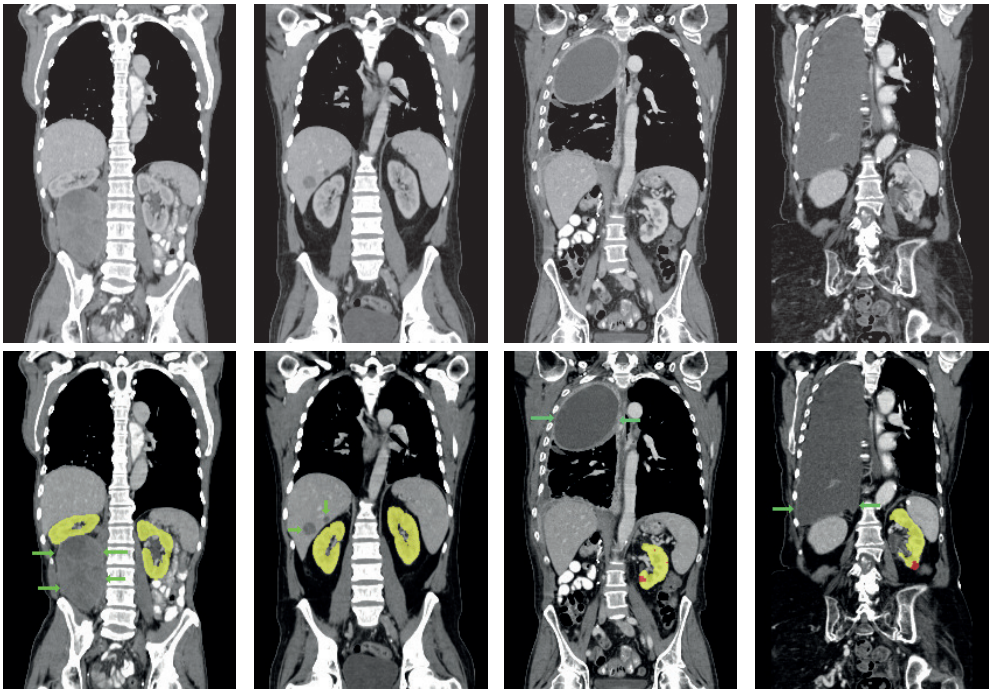
gave us 132 cases for analysis, which were then balanced with additional 133 random patient studies without kidney abnormalities, for a total of 265 CT scans from 265 patients. The patient cohort contains 56% males; the average age was 60 years, and the age ranged from 22 to 84. We divided this set into 215 CT scans for training (dataset A) and 50 for testing (dataset B). The test set was further subdivided, with 60% (30/50) containing abnormalities (dataset B_{30}) and the remaining 40% (20/50) devoid of abnormalities (dataset B_{20}). The distribution of the five types of abnormalities (tumors, cysts, masses, lesions, and metastases) was proportional among the 30 cases in dataset B_{30} (six cases per abnormality), which were randomly selected.

In the test set, two and six patients had undergone left and right nephrectomy, respectively, while the training set included seventeen and eighteen patients who had undergone left and right nephrectomy, respectively.

4.2.2 Annotation procedure

Four medical students manually segmented the kidney's parenchyma and kidney abnormalities. They were trained by an experienced radiologist (EthS) and consulted the radiologist whenever needed throughout the annotation process. Adhering to a standardized protocol, the medical students annotated the kidney parenchyma as the region composed of the renal cortex, renal medulla, and renal pyramid. The renal hilum, collecting system, and (major and minor) calyces were excluded as much as possible from the kidney parenchyma annotations. We grouped cysts, lesions, masses, metastases, and tumors connected to the kidney parenchyma as kidney ab-

normalities. The protocol excluded cases with abnormalities in the collecting system.



(a) Patients without kidney abnormalities. **(b)** Patients with kidney abnormalities.

Figure 4.3: Four examples of CT scans from the training set (dataset A) showing coronal sections with annotations in format 2 (see Figure 4.2c) where yellow and red overlays represent annotations of the parenchyma and kidney abnormalities, respectively. Note that all the patients have anomalies in the body (green arrows in the body), and both cases of (b) have only one kidney and contain kidney abnormalities. All the slices have a window center of 60 HU and a window width of 360 HU.

Annotators used an in-house tool based on MeVisLab[105] to fully delineate the contours of the structures in 2D orthogonal planes. Our tool was designed to reduce the manual annotation time by interpolating unannotated contours between two manually delineated contours. The kidney parenchyma of the training set was annotated using an active learning process, with medical students correcting the kidney parenchyma predictions made by a pre-trained 3D U-Net (it used 50 CT scans from dataset A); the kidney abnormalities were annotated from scratch. The test set was manually annotated (i.e. the contour interpolation option of our tool was disabled) by two medical students. One of these was considered as the reference standard and the other one as the second observer ■. The latter was the most experienced

among the medical students and was not allowed to consult the experienced radiologist during these annotations. The annotations of the second observer ■ served as a benchmark for human performance. The annotations were initially obtained in the axial plane, followed by corrections in coronal and sagittal planes to keep the annotation consistent in all orthogonal directions.

This study utilized two annotation formats, format 1 and format 2, to store the annotations. Format 1 considers the kidney parenchyma and kidney abnormalities as a single class (see Figure 4.2b) while format 2 separates them into two classes (see Figure 4.2c).

Samples of CT scans from patients included in this study can be seen in Figure 4.3. While Figure 4.3a depicts patients without kidney abnormalities, it highlights the presence of abnormalities in other parts of the body, such as liver tumors. Figure 4.3b shows patients with kidney abnormalities, as well as other abnormalities in the body, such as nephrectomy and collapsed lung.

4.2.3 Segmentation network

We present an end-to-end method for segmenting renal parenchyma and abnormalities in CT scans. We depict our architecture in Figure 4.4. It consists of two segmentation networks, a multi-resolution network for kidney segmentation (annotations in format 1, one voxel represents 4×4×4mm) and a high-resolution network (annotations in format 2, one voxel represents 1×1×1mm). The multi-resolution network is designed to first provide a rough localization of the kidney by processing a low-resolution version of the CT scan. This defines an ROI for the high-resolution network to refine the segmentation of the kidneys and kidney abnormalities.

Pre-processing

The CT scans and annotations were resampled to $1\times1\times1mm$ (for high-resolution segmentation using annotations in format 2) and $4\times4\times4mm$ (for multi-resolution segmentation using annotations in format 1) resolutions (see Figure 4.4a). Scans and annotations were resampled using cubic and nearest-neighbor interpolation, respectively. We clipped the Hounsfield Units to the range [-500,400].

Multi-resolution network

We present an end-to-end cascade method for parenchyma and kidney abnormality segmentation. Unlike traditional cascade networks, which use two separate networks and do not allow for backpropagation, our approach uses a single network

composed of two sub-networks. The first sub-network is a 3D U-Net with 16 filters that performs multi-resolution segmentation and defines an ROI. This network takes 3D patches of 108×108×108 voxels, with each voxel representing 4×4×4 mm, as input using annotations in format 1 (kidney parenchyma + kidney abnormalities) and outputs 20×20×20 voxels. The output is then up-sampled 4 times and padded with zeros to match and mask out the high-resolution input image in millimeters (108×108×108 mm, one voxel represents 1×1×1 mm). The masked-out image serves as an additional input to the second sub-network, the high-resolution segmentation network, which uses a 3D U-Net with 32 filters and serves to fine-segment the kidneys and kidney abnormalities (see Figure 4.4b). Figures 4.4a and Figure 4.4b illustrate our approach and the connection between the multi-resolution and the high-resolution segmentation network, respectively.

Data augmentation

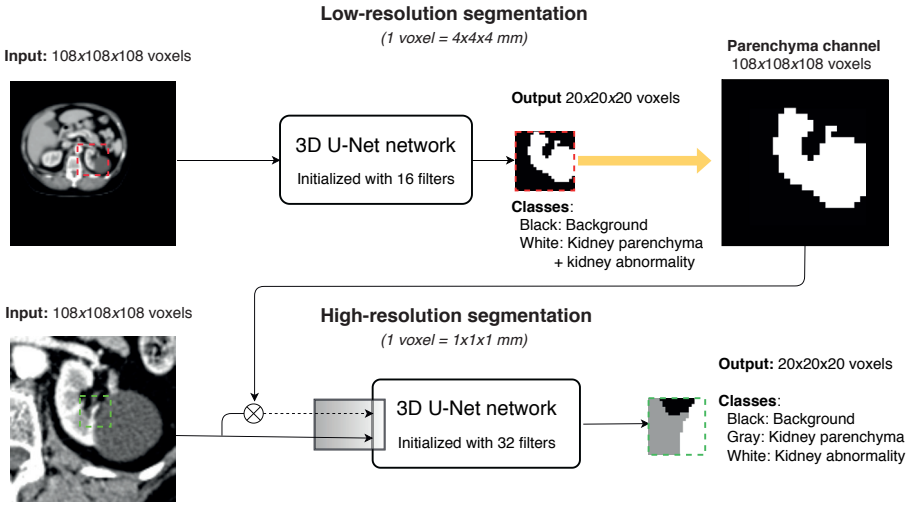
Data augmentation was applied randomly to 70% of the training samples using scaling, rotation, Gaussian blurring, image intensity variation, and elastic deformation. Up to three of these data augmentation methods were applied randomly to each training sample, to prevent too much data distortion. When elastic deformation was used, it was only performed in conjunction with Gaussian blurring and image intensity variation. Interpolation methods of cubic and nearest neighbor were used for CT scans and reference standards, respectively. The scaling factor ranged from 0.95 to 1.05, with rotations of up to two planes of -5° to 5° degrees. Gaussian blurring had a sigma range of 0.2 to 1.0, and image intensity variation varied between -20 and 20 HU. We performed elastic deformation by placing ten control points in a grid, randomly perturbed by up to 5 voxels that were used as input to cubic B-spline interpolation.

Spatial dropout

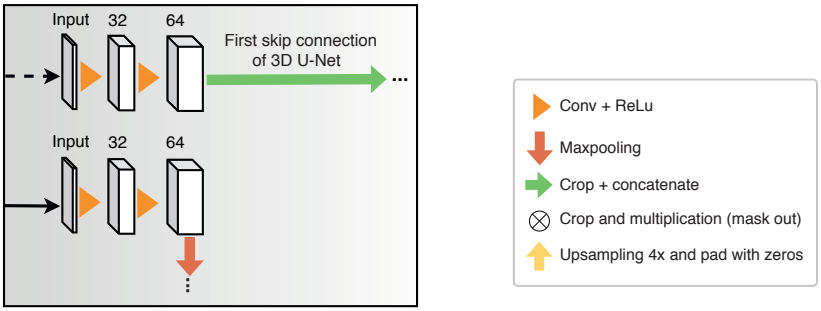
We applied spatial dropout[106], a regularization technique that is different from traditional dropout. Spatial dropout drops feature maps instead of individual neurons to enforce independence among feature maps, encouraging the network to learn more robust and generalizable features. We randomly dropped 10% of the feature maps per layer.

Loss function

The loss function determines how the network's weights are optimized after a forward pass. In our experiments, we used a combination of weighted categorical cross-



(a) Diagram of the proposed multi-resolution network.



(b) Input of the high-resolution segmentation.

Figure 4.4: (a) Diagram of the proposed network. The multi-resolution segmentation network uses a 3D U-Net network initialized with 16 filters. It processes blocks of 108×108×108 voxels and outputs the central 20×20×20 voxels (represented by the dashed red square). One voxel corresponds to a resolution of 4×4×4mm, giving the network a receptive field of 88×88×88 voxels or 352×352×352mm. The kidney parenchyma and kidney abnormalities are considered a single class in the multi-resolution network (see Figure 4.2b). The high-resolution segmentation network uses a 3D U-Net architecture initialized with 32 filters, with each voxel representing 1×1×1mm. Its receptive field is 88×88×88 mm and it segments the parenchyma and the kidney abnormalities as different classes (see Figure 4.2c). (b) Shows how the multi-resolution and the high-resolution networks are connected.

entropy and dice loss in the experiments.

$$\text{Combined loss} \quad = \quad \alpha * diceLoss + \gamma * TopK(weightedCrossentropy) \qquad (4.1)$$

where $\alpha = 0.3$ and $\gamma = 0.7$ were used in all the experiments. Top-k[107] sorts the voxel-wise loss in descending order and keeps the top $k\%$ to compute the final mean loss; this approach emulates an online voxel-wise hard-mining per sample.

Post-processing

The output of the networks was post-processed to eliminate false positives. The end-user prediction was reconstructed by stitching together the predictions. In all the networks, the output was thresholded at 0.5 to get a binary prediction. The predictions of the multi-resolution network were up-sampled four times and dilated five times to mask out the predictions of the high-resolution segmentation network. Only the kidney abnormalities that were connected to the kidney parenchyma were kept, to ensure that there were no spurious kidney abnormality candidates outside the kidney region.

CNN Settings

Due to the large footprint of the network, scans were divided into 3D patches to train the 3D network. Each training sample consisted of a patch of 108×108×108 voxels from the CT scan and a 20×20×20 voxel reference standard. During training, the reference standard patches were sampled every ten voxels in all the orthogonal planes with up to 50% overlap among surrounding patches. During inference, the cubes do not overlap. Patches at the border of the CT scan were mirrored to match the input network size.

The Glorot uniform algorithm[108] was used to initialize the weights of the network. The weight-map w compensated for the high-class imbalance between the classes. The background, parenchyma, and kidney abnormality classes had empirically defined weights of 0.05, 0.10, and 0.99, respectively. We used Adam[109] as optimization function with learning rate$= 0.00001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The training stopped when the performance on the validation set stopped improving for ten epochs, and the model with the highest average Dice score on the validation set was selected as the optimal model.

Implementation of the CNN

The networks were implemented using Keras and TensorFlow as backend in Python 3.6. The segmentation experiments were executed on a cluster of computers equipped

with GTX1080 and GTX1080ti graphics cards, each with 256GB of CPU RAM.

4.2.4 Evaluation

The end-user segmentation obtained by our networks was compared to the reference masks using the Dice score.

$$\text{Dice score} = \frac{2 * volume(X \cap Y)}{volume(X) + volume(Y)} \tag{4.2}$$

where X is the prediction, and the Y is the reference standard.

4.2.5 Ablation study

In this section, we conducted a step-by-step evaluation of the impact of each module (multi-resolution, data augmentation, top-k, and spatial dropout) in our proposed network. The backbone architecture for this ablation study was the 3D U-Net[9]. Our experiments setup started with a 3D U-Net, and additional modules were added one by one in subsequent experiments (see the left side of Table 4.1). In order to evaluate the impact of each module on the network performance, we conducted an ablation study by adding modules to the 3D U-Net backbone architecture one by one. The baseline network, referred to as experiment 5 ■, only used the 3D U-Net initialized with 32 filters and had a single input of 108×108×108 voxels with 1×1×1 mm per voxel, producing 20×20×20 voxels. The subsequent experiments added the multi-resolution module (experiment 4 ■), data augmentation module (experiment 3 ■), top-k module (experiment 2 ■), and spatial dropout module (experiment 1 ■) to the network. The input and output sizes and formats were consistent across all experiments except experiment 5 ■; networks receive two inputs of 108×108×108 voxels each, one input of 1×1×1 mm and one input of 4×4×4 mm per voxel for high-resolution (input of 108×108×108 mm using annotation format 2) and multi-resolution segmentation (input of 432×432×432 mm using annotation format 1), respectively. The difference in performance between experiment 1 ■ (experiment with spatial dropout) and experiment 2 ■ (experiment without spatial dropout) showed the influence of the spatial dropout module, for example. As an initial step, we first trained the multi-resolution module independently to reach its optimal sub-model. Afterward, we froze the weights of the multi-resolution sub-model, except for the last three layers to allow back-propagation from the high-resolution segmentation network. All the experiments used 80% of dataset A for training and 20% for validation. Each experiment was trained independently to find the optimal model. The best model from each experiment was evaluated using test sets B_{20} and B_{30}.

4.2.6 nnUNet

We conducted experiments with nnUNet[76] to compare its performance with our methods. Unlike our approach, nnUNet processes CT scans without any preprocessing step, while we resample the CT scans to an isotropic resolution and clip the HU range. To gain insight about the benefits of ensemble networks, we ensembled nnUNet with our two highest-performing methods, one at a time. As nnUNet only uses thresholding as postprocessing, we analyzed the impact of our dedicated postprocessing method on performance. Note that our postprocessing eliminates false-positive kidney abnormalities that are not attached to the parenchyma.

4.3 Results

The results of the ablation study conducted on the test sets (dataset B_{20} and B_{30}) are shown in Figure 4.5. These results are also summarized in Table 4.1, which includes asterisks (*) to indicate statistical significance (P-value < 0.05) between experiment 1 ■ and other experiments, as determined by a two-tailed Mann-Whitney U test. We evaluated the predictions of each experiment per class to show more insights into the results of our experiments. Furthermore, we combined the prediction of both classes (annotation format 2) as a single structure (annotation format 1) and computed its Dice score; this helps to make our results comparable to methods that reported kidney dice only.

Dataset B_{30}: The presence of kidney abnormalities characterizes the patients in this dataset (see Figure 4.3b). The results of our experiments on dataset B_{30} are displayed in Figures 4.5d, 4.5b, and 4.5c. First, we evaluated the performance of the methods in segmenting the kidney abnormalities class only. The results are shown in Figure 4.5d and in the column "Dataset B_{30}/Abnormalities class" of Table 4.1. The second observer ■ and experiment 1 ■ achieved the two highest scores, 0.664±0.274 and 0.487±0.314, respectively. Experiment 5 ■ obtained 0.390±0.315 Dice, the lowest score when segmenting the kidney abnormalities only. Next, we evaluated the performance of the methods in segmenting the parenchyma class only. The results are shown in Figure 4.5b and in the column "Dataset B_{30}/Parenchyma class" of Table 4.1. The two highest scores were obtained by Experiment 2 ■ and experiment 4 ■ with 0.938±0.051, 0.936±0.058, respectively, while the second observer ■ obtained the lowest score with 0.925±0.051. Finally, we evaluated the performance of the methods when segmenting both the parenchyma and the kidney abnormalities class as a single structure (annotation format 1). The results are shown in Figure 4.5c and in column "Dataset B_{30}/Parenchyma + abnormalities class" of Table 4.1. The two

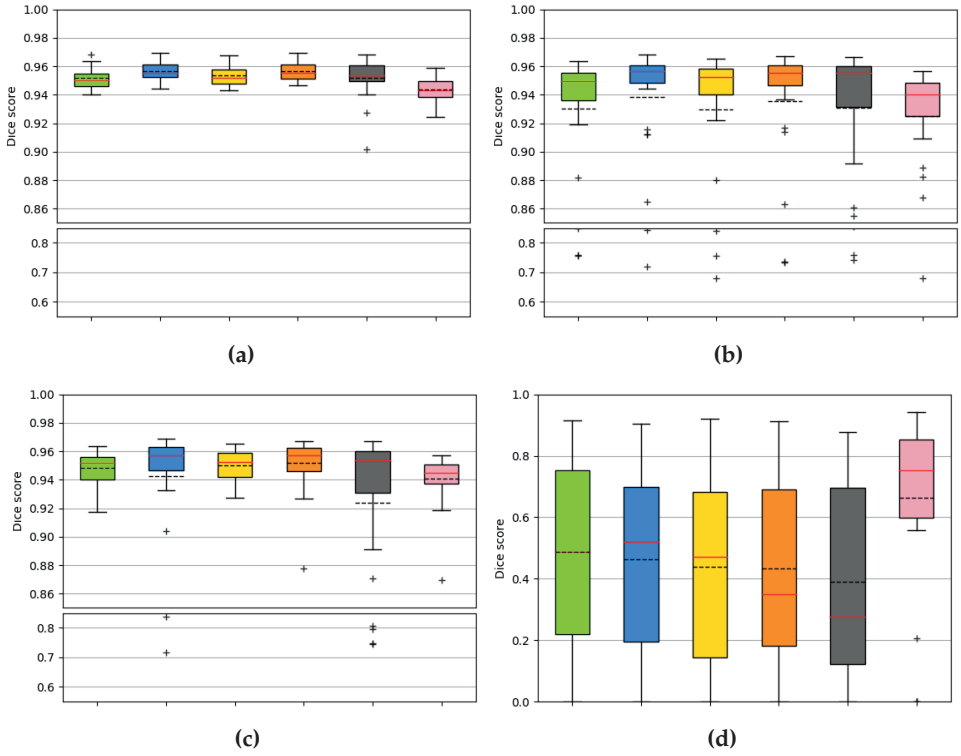


Figure 4.5: Performance comparison of our methods and the second observer ■ on datasets B_{20} and B_{30} using boxplots. The red and black lines represent the median and the mean, respectively. Boxplot (a) shows results for class parenchyma only on the dataset B_{20} (twenty cases without abnormalities). Boxplot (b) shows results for class parenchyma only on the dataset B_{30} (thirty cases with abnormalities). Boxplot (c) displays the results for class parenchyma plus abnormalities as a single structure on dataset B_{30} (thirty test cases with abnormalities). Boxplot (d) shows results for Class abnormalities only on the dataset B_{30} (thirty cases with abnormalities). Note that the scale in the y-axis is different for boxplot (d). The modules for each experiment are represented by the same color coding as in Table 4.1: experiment 1 ■, experiment 2 ■, experiment 3 ■, experiment 4 ■, experiment 5 ■, and second observer ■ .

highest scores were achieved by Experiment 4 ■ and experiment 3 ■ with Dice scores 0.952±0.017 and 0.950±0.010, respectively. Experiment 5 ■ obtained the lowest Dice score with 0.924±0.065.

Dataset B_{20}: The patients in this dataset do not present kidney abnormalities, but it is probable that they have other anomalies in the body (see Figure 4.3a). The results on the test set B_{20} are depicted in Figure 4.5a and in Table 4.1 under the column

Table 4.1: Summary of performance of our methods and the second observer ■ on the test sets (dataset B_{20} and B_{30}) using Dice score. The upper part of the table shows Mean and SD, while the bottom part shows 95% Confidence Intervals. We analyzed the performance per class (parenchyma and kidney abnormalities) and when both classes are combined (parenchyma + kidney abnormalities). An asterisk (*) indicates that a value is significantly better or worse than the value of the reference system (experiment 1 ■). The best Dice score values per column are in bold.

Experiment	Test set: dataset B_{20}		Test set: dataset B_{30}	
	Parenchyma class (Fig. 4.5a)	**Parenchyma class** (Fig. 4.5b)	**Parenchyma + abnormalities class** (Fig. 4.5c)	**Abnormalities class** (Fig. 4.5d)
		Mean ± SD		
Exp. 1 ■	0.952±0.008	0.930±0.053	0.948±0.011	0.487±0.314
Exp. 2 ■	*0.957±0.006	*0.938±0.051	*0.942±0.049	0.464±0.320
Exp. 3 ■	0.954±0.007	0.930±0.064	0.950±0.010	0.440±0.310
Exp. 4 ■	*0.956±0.007	0.936±0.058	*0.952±0.017	0.434±0.311
Exp. 5 ■	0.952±0.015	0.931±0.057	0.924±0.065	0.390±0.315
Second observer ■	*0.944±0.009	*0.925±0.051	*0.941±0.017	**0.664±0.274**
nnUNet	*0.960±0.015	*0.940±0.069	*0.931±0.099	0.521±0.303
Ens. nnUNet Exp. 1 ■	*0.962±0.007	*0.941±0.058	*0.927±0.125	0.526±0.305
Ens. nnUNet Exp. 2 ■	*0.964±0.006	*0.944±0.055	*0.928±0.129	0.507±0.318
nnUNet PP	*0.964±0.006	*0.941±0.068	*0.955±0.042	0.576±0.290
Ens. nnUNet Exp. 1 ■ PP	*0.962±0.007	*0.942±0.058	***0.960±0.009**	0.585±0.293
Ens. nnUNet Exp. 2 ■ PP	***0.965±0.006**	***0.947±0.050**	*0.958±0.032	0.566±0.309
		Confidence Intervals 95%		
Exp. 1 ■	0.948-0.955	0.910-0.950	0.945-0.952	0.369-0.604
Exp. 2 ■	0.954-0.960	0.920-0.957	0.924-0.961	0.345-0.584
Exp. 3 ■	0.950-0.957	0.906-0.954	0.946-0.954	0.324-0.555
Exp. 4 ■	0.953-0.959	0.914-0.957	0.945-0.958	0.318-0.550
Exp. 5 ■	0.945-0.959	0.910-0.952	0.900-0.948	0.273-0.508
2nd observer	0.939-0.948	0.906-0.944	0.935-0.947	0.561-0.766
nnUNet	0.953-0.965	0.913-0.961	0.892-0.959	0.411-0.625
Ens. nnUNet Exp. 1 ■	0.959-0.965	0.918-0.958	0.876-0.959	0.416-0.627
Ens. nnUNet Exp. 2 ■	0.961-0.967	0.923-0.961	0.876-0.961	0.397-0.613
nnUNet PP	0.961-0.967	0.915-0.962	0.938-0.965	0.469-0.673
Ens. nnUNet Exp. 1 ■ PP	0.959-0.966	0.918-0.959	0.957-0.963	0.482-0.687
Ens. nnUNet Exp. 2 ■ PP	0.962-0.967	0.926-0.963	0.945-0.965	0.454-0.672

"Dataset B_{20}/Parenchyma class". Experiment 2 ■ and experiment 4 ■ obtained the highest Dice scores, 0.957±0.006 and 0.956±0.007, respectively. The second observer ■ obtained the lowest Dice score with 0.944±0.009.

nnUNet: In our experiments, nnUNet obtained slightly better results in the parenchyma class of datasets B_{20} and B_{30} compared to our experiments, a Dice score of 0.521 ± 0.303 in the kidney abnormality class, which was higher by +0.034 Dice than our experiment 1 ■. To further analyze the differences between nnUNet and our experiments, we ensembled the predictions of nnUNet with either experiment 1 ■

or experiment 2 ■ by averaging their probabilities. The ensemble nnUNet with experiment 2 ■ slightly improved the results of nnUNet in the parenchyma class of both datasets but decreased in -0.014 Dice score in the abnormality class, while the ensemble nnUNet with experiment 1 ■ slightly improved in +0.004 dice score compared to nnUNet in the abnormality class. The ensemble nnUNet with experiment 2 ■ performed slightly better than the ensemble nnUNet with experiment 1 ■ in all classes, except the abnormality class, where the ensemble with experiment 1 ■ had a Dice score of 0.526 ± 0.306, and the ensemble with experiment 2 ■ obtained 0.507 ± 0.318. Since nnUNet only uses thresholding for post-processing, we applied our dedicated post-processing to the nnUNet predictions to remove kidney abnormalities that are not attached to the kidney, which resulted in notable improvements of +0.055, +0.059, and +0.059 for nnUNet, ensemble nnUNet with experiment 1 ■, and ensemble nnUNet with experiment 2 ■, respectively. As a result, the ensemble nnUNet with experiment 1 ■ and our dedicated post-processing was the highest-performing experiment in the abnormality class, with a Dice score of 0.585 ± 0.293.

Table 4.2 compares our results with other methods published in the literature. Some of the methods report the Dice scores for the left and right kidneys separately, while others report a single score for both kidneys combined. To make our results comparable to these methods, we post-processed our predictions to obtain the Dice scores for both the left and right kidneys.

4.4 Discussion

In this paper, we presented an automatic method for the segmentation of the (kidney) parenchyma and kidney abnormalities. We conducted experiments in an ablation study fashion to evaluate the contribution of each module to the performance (see Section 4.2.5). For instance, the comparison between experiment 5 ■ and experiment 4 ■ in Figure 4.5 shows the influence of the multi-resolution module. Figure 4.5a shows that all of our experiments outperformed the second observer ■ when segmenting the kidney parenchyma in dataset B_{20} (patients without kidney abnormalities). While the presence of kidney abnormalities affected the performance of kidney (parenchyma + abnormalities) segmentation; see the difference of outliers between Figure 4.5a (dataset B_{20}) and Figure 4.5c (dataset B_{30}: patients with kidney abnormalities). One of the reasons for this behavior may be the difficulty in defining the boundary between the parenchyma and the kidney abnormality. When comparing the boxplots, the interquartile range of Experiment 5 ■ and experiment 2 ■ obtained the largest and the smallest interquartile range, respectively, indicating that the combination of multi-resolution, data augmentation, and top-k modules positively im-

Table 4.2: Performance comparison between our methods (experiment 1 ■ and experiment 2 ■) and previous work of kidney segmentation and kidney tumor/abnormality segmentation using the mean Dice score as the metric. The methods listed below the line reported the presence of cases with kidney tumors/abnormalities in their datasets. The values marked with † are obtained after post-processing the prediction masks from 'Both kidneys' column for comparability with other methods.

Method	Kidney abnormalities	Kidneys: Left kidney	Kidneys: Right kidney	Kidneys: Both kidneys	Num. test scans	Deep Learning	Description
Methods that reported cases without kidney abnormalities or did not report them							
Jackson et al. [98]	N/A	0.860	0.910	-	24	Yes	
Gibson et al. [66,110]	N/A	0.930	-	-	*10*	Yes	*9-folds cross-validation.*
Badura and Wieclawek [88]	N/A	0.938	0.944	-	20	No	
Heinrich et al. [111]	N/A	0.942	-	-	*10*	Yes	*4-fold cross validation.*
Wang et al. [112]	N/A	0.956	0.954	-	*30*	Yes	*4-fold cross-validation.*
Khalifa et al. [90]	N/A	-	-	**0.973**	*60*	No	*Leave-one-out.*
Exp. 1 ■	N/A	0.953†	0.951†	0.952	20	Yes	Test set: Dataset B_{20}, train set: Dataset A.
Exp. 2 ■	N/A	0.960†	**0.955**†	0.957	20	Yes	Test set: Dataset B_{20}, train set: Dataset A.
Second observer ■	N/A	0.943†	0.945†	0.944	20	N/A	Test set: Dataset B_{20}, train set: N/A.
nnUNet	N/A	0.956†	0.959†	0.960	20	Yes	Test set: Dataset B_{20}, train set: Dataset A.
Ens. nnUNet Exp. 1 ■ PP	N/A	0.962†	0.944†	0.962	20	Yes	Test set: Dataset B_{20}, train set: Dataset A.
Ens. nnUNet Exp. 2 ■ PP	N/A	**0.964**†	0.946†	0.965	20	Yes	Test set: Dataset B_{20}, train set: Dataset A.
Methods that reported cases with kidney abnormalities							
Turco et al. [87]	-	-	-	0.800	55	No	Polycystic kidneys only.
Sharma et al. [95]	-	-	-	0.860	81	Yes	Polycystic kidneys.
Skalski et al. [84]	-	-	-	0.862	10	No	Kidney cancer.
Blau et al. [102]	-	0.870	0.870	-	46	Yes	Cysts.
Lin et al. [80]	-	0.873	0.886	-	30	No	Two cases with tumor, one with a cyst.
Zheng et al. [94]	-	0.890	0.920	-	78	Yes	Kidney tumors.
Wieclawek [91]	-	-	-	0.917	170	No	Cysts and kidney tumors.
Yang et al. [100]	0.802	-	-	0.931	50	Yes	Kidney tumors.
Yu et al. [99]	**0.913**	-	-	-	36	Yes	Kidney tumors.
Exp. 1 ■	0.488	0.949†	0.951†	0.948	30	Yes	Test set: Dataset B_{30}, train set: Dataset A.
Exp. 2 ■	0.464	0.956†	0.936†	0.942	30	Yes	Test set: Dataset B_{30}, train set: Dataset A.
Second observer ■	0.664	0.939†	0.943†	0.941	30	N/A	Test set: Dataset B_{30}, train set: N/A.
nnUNet	0.521	0.928†	0.951†	0.931	30	Yes	Test set: Dataset B_{30}, train set: Dataset A.
Ens. nnUNet Exp. 1 ■ PP	0.585	0.960†	**0.960**†	**0.960**	30	Yes	Test set: Dataset B_{30}, train set: Dataset A.
Ens. nnUNet Exp. 2 ■ PP	0.566	**0.963**†	0.955†	0.958	30	Yes	Test set: Dataset B_{30}, train set: Dataset A.

pacted the segmentation of the kidneys (parenchyma + abnormalities). Note the spatial dropout module (difference between experiment 1 ■ and experiment 2 ■) was beneficial only to the kidney abnormality class (see Figure 4.5). Furthermore, Figure 4.5d shows that the mean Dice score (black dashed line in boxplots) of our experiments gradually increases when adding more modules (experiment 5 ■ to experiment 1 ■) when segmenting the kidney abnormality class. This highlights the positive impact of each module in this ablation study on the segmentation of kidney abnormalities.

Additionally, we trained nnUNet, a state-of-the-art segmentation method, on our data and obtained results that were consistent with our previous experiments, except for the kidney abnormality class where nnUNet achieved a 0.521 Dice score compared to 0.488 obtained by experiment 1 ■. To explore further improvements, we combined nnUNet predictions with our best-performing experiments, resulting in an ensemble nnUNet + experiment 1 ■ that achieved 0.526 Dice score for the kidney abnormality class. Since nnUNet uses only thresholding as postprocessing, we investigated whether postprocessing nnUNet predictions with our dedicated postprocessing could result in better performance. This additional postprocessing yielded a 0.585 Dice score, an improvement of +0.064 compared to the original nnUNet with 0.521 Dice score. While nnUNet is a state-of-the-art segmentation method, our dedicated postprocessing method contributed to further improvement in discarding false positive regions.

We note that the performance of the second observer ■ is substantially better than any of our experiments when segmenting only the kidney abnormalities, with an average 0.664 Dice score. Figure 4.5d shows four outliers for the second observer ■ , three of these cases obtained a Dice score of zero and one case 0.207. The volume of these four outliers is 29, 197, 282, and 5769 mm^3, three of them are below the median kidney abnormality volume in dataset B_{30} (1421 mm^3). This demonstrates the difficulty of kidney abnormality segmentation, even for experienced radiologists. The fact that we annotated multiple classes of kidney abnormalities (e.g. tumors, cysts, lesions, and masses) as a single class and the diverse patient anatomy in patients with kidney abnormalities may have contributed to the gap in performance.

Table 4.2 compares the Dice score obtained by previous work and our methods; the middle line separates methods that segmented kidneys without abnormalities and kidneys with abnormalities. While some methods reported Dice score for both kidneys as a single score as reported in this paper, others reported Dice scores for the left and right kidneys separately; then, we postprocessed our predictions to the same format and have a better comparison. Most of the methods trained without kidney abnormalities achieved higher Dice scores in the kidney parenchyma than those trained with kidney abnormalities (below the middle line) due to the more complex task. Although the performance of experiment 1 ■ for kidney abnormality segmentation was the lowest (0.487) among the previous work, the performance of the second observer ■ (0.664) was also below the previous work where Yu et al.[99] obtained 0.913 and Yang et al.[100] 0.802 Dice score. This disparity could be due to the fact that we grouped different types of kidney abnormalities including cysts, lesions, masses, metastases, and tumors into a single class while Yu et al.[99] and Yang et al.[100] discarded other abnormalities different than kidney tumors. Our set of kidney ab-

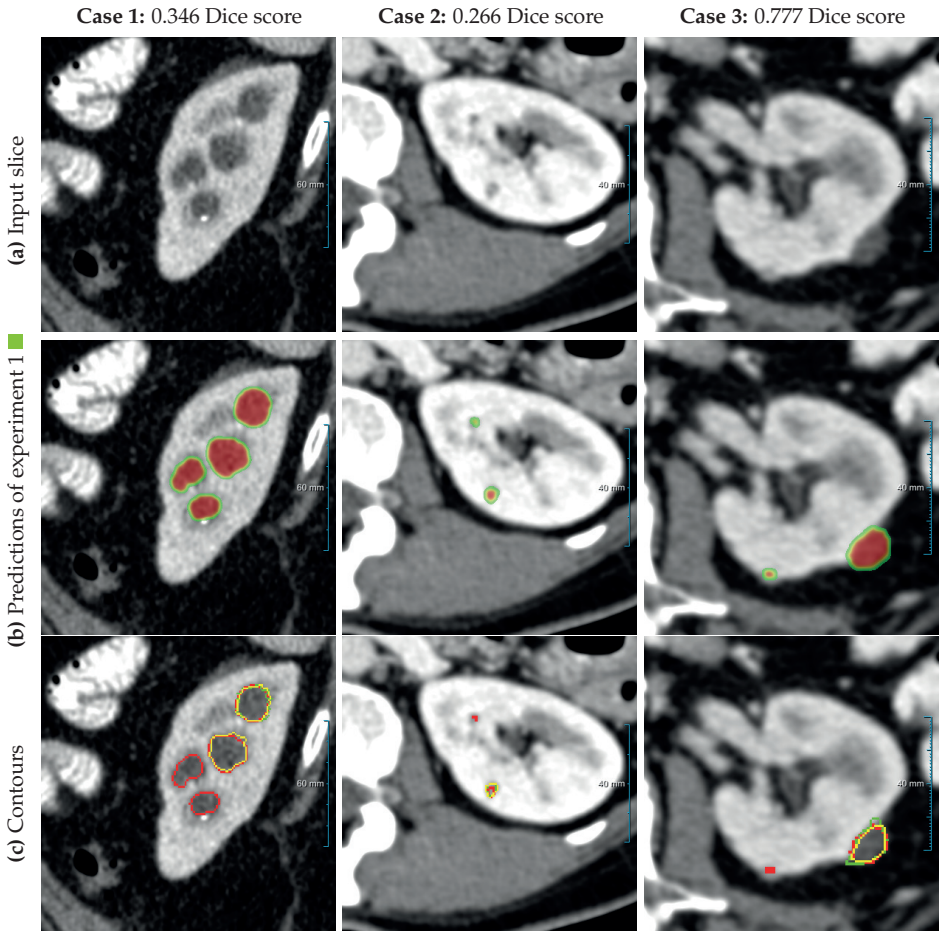


Figure 4.6: Comparison of three cases on the test set B_{30} between experiment 1 ■, the reference standard, and the second observer. (a) shows the original slice. (b) shows the heatmaps (predictions prior to post-processing, using a color table mapping [0,1] from transparent to green to red) of experiment 1 ■. (c) shows the final predictions (red contours) of experiment 1 ■, the reference standard (green contours), and the second human observer (yellow contours). The window center and window width used for all slices were 60 HU and 360 HU.

normalities is diverse in terms of volume, texture, image intensity, and location in the kidney, which makes network learning difficult.

Segmenting kidney abnormalities is challenging due to the similarity between tumors in the collecting system and kidney cysts. For instance, Figure 4.6 shows three cases from dataset B_{30} where our method returned some false positives due to the similarity with tumors in the collecting system. Each case shows the kidney abnor-

mality predictions of experiment 1 ■ prior to post-processing in the second row as heatmaps. While the third row shows the post-processed segmentation, reference standard, and second observer as red, green, and yellow contours, respectively. In all three cases, a false positive by our method is present, indicated by an isolated red contour. In case 1, the false positives are abnormalities in the collecting system, which have a similar image intensity as the cysts, similarly, the second observer also segmented one of these abnormalities in the middle region. In case 2, the false positive appears as a small cyst-like region, while in case 3, it resembles an irregular region in the kidney. Figure 4.7 shows a comparison of the final prediction in annotation format 1 of experiment 1 ■, the reference standard, and the second observer represented as red, green, and yellow contours, respectively. This figure shows the best and median cases of datasets B_{20} and B_{30} and the Dice score of each case computed between experiment 1 ■ and the reference standard.

A limitation of our study is that we excluded patients with unusual anatomy and with abnormalities in the collecting system.

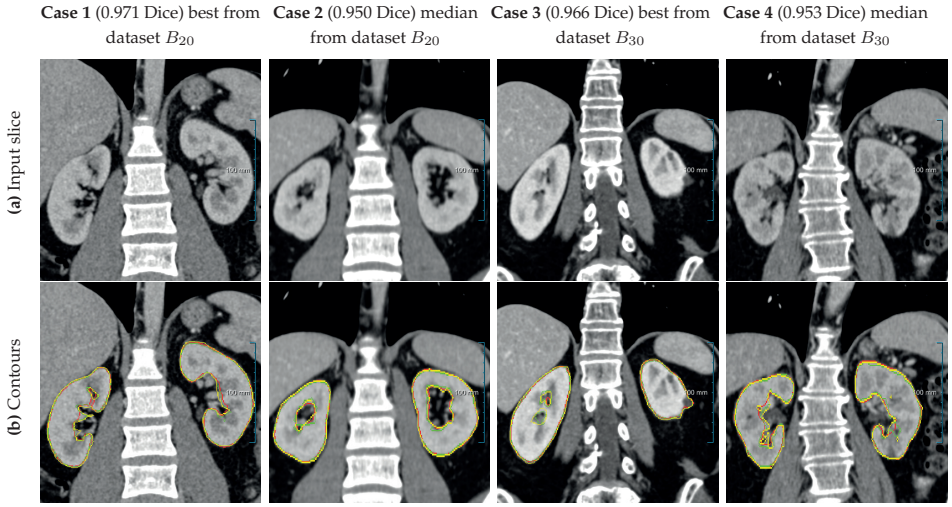


Figure 4.7: Comparison of four cases between experiment 1 ■, the reference standard, and the second observer on the test set B_{30} in annotation format 1. (a) shows the original slice and (b) shows the final predictions (red contours) of experiment 1 ■, the reference standard (green contours), and the second human observer (yellow contours). All the slices have a window center of 60 HU and a window width of 360 HU.

4.5 Conclusions

In conclusion, our ablation study and nnUNet showed that segmenting kidney abnormalities in challenging scenarios is possible, and improved performance can be achieved by an ensemble of different methods and dedicated postprocessing. The results show that our method has the potential to be a valuable tool for clinicians in detecting and monitoring kidney abnormalities. An ablation study was conducted to better understand the impact of the different modules of our method on its performance. Further research is needed to optimize the performance of experiment 1 and nnUNet to test their ability to generalize to other datasets. Overall, our work contributes to the ongoing efforts to develop accurate and reliable computer-aided diagnosis systems for detecting and quantifying renal abnormalities.

Transfer learning from a sparsely annotated dataset of 3D medical images

5

G.E. Humpire Mamani, C. Jacobs, M. Prokop, B. van Ginneken, N. Lessmann

Original title: Transfer learning from a sparsely annotated dataset of 3D medical images

Published in: arXiv:2311.05032, 2023

Abstract

Transfer learning is a technique used in machine learning where features learned by a model on a large annotated dataset are transferred and leveraged when training a new model for other tasks. This technique can save substantial time and computational resources compared to training models from scratch. In addition, performance may also improve when leveraging pre-trained features. Due to the lack of large datasets in the medical imaging domain, transfer learning from one medical imaging model to other medical imaging models has not been widely explored. This study explores the use of transfer learning to improve the performance of deep convolutional neural networks for organ segmentation in medical imaging. A base segmentation model (3D U-Net) was trained on a large and sparsely annotated dataset; its weights were used for transfer learning on four new down-stream segmentation tasks for which a fully annotated dataset was available. We analyzed the training set size's influence to simulate scarce data. The results showed that transfer learning from the base model was beneficial when small datasets were available, providing significant performance improvements; where fine-tuning the base model is more beneficial than updating all the network weights with vanilla transfer learning. Transfer learning with fine-tuning increased the performance by up to 0.129 (+28%) Dice score than experiments trained from scratch, and on average 23 experiments increased the performance by 0.029 Dice score in the new segmentation tasks. The study also showed that cross-modality transfer learning using CT scans was beneficial. The findings of this study demonstrate the potential of transfer learning to improve the efficiency of annotation and increase the accessibility of accurate organ segmentation in medical imaging, ultimately leading to improved patient care. We made the network definition and weights publicly available to benefit other users and researchers.

5.1 Introduction

Transfer learning is a widely used strategy when developing image analysis models based on deep learning. The weights of a model trained on a large dataset are transferred to a new model that learns a different task. This transfer initializes the weights of the new model with well-converged and meaningful filters, which gives the new model a head start compared with random initialization and especially improves the performance of models trained with small datasets[113]. While transfer learning is therefore useful in many computer vision and medical image analysis applications, the vast majority of pretrained models available for transfer learning are models trained with two-dimensional non-medical images, such as the ImageNet dataset[114]. Transfer learning from these models can lead to adverse results when training medical image analysis models[115]. However, models pretrained on a medical dataset with a large number of three-dimensional medical images and a diverse set of labeled objects are currently not readily available.

Although various annotated datasets with three-dimensional medical images are available under licenses that permit their use for pretraining, most of these datasets are small, and annotations are usually sparse. In virtually all segmentation datasets, only a single or a few anatomical structures are delineated because manual segmentation in three-dimensional datasets is time-consuming and expensive. Most projects are also focused on a specific task that does not require exhaustive annotation of all structures. However, datasets in which many visible anatomical structures are delineated would be best suited for training generic models for transfer learning.

There have been a few efforts to assemble datasets with a larger number of structures annotated, such as the VISCERAL dataset[116]. Other efforts focused on combining datasets and expanding annotations to additional structures, such as Gibson et al.[117], who combined two publicly available datasets[118,119] and expanded the annotations to 14 structures. However, increasing the number of annotated structures in a set of scans usually comes at the expense of the number of scans in the dataset.

This paper explores an alternative strategy for training a generic base model for segmentation tasks in medical images that does not require a fully annotated dataset. Instead, we propose to train the base model with a large but sparsely annotated dataset. This dataset is assembled from multiple publicly available datasets with CT images and reference segmentations of various anatomical structures. While each image has at least one delineated structure, we relax the requirement that all structures be delineated in all images and propose a method for training a deep neural network with this kind of sparse annotation. We investigate whether using this base model to initialize the network for a new segmentation task improves the perfor-

mance. We evaluate whether the size of the training dataset for the new tasks is related to the efficacy of transfer learning. To enable others to use this base model for transfer learning, the network is based on the commonly used 3D U-Net architecture[9], and the code and weights are made available online.[1]

5.2 Related work

5.2.1 Transfer learning

While transfer learning is widely applied to 2D data, it is not commonly applied in 3D medical imaging due to the lack of large 3D datasets. Multiple large 2D fully-annotated datasets (ImageNet, MS-COCO, and CIFAR) and spatiotemporal 2D datasets, such as Kinetics, are available in comparison to the small 3D medical imaging datasets[120].

Regardless of the image domain, methods trained on 2D images used transfer learning to 3D images [121–126] by decomposing the 3D image into a sequence of 2D images. In Conze et al.[122], the features of a pretrained network on ImageNet segmented healthy and unhealthy shoulder muscles in MRI using transfer learning. Similar to action recognition in videos, a pretrained network on Kinetics (large dataset for action recognition in videos) initialized a 3D network to diagnose appendicitis in CT scans[121]. The network expects a sequence of frames to recognize an action in a video; the sequence of frames was replaced by a sequence of 2D slices to recognize abnormal regions in a CT scan. In Yang et al.[123], 2D pretrained networks are converted to 3D networks; this approach benefits from the large-scale 2D datasets and the 3D context that 3D networks offer.

Although different-domain transfer learning methods showed higher results than methods trained from scratch, same-domain transfer learning showed more reliable results for medical imaging[65,127–130]. Roth et al.[65] trained a cascade of 3D U-Nets using a fully annotated dataset to make 3D medical imaging more accessible for other researchers. In Zhou et al.[128], a self-supervised learning method learns from unannotated medical data (LUNA16 dataset without annotations) to obtain a pretrained model, which can be fine-tuned for classification and segmentation. The method uses an encoder-decoder architecture to perform medical image restoration, which learns the texture and features of organs. Similar to our approach, Chen et al.[129] joined data from three medical challenges (MRI and CT) to compose a partially annotated dataset and trained 3D convolutional networks for segmentation and classification. In a recent study, Ji et al.[130] trained a nnU-Net on a large-scale medical

[1] https://github.com/DIAGNijmegen/MedicalTransferLearning3D-UNet

dataset for organ segmentation; transfer learning from that model increased the performance in unseen segmentation tasks from the Medical Segmentation Decathlon challenge[131]. The network weights obtained by Roth et al.[65], Chen et al.[129], and Zhou et al.[128] were released and are publicly available. Additionally, our method uses a partially annotated dataset composed of 6 publicly available datasets.

5.2.2 Learning from sparsely annotated data

Combining multiple medical annotated datasets could create a large but partially annotated dataset; this data cannot be directly used by methods that depend on fully annotated datasets. Only few researchers focused on medical image segmentation with partially annotated datasets where methods obtain pseudo-labels from unlabeled images to train networks. Pseudo-labels can be obtained by approximating the shape and position of a missing label[132–134], relabeling[135], weak annotations[136,137], or by adding constraints such us anatomical prior organ size[138]. By adapting the cross-entropy to learn more from the foreground than the background, Jin et al.[139] trained a 3D network to add more airway branches to the airway segmentation obtained by a previous method. Multi-stage approaches (i.e., per groups of fully annotated data) served as a multi-organ segmentation network[129,135,140]. In an end-to-end solution, Shi et al.[141] proposes two losses to train a network using a partially annotated dataset (data from multiple datasets). The first loss merges all unlabeled as a single label and the second loss assumes organs are non-overlapped to differentiate between labeled organs and estimated predictions. Liu et al.[142] used incremental learning to train the network on a different organ in each stage. After four stages, the network segments the organs from four datasets. The method uses a corrective loss to remove low-confidence output. Zhang et al.[143] proposed DoDNet, a network that emulates a multi-head network (each head for a different task) by proposing a dynamic single-head network.

5.3 Dataset

We combined a number of datasets to assemble a large but sparsely annotated training set for the generic base model, and multiple additional datasets for transfer learning experiments. See Figure 5.1 for examples of some of the datasets. Many of these datasets have previously been made publicly available as part of segmentation challenges. Because the field of medical imaging is so broad with various modalities such as CT and MRI and various imaging modes per modality such as contrast-enhanced or ECG-gated CT imaging, we limited our experiments to the datasets consisting

of thoracic and abdominal CT scans. The evaluation also includes a transfer learning experiment with an MR dataset to investigate whether cross-modality transfer learning is effective. We used the following datasets in our study:

- The 2019 Kidney Tumor Segmentation challenge (KiTS19) dataset[104] comprises 300 abdominal CT scans from a single medical center. In all scans, the kidneys and kidney tumors were manually delineated and post-processed to remove fat tissue. We included only the 210 scans that were originally made available for model training and validation.

- The Liver and Liver Tumor Segmentation challenge (LiTS) dataset[96] comprises 200 thoracic-abdominal CT scans from several medical centers. In the scans, the liver and liver tumors were manually delineated. We included only the 131 scans that were originally made available for model training and validation.

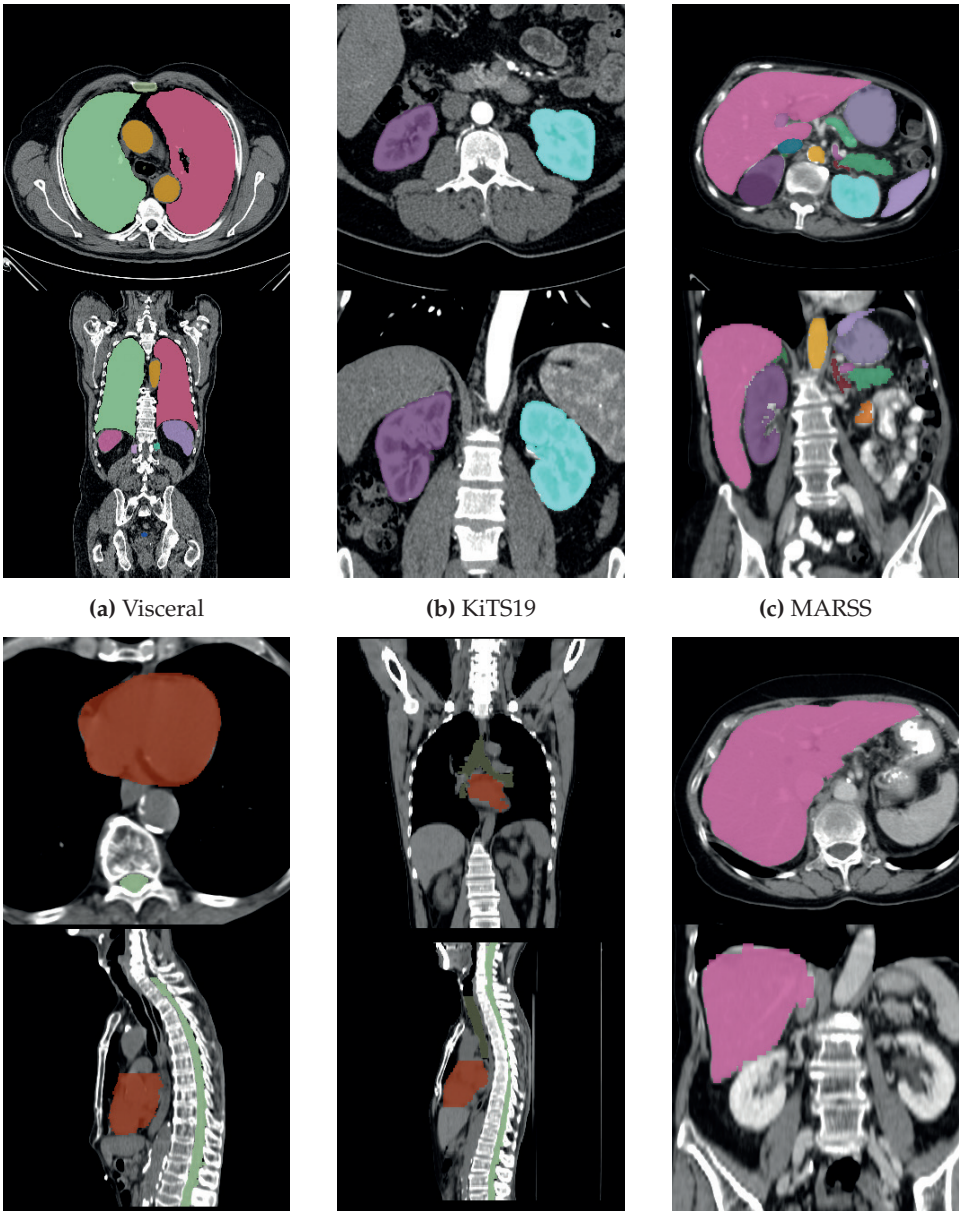
- The Multi-organ Abdominal CT Reference Standard Segmentations (MARSS) dataset[117] is itself based on two other datasets, namely 47 images from the Multi-atlas Labeling Beyond the Cranial Vault challenge[119] and 43 images from The Cancer Image Archive Pancreas-CT dataset[118]. In these 90 abdominal CT scans, a total of 14 structures were manually delineated, but 6 of them were only in the Cranial Vault dataset (see Table 5.1).

- The dataset from the Automatic Structure Segmentation for Radiotherapy Planning challenge (StructSeg2019)[2] contains CT scans of the head and neck and the thorax. We used only the 50 thoracic CT scans from the "organ-at-risk" segmentation subtask, for which delineations of the lungs, heart, esophagus, and spinal cord are available.

- The AAPM Thoracic Auto-Segmentation Challenge (TASC) dataset[144] comprises 36 thoracic CT scans with delineations of the esophagus, heart, lungs, and spinal cord.

- The Visceral dataset[116] comprises 40 thorax-abdomen CT scans with delineations of 20 structures (see Table 5.1). We disregarded four of these structures: the left and right rectus abdominis muscles because they are thin and often difficult to segment structures, the thyroid because we found that the segmentations were of lower quality compared with other structures, and the L1 vertebrae since only a single vertebra was delineated while we used other datasets (see below) with segmentations of all visible vertebrae.

[2] https://structseg2019.grand-challenge.org/



(a) Visceral **(b)** KiTS19 **(c)** MARSS

(d) TASC **(e)** StructSeg2019 **(f)** LiTS

Figure 5.1: Examples CT scans of publicly available datasets. Colored regions represent annotated (reference masks) regions per dataset. The joined annotations of all those datasets compose the partially annotated dataset used for the training of the base model (Exp0). This figure uses the lung window level.

- The dataset from the 2014 vertebra segmentation challenge at the Computational Methods and Clinical Applications for Spine Imaging (CSI) workshop comprises 20 spine-focused thorax-abdomen CT scans with delineations of the thoracic and lumbar vertebrae[145].

- The Large Scale Vertebrae Segmentation challenge (VerSe19) dataset[146,147] comprises 180 CT scans of the spine, including thoracic and abdominal scans but also cervical spine scans. We used the 80 scans from the first two training batches and the corresponding delineations of all visible vertebrae.

- The COPDGene study[148] is a clinical trial that enrolled 10,000 patients with mild to severe COPD who received a thoracic CT scan in one of 21 medical centers in the United States. We used a randomly selected subset of 100 CT scans for which we had access to delineations of the pulmonary lobes.

- The PROMISE12 challenge dataset[149] comprises 50 T2-weighted MR scans with delineations of the prostate.

Table 5.1 provides an overview of the number of scans and the annotated anatomical structures in the individual datasets.

5.3.1 Harmonization of the reference delineations

In the large dataset that we created by combining multiple datasets, each dataset contains several structures annotated in more than one of the source datasets. The annotation protocols sometimes differed slightly, and we, therefore, post-processed some of the annotations to harmonize the reference data.

Three datasets delineated the kidneys: MARSS, visceral, and KiTS19. In the KiTS19 dataset, kidney tumors were annotated with a separate label, while this was not the case in the other two datasets. We, therefore, decided not to consider kidney tumors a separate class and added them to the kidney class in the KiTS19 dataset. Additionally, the KiTS19 dataset does not distinguish between left and right kidneys, while the other two kidney datasets use two different labels. Therefore, we identified the left and right kidneys in the KiTS19 dataset and labeled them the same way as in the other two datasets.

Similarly, the LiTS dataset contains delineations of the liver as well as liver tumors, while other datasets did not separately label liver tumors (MARSS and visceral). Therefore, we combined the liver and liver tumor labels in the LiTS dataset.

In the CSI and VerSe19 datasets, we simplified the vertebra segmentation task by removing the anatomical identification task. Instead, we assigned all vertebrae to

Table 5.1: Summary of publicly available datasets considered in this paper and the number of CT scans annotated per dataset and organ. The datasets above the middle line (except annotations of the esophagus) served to compose the partially annotated dataset to train Exp0. The datasets below the middle line (+ subset of masks of the esophagus) are fully annotated datasets and were used to analyze the influence of transfer learning on new segmentation tasks.

Dataset	Organ/structure name																														
	Left adrenal gland	Right adrenal gland	Aorta	Bladder	Duodenum	**Esophagus**	Gallbladder	Heart	Left kidney	Right kidney	Liver	**Lung lobes**	Left lung	Right lung	Pancreas	Portal & splenic vein	**Prostate**	Left psoas major	Right psoas major	Left rectus abdominis	Right rectus abdominis	Spinal cord	Spleen	Sternum	Stomach	Thyroid	Trachea	**Vertebrae**	Vertebra L1	Vena cava
KiTS19									210	210																					
LiTS											131																				
MARSS	47	47	47		90	**90**	90		90	47	90				90	47							90		90						47
StructSeg2019						**50**		50					50	50								50					50				
TASC						**36**		36					36	36								36									
Visceral	30	27	40	39			38		40	40	40		40	40	38			40	40	40	40		40	40		32	40		40		
CSI																												**20**			
VerSe19																												**80**			
COPDgene												**100**																			
PROMISE12 (MR)																	**50**														
Base model	77	74	87	39	90	–	128	86	340	297	261	–	126	126	128	47	–	40	40	–	–	86	130	40	90	–	90	–	–	–	47

the same class. This also helped to eliminate any labeling discrepancies between the datasets.

5.3.2 Base model and transfer learning tasks

The combined dataset contains a total of 736 CT scans and 50 MR scans with delineations of 26 anatomical structures, where each structure is annotated in at least 39 and in up to 340 scans (see Table 5.1). For the generic base model training set, we selected 22 of these structures; this corresponds to a set of 556 CT scans of which 90% were used for training and 10% for validation.

We used the remaining four structures (esophagus, vertebrae, lung lobes, and prostate) for transfer learning experiments with the generic base model (Figure 5.2). The esophagus was annotated in three datasets already included in the training set, i.e., this new target structure was not part of the annotations in the training set, but the same CT scans were used to train the generic base model. This was not the case for the vertebrae and the lung lobes, which were annotated in scans that were not

part of the training set. Finally, the prostate was annotated in MR scans rather than CT scans. For all four tasks, 10% of the available scans were set aside for evaluation of the segmentation performance.

5.4 Method

The proposed strategy for training a generic base model using a sparsely annotated dataset does not require a specific network architecture. However, networks initialized with the weights of this base model will typically use the same or a very similar architecture since the learned weights are coupled to the size and order of the individual layers in the original architecture. We used the 3D U-Net[9] architecture both for the generic base model and for all transfer learning experiments because this architecture is particularly popular for medical image segmentation tasks. However, to avoid potential issues with normalization layers that become too data-specific and thus hinder transfer learning, we opted to remove batch normalization from this architecture. The 3D U-Net architecture used in this paper uses four resolutions, i.e., contains three pooling layers in the compression path. The number of filters in the convolutional layers starts at 32 filters and doubles after each pooling layer.

The 3D U-Net is a patch-based segmentation network. The models trained with CT images used an input patch size of 132×132×132 voxels. Since the network does not make use of padding in the convolutional layers, the size of the output patches is smaller, namely 44×44×44 voxels. The models trained with MR images from the PROMISE12 dataset used a different patch size of 108×108×108 voxels because the images had a substantially smaller field of view compared with the CT scans so that 132×132×132 voxel patches would have been often larger than the entire image. The corresponding output patch size was 20×20×20 voxels. Note that the network does not contain any fully-connected layers but only convolutions and pooling layers, which makes it possible to change the input patch size without affecting the model.

5.4.1 Pre-processing

All networks were trained with isotropically resampled images and reference segmentation masks. Images were resampled using cubic interpolation and reference segmentation masks using nearest neighbor interpolation. The combination of various datasets with images acquired in different institutions and for different purposes resulted in a dataset with a wide range of different image resolutions. For instance, the spacing between slices ranged from 0.5 mm (PancreasCT) to 5 mm (KiTS19, LiTS and Multi-atlas Labeling Beyond the Cranial Vault challenge (Cranial Vault)). We

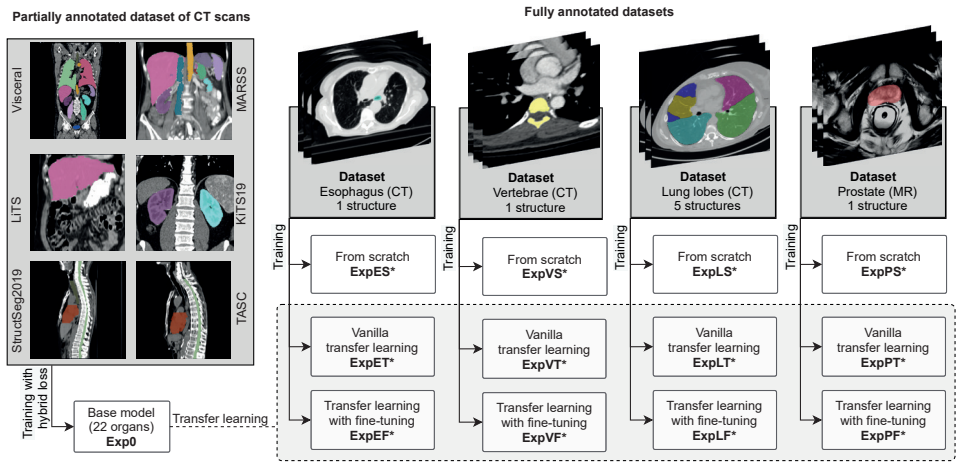


Figure 5.2: This paper has two types of datasets, the large sparsely annotated dataset–obtained by joining multiple publicly available datasets–and the fully annotated datasets (vertebrae/CT, esophagus/CT, lung lobes/CT, and prostate/MR). The large sparsely annotated dataset served to train Exp0, which learned from annotated regions only. The fully annotated datasets were used for the evaluation of three training strategies (scratch, vanilla transfer learning, and transfer learning with fine-tuning); each of these experiments was evaluated on different training set sizes. The weights of Exp0 initialized the weights of the networks that use transfer learning and continue training on the new fully annotated datasets.

resampled all images using cubic interpolation and all segmentation masks using nearest-neighbor interpolation to 1mm $\times$ 1mm $\times$ 1mm isotropic resolution.

Image intensities were normalized differently in CT and MR images. In CT images, the image values in Hounsfield Units (HU) were clipped to the range [-500,400], which roughly corresponds to the abdominal window and level settings used to view abdominal CT images. In MR images, we clipped the intensity values to the 5% and 95% percentiles of the image values (per image) and scaled this interval to the range [-500, 400] to match the range used for the CT data.

5.4.2 Base model (Exp0)

We refer to the generic base model that was trained with a dataset with sparse annotations of 22 structures as Experiment Zero (Exp0). This model was trained using patches evenly sampled from the training dataset with a stride of 30mm, a parameter determined through empirical experimentation. The final layer of the network is a softmax layer with 23 classes corresponding to 22 foreground structures and the

Table 5.2: Data distribution per dataset for the four new tasks for the transfer learning experiments.

Segmentation task	Data source	Modality	Number of scans		
			Training	Validation	Total
Esophagus	MARSS, StructSeg2019, and TASC	CT	158	18	176
Vertebrae	CSI and VerSe19	CT	90	10	100
Lung lobes	COPDgene	CT	90	10	100
Prostate	PROMISE12	MR	45	5	50

background. The background contains all the structures that were not annotated in any of the images in the training set. Due to the sparse nature of the annotations, there are no images with annotations of all foreground structures, and hence no images in which the background voxels are known. For example, voxels that were not annotated as kidneys in the KiTS19 images might be background voxels but might just as well belong to one of the other foreground classes that were not annotated in the KiTS19 subset, e.g., the liver. To mitigate this problem, we trained the base model using a hybrid masked loss function composed of two terms, an average Dice score term and a cross-entropy term. These were computed only for the present foreground structures, ignoring classes that were not annotated in an image.

Given a binary flag $\delta_c \in \{0, 1\}$ that indicates whether structure c is one of the $N \in \{1, \ldots, 22\}$ annotated structures in the present image and a weight map ω_c (defined below), the hybrid masked loss function is defined as:

$$\mathcal{L} = \left(1 - \frac{1}{N}\sum_{c=1}^{22} \delta_c \cdot \mathrm{Dice}_c\right) + \left(-\sum_{c=1}^{22} \delta_c \cdot \omega_c \cdot \mathrm{CCE}_c\right),$$

where Dice_c and CCE_c correspond to the soft Dice score and the categorical cross-entropy for class c, respectively. Both loss components are never computed for the 23rd class, the background since background labels are unknown in any images. The factor δ_c ensures that the loss components are zero for structures that were not annotated in the image. Note that this does, in principle, not have any effect on the cross-entropy term since the cross-entropy for structure c is already zero for voxels with another label.

The Dice score component is crucial for training with sparsely annotated data without explicit background labels. In each image, N of the 22 structures are annotated, and the corresponding voxels' labels are thus known. For the remaining voxels, we only know that their labels are not among the N annotated classes. The Dice score as a volume overlap measure penalizes false positive predictions and thus

penalizes the classification of unlabeled voxels as one of the N known structures. By presenting in consecutive training steps images from different subsets, i.e., with different structures annotated, the network can learn to recognize all 22 structures. The use of a softmax layer forces the network to resort to the background class for voxels that it does not recognize as any of the 22 structures; the probability of the background class will automatically increase when the network assigns low probabilities to all other classes.

The cross-entropy component is computed based on only the labeled voxels and thus only penalizes false-negative predictions. In combination with the Dice score component, it provides an additional penalty for incorrect classifications of the labeled voxels in a training image, which boosts the importance of the strong labels in the training data, i.e., the annotated foreground objects in each image. Because cross-entropy terms are sensitive to class imbalances, we also introduce a weight map ω_c for each class. The weight of each class corresponds to the inverse sampling probability across the entire training set[150].

The weights of the model were initialized using the Glorot uniform initializer[108]. To optimize the weights, we used the Adam optimizer[109] and trained three networks per experiment using different learning rates: 1×10^{-4}, 1×10^{-5}, and 5×10^{-5}. The training was stopped once the mean dice score in the validation set did not improve for ten epochs. An epoch was defined as the full iteration of positive patches (containing annotations) from all CT images in the training set.

5.4.3 Data augmentation

During network training, 70% of the samples in each epoch were subject to slight random transformations to augment the training data. We applied a combination of up to three of the following data augmentation techniques to the samples: 3D scaling, 3D rotation, Gaussian blurring, image intensity variation, and elastic deformations. Scaling was between -5% and 5%, rotation in up to two planes between -5° and 5° degrees, Gaussian blurring with sigma between 0.2 and 1.0, and image intensity variation between -20 and 20 HU, which was applied to the entire image. The elastic deformation method used a ten-control point grid on the sample where every control point was randomly shifted up to 5 voxels. Because the combination of elastic deformations and scaling or rotation frequently resulted in unrealistic images, we allowed only the combination with Gaussian blurring or image intensity variation.

Additionally, to combat the difference in image quality across scans from the various datasets, we randomly applied salt and pepper noise to 20% of the voxels followed by Gaussian smoothing ($\sigma = 0.9$mm), which results in CT scans that look

similar to scans acquired with lower radiation, i.e., lower mAs and kVp.

5.4.4 Post-processing

The image was divided into non-overlapping patches corresponding to the network output to obtain segmentations of an entire image. The predictions for these patches were stitched together to form a complete segmentation mask and were thresholded at 0.5, assigning the background label if none of the foreground classes reached a probability above 0.5. Connected component analysis was used to remove all but the largest structure for each foreground class. Finally, the predicted segmentation masks were resampled to the image's original resolution.

5.4.5 Transfer learning (ExpXYZ)

The effect of transfer learning when training a 3D U-Net for segmentation of a new structure was evaluated for four new segmentation tasks (Table 5.2). For each task, we compared three training strategies: (1) training from scratch, where the weights of the network were randomly initialized, and all network weights were updated during training; (2) vanilla transfer learning, where the weights of the network were initialized with the weights of the trained base model (Exp0), and all networks weights are updated during training; and (3) transfer learning in combination with fine-tuning, where the base model was used to initialize the network, followed by training only the last three layers for the new task while the other layers remained fixed. The intention was to prevent a form of catastrophic forgetting where useful low-level filters in the first layers that were inherited from the base model might be forgotten when switching abruptly to a new task. By first adjusting only the last few layers to the new task and then fine-tuning the entire network, the network might profit more from transfer learning in a second step. In each step, the network was trained until convergence.

The output layer of the base model has 23 channels, corresponding to a background class and 22 foreground classes. New tasks will usually have fewer classes, e.g., three of our example tasks are binary segmentation problems with only one foreground class, and the lung lobe segmentation task has five foreground classes. When initializing a new model with the weights of the base model, we retain the background class and reduce the number of foreground classes to the required number of foreground classes by dropping channels from the output layer.

The new models were trained in the same way as the base model with the exception of the loss function, which did not use the class presence factor δ_c and included the background class in the cross entropy computation (i.e., an unmodified

weighted cross-entropy term was used in combination with an unmodified soft Dice score term). In addition, we extracted patches with a stride of 10mm from the images because there were fewer images in the training sets than in the base model training set which used a stride of 30mm.

Table 5.3: List of experiment IDs for the four additional segmentation tasks (vertebrae, esophagus, lung lobes, and prostate), training strategy (scratch, vanilla transfer learning, and transfer learning with fine-tuning), and training set size ($Z = [10, 20, 30, 40, 50,$ and full] CT scans). For instance, the experiment ExpVT40 trained on 40 CT scans using vanilla transfer learning to segment the vertebrae. The results show the mean Dice score and standard deviation.

	Training from scratch		**Vanilla transfer learning**		**Transfer learning with fine-tuning step**	
	Experiment	Dice score	Experiment	Dice score	Experiment	Dice score
	ExpES10	0.459 ± 0.245	ExpET10	0.548 ± 0.214	ExpEF10	0.588 ± 0.193
Esophagus	ExpES20	0.579 ± 0.194	ExpET20	0.617 ± 0.201	ExpEF20	0.641 ± 0.197
	ExpES30	0.590 ± 0.204	ExpET30	0.632 ± 0.197	ExpEF30	0.652 ± 0.194
Modality: CT	ExpES40	0.624 ± 0.195	ExpET40	0.659 ± 0.195	ExpEF40	0.666 ± 0.197
Structures: 1	ExpES50	0.627 ± 0.213	ExpET50	0.653 ± 0.196	ExpEF50	0.668 ± 0.197
	ExpES158	0.694 ± 0.196	ExpET158	0.684 ± 0.206	ExpEF158	0.696 ± 0.206
	ExpVS10	0.920 ± 0.036	ExpVT10	0.926 ± 0.031	ExpVF10	0.929 ± 0.028
Vertebrae	ExpVS20	0.931 ± 0.030	ExpVT20	0.935 ± 0.029	ExpVF20	0.928 ± 0.029
	ExpVS30	0.939 ± 0.023	ExpVT30	0.942 ± 0.025	ExpVF30	0.939 ± 0.025
Modality: CT (unseen)	ExpVS40	0.943 ± 0.022	ExpVT40	0.944 ± 0.025	ExpVF40	0.943 ± 0.026
Structures: 1	ExpVS50	0.942 ± 0.024	ExpVT50	0.952 ± 0.020	ExpVF50	0.949 ± 0.023
	ExpVS90	0.956 ± 0.020	ExpVT90	0.955 ± 0.022	ExpVF90	0.955 ± 0.016
	ExpLS10	0.917 ± 0.027	ExpLT10	0.930 ± 0.028	ExpLF10	0.941 ± 0.030
Lung lobes	ExpLS20	0.939 ± 0.035	ExpLT20	0.948 ± 0.026	ExpLF20	0.950 ± 0.025
	ExpLS30	0.951 ± 0.024	ExpLT30	0.955 ± 0.022	ExpLF30	0.956 ± 0.023
Modality: CT (unseen)	ExpLS40	0.959 ± 0.016	ExpLT40	0.961 ± 0.016	ExpLF40	0.961 ± 0.018
Structures: 5	ExpLS50	0.961 ± 0.015	ExpLT50	0.965 ± 0.014	ExpLF50	0.964 ± 0.013
	ExpLS90	0.969 ± 0.010	ExpLT90	0.970 ± 0.011	ExpLT90	0.969 ± 0.013
	ExpPS10	0.816 ± 0.020	ExpPT10	0.813 ± 0.014	ExpPF10	0.818 ± 0.019
Prostate	ExpPS20	0.819 ± 0.062	ExpPT20	0.851 ± 0.032	ExpPF20	0.862 ± 0.018
	ExpPS30	0.844 ± 0.053	ExpPT30	0.854 ± 0.018	ExpPF30	0.870 ± 0.015
Modality: MR (unseen)	ExpPS40	0.852 ± 0.070	ExpPT40	0.861 ± 0.034	ExpPF40	0.869 ± 0.016
Structures: 1	ExpPS50	–	ExpPT50	–	ExpPF50	–
	ExpPS45	0.852 ± 0.031	ExpPT45	0.882 ± 0.010	ExpPF45	0.884 ± 0.012

5.5 Results

We performed experiments with the base model and with models trained for four new segmentation tasks, where these models were either trained from scratch or ini-

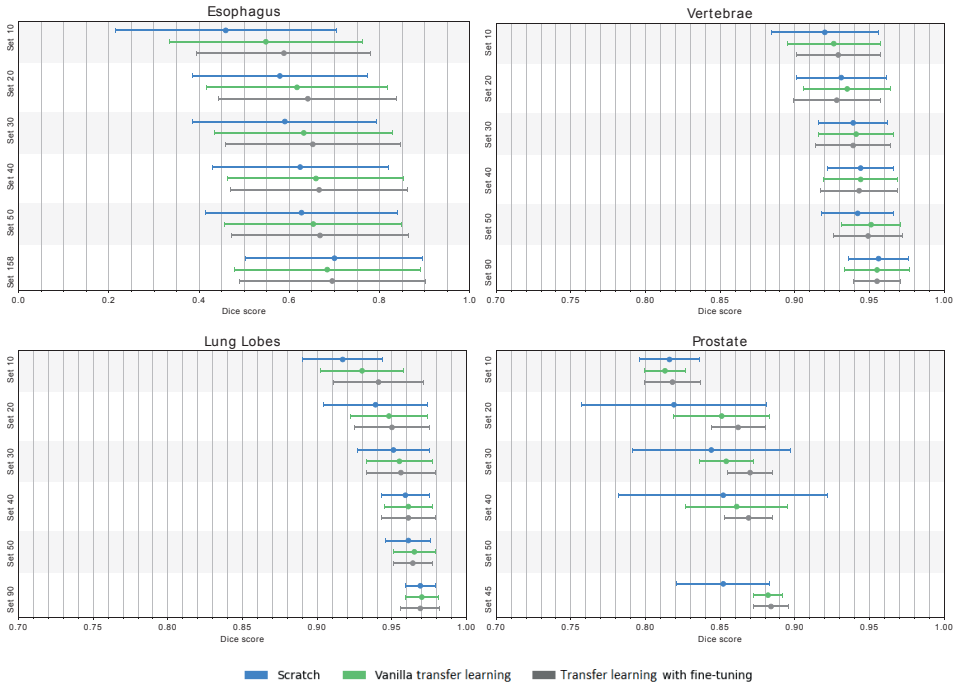


Figure 5.3: Bar plots of Dice score obtained by the experiments reported in Table 5.3 respectively. Note that the Dice score ranges from 0.7 to 1.0 in the subfigures for better visualization, except for the esophagus.

tialized with the base model. The new models were trained with different training set sizes by randomly selecting a subset of the available training data to simulate limited training data. In the following, we refer to the generic base model as Exp0 and use the following naming convention for the experiments of the four new tasks: ExpXYZ, where X represents the segmentation task (E=esophagus, V=vertebrae, L=lung lobes, and P=prostate), Y represents the training strategy (S=scratch, T=vanilla transfer learning, F=transfer learning with fine-tuning step), and Z represents the training set size (10, 20, 30, 40, 50, and all the images available in the dataset).

5.5.1 Base model (Exp0)

The base model was trained with a sparsely annotated dataset and evaluated on a randomly selected subset of this dataset, which was not used for training. Note that there was no separate test set but that the performance on the validation set is reported. We chose not to reserve a test set for evaluating the base model because the transfer learning experiments are the focus of this paper, and the base model's

performance is only of secondary interest. Since the base model was evaluated with sparsely annotated data, we ignored structures without annotations and calculated the Dice volume overlap score for each image between the available reference segmentations and the automatic segmentation results for these structures. The scores were then averaged across the dataset, resulting in an average Dice score of 0.725 $\pm$ 0.195 for segmenting 22 structures across 54 images. The highest scores were achieved for large structures like the lungs (0.967 and 0.966 for the right and left lungs, respectively), while small and irregular structures like the portal and splenic veins (0.349) achieved the lowest scores.

5.5.2 Transfer learning from Exp0

To evaluate whether transfer learning from the generic base model (Exp0) is beneficial, we trained segmentation networks for four additional tasks using fully annotated rather than sparsely annotated datasets (Table 5.2). The performance of the models when trained with the full datasets and smaller subsets of the data are listed in Table 5.3 and visualized in Figure 5.3. Figures 5.4 to 5.7 show examples of segmentations; the green and red regions represent the annotations and the predictions, respectively, both with transparency to visualize the overlap (dark green) between regions.

The segmentation models generally performed better when trained with larger training sets, regardless of the training strategy. When training from scratch, the increase in performance when training with more data was largest in the low data regime. For example, the Dice score for the challenging esophagus segmentation task increased from 0.459 to 0.579 (diff. +0.120) when training with 20 instead of 10 scans, but adding ten additional scans and training with 30 scans in total only resulted in a further increase of 0.11 in Dice score. For the pulmonary lobe segmentation task, where large parts of the object are well recognizable but where the boundaries are challenging to delineate precisely, ten scans were sufficient to reach a Dice score of 0.917. Adding more scans resulted in a steady increase in performance, reaching 0.969 when training with 90 scans.

Transfer learning by initializing the network weights with the weights of the base model resulted, in most cases, in a better segmentation performance, both vanilla and with the fine-tuning step. The impact differed per task and training set size. Models that were trained with a small dataset and reached low segmentation performance generally profited more from transfer learning, such as the esophagus segmentation task where the model trained with only ten scans improved from 0.459 to 0.548 (diff. +0.089) when using vanilla transfer learning and to 0.588 when using

transfer learning with fine-tuning step. On the other hand, lung lobes and vertebra segmentation models trained with ten scans already reached a Dice score of more than 0.900, and transfer learning had a limited impact on the performance of these models. Overall, we observed that transfer learning did usually not contribute anymore to the increase in performance when approximately 30 training scans were available. At the same time, transfer learning usually did not hurt the performance.

5.5.3 Cross-modality transfer learning

To evaluate whether transfer learning from the base model (Exp0, trained on CT scans) is beneficial for other modalities, we trained networks using MR scans (prostate segmentation task) as input instead of CT scans. Although all the training strategies trained on ten images obtained similar results, both transfer learning strategies obtained higher results than experiments trained from scratch when training on 20 (and more) images. When training with 20 images, the vanilla transfer learning experiment obtained 0.851 Dice, while the experiment trained from scratch obtained 0.816 (diff: +0.032); the fine-tuning experiment obtained 0.862 (diff with scratch +0.062). As with the experiments with CT scans (esophagus, lung lobes, and vertebrae segmentation tasks), the performance increases when adding more images to the training set regardless of the training strategy. We observed that the difference among training strategies gets smaller with more training data for CT data; this was different for the prostate segmentation task, where the transfer learning experiments kept a steady difference with the experiments trained from scratch while adding more training data. For instance, the difference in Dice score between vanilla transfer learning and scratch experiments ranged from 0.010 to 0.032 (average diff: 0.020) when the training set size increased from 10 to 45 MR scans. While the difference in Dice score between transfer learning with fine-tuning and scratch experiments ranged from 0.026 to 0.043 (average diff: 0.029) when increasing the training set size from 10 to 45 MR scans.

5.5.4 Transfer learning with fine-tuning step from Exp0

Large network weight changes may happen when a pretrained network is re-trained on a different task. We conducted experiments where the pretrained network gradually adapted to the new task by allowing weight changes to a certain number of layers. Subsequently, changes to all the network weights are allowed for further specialization; we refer to this procedure as transfer learning with fine-tuning step.

Overall, transfer learning with fine-tuning obtained higher results than the other two training strategies (vanilla transfer learning and scratch). For instance, the ex-

periment to segment the esophagus trained on 10 CT scans using transfer learning with fine-tuning (ExpEF10) obtained 0.588 Dice score, +0.129 than the experiments trained from scratch ExpES10. Adding more images to the training set gradually reduces the performance difference from 0.129 to 0.002 (average diff: 0.056). The vertebrae segmentation tasks slightly increased the performance of the experiments trained from scratch from 0.920 to 0.929 (diff: +0.009) when training on 10 CT scans; adding more CT scans made the experiments trained from scratch slightly better than transfer learning. Similarly, the lung lobe segmentation task showed minor improvements, where fine-tuning with 10 CT scans obtained 0.941 Dice, which is only +0.024 higher than the experiment trained from scratch.

When comparing the transfer learning strategies, we observe that transfer learning with fine-tuning gets higher results than vanilla transfer learning. The esophagus was the most benefited segmentation task with the transfer learning with fine-tuning strategy. For instance, for a training set of 10 CT scans, the performance increased in +0.040 Dice score when using fine-tuning compared to vanilla transfer learning. Fine-tuning got slightly higher results for the lung lobes than vanilla transfer learning when training with up to 30 images, while more images made vanilla transfer learning slightly higher. For the vertebrae experiments, the vanilla transfer learning was slightly better than fine-tuning; the difference in performance was between -0.0066 to +0.0026 (average diff: -0.0017) for the experiments with different training set sizes.

5.5.5 Vanilla transfer learning from a simpler model

To evaluate whether transfer learning from Exp0 (trained on a large sparsely annotated dataset) is more beneficial than a simpler model, we compared the results of transfer learning from a simpler model to transfer learning from Exp0. We conducted experiments using vanilla transfer learning from:

Vertebrae to other segmentation tasks

We picked one of the models of the new segmentation tasks trained from scratch with all the images available for further analysis. The weights of the model ExpVS90 (vertebrae trained from scratch using 90 CT scans) initialized networks to perform vanilla transfer learning on the esophagus, lung lobes, and prostate datasets (see Table 5.4). The esophagus experiment (initialized with the simpler model ExpVS90) trained on 10 CT scans obtained 0.475 Dice, while ExpET10 (initialized with Exp0) obtained 0.548 Dice (diff: +0.073). The difference in all the training set sizes for the esophagus ranged from -0.004 to 0.073 (average diff among five experiments: 0.036).

Table 5.4: Results of vanilla transfer learning from a simpler model (ExpVS90) to the esophagus, lung lobes, and prostate segmentation tasks.

Dataset	Set: 10 images	Set: 20 images	Set: 30 images	Set: 40 images	Set: 50 images	Set: all images available
Esophagus	0.475 ± 0.243	0.550 ± 0.254	0.617 ± 0.199	0.616 ± 0.218	0.633 ± 0.199	0.688 ± 0.209
Lung lobes	0.908 ± 0.028	0.934 ± 0.029	0.946 ± 0.029	0.956 ± 0.017	0.964 ± 0.014	0.968 ± 0.016
Prostate	0.794 ± 0.043	0.844 ± 0.035	0.870 ± 0.017	0.881 ± 0.026	–	0.882 ± 0.017

Similarly, for the lung lobe segmentation task, the difference in performance among training set sizes ranges from 0.001 to 0.022 (average diff among five experiments: 0.009). In contrast, the prostate segmentation task obtained differences in performance ranging from -0.02 to 0.019 (average diff among four experiments: -0.002).

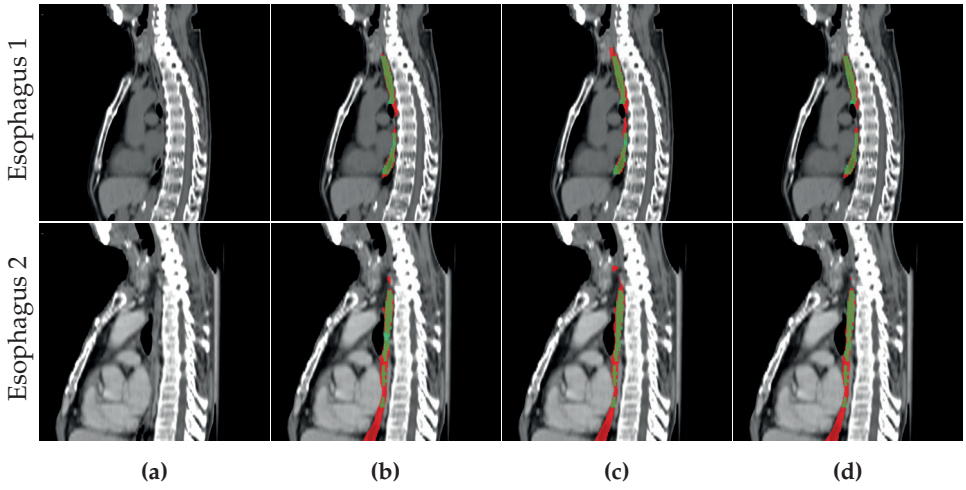


Figure 5.4: Predictions of the experiments of the esophagus segmentation task trained on 30 CT scans. (a) Shows the original slice, and the training strategies (b) scratch ExpES30, (c) vanilla transfer learning ExpET30, and (d) transfer learning with fine-tuning ExpEF30.

5.6 Discussion

This study evaluated transfer learning for segmentation tasks in medical imaging. For this study, we focused on CT imaging and collected a large dataset consisting of 556 CT scans from 6 publicly available datasets. We trained a base segmentation model using this large but sparsely annotated dataset and evaluated whether

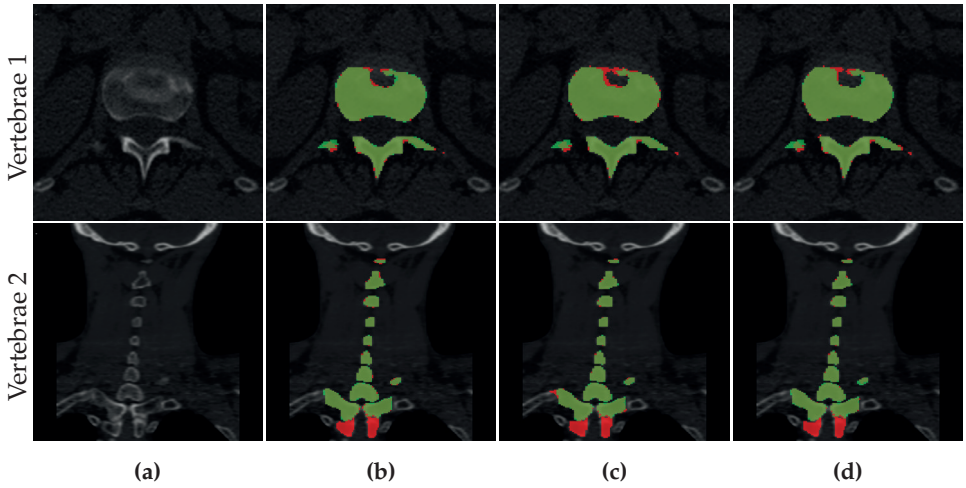


Figure 5.5: Predictions of the experiments of the vertebrae segmentation task trained on 30 CT scans. (a) Shows the original slice, and the training strategies (b) scratch ExpVS30, (c) vanilla transfer learning ExpVT30, and (d) transfer learning with fine-tuning ExpVF30.

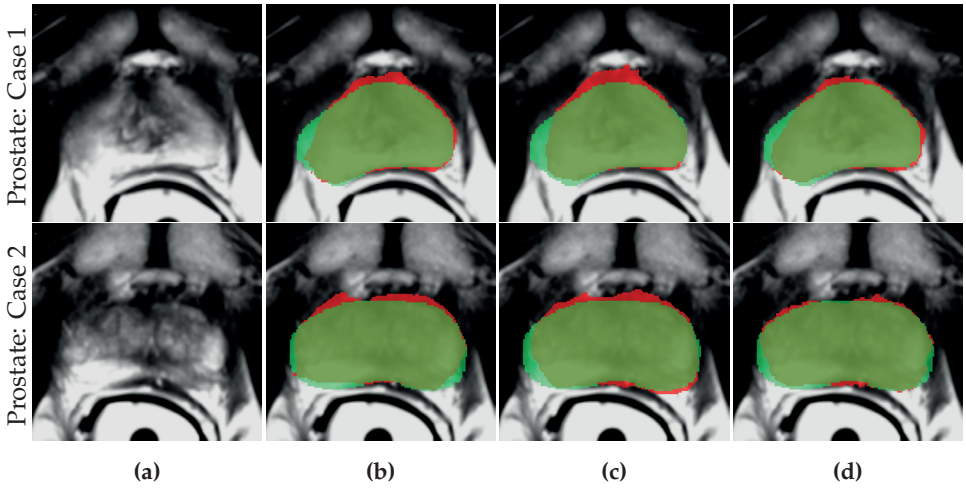


Figure 5.6: Predictions of the experiments of the prostate segmentation task trained on 30 MR scans. (a) Shows the original slice, and the training strategies (b) scratch ExpPS30, (c) vanilla transfer learning ExpPT30, and (d) transfer learning with fine-tuning ExpPF30.

transfer learning from the base model benefits four new segmentation tasks (esophagus, lung lobes, prostate, and vertebrae). To utilize this limited annotated dataset,

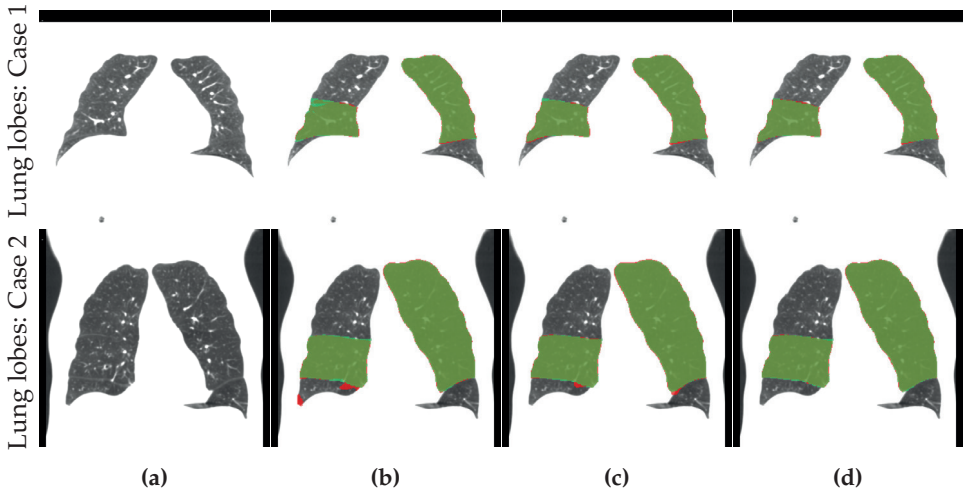


Figure 5.7: Predictions of the experiments of the lung lobes segmentation task trained on 30 CT scans. (a) Shows the original slice, and the training strategies (b) scratch ExpLS30, (c) vanilla transfer learning ExpLT30, and (d) transfer learning with fine-tuning ExpLF30.

we trained the base model to learn from only the annotated regions and ignored other regions which could belong to one of the other classes (e.g., kidneys in scans in which only annotations of the liver are present). We analyzed different training strategies (training from scratch, vanilla transfer learning, and transfer learning with fine-tuning) and the influence of the training set size. We found that initializing a 3D U-Net with the learned parameters of the base model is beneficial, especially when only a limited number of annotated scans for the new task are available. Models trained with small datasets (10 scans) that use transfer learning performed comparable with models trained from scratch with 40 or more scans. This finding aligns with previous research that showed that transfer learning reduces the need for large amounts of annotated data and obtains better performance than networks trained from scratch. Most importantly, for datasets with up to 50 annotated images, transfer learning from the generic base model never hurt performance and can therefore be generally recommended. However, transfer learning might not be necessary if a larger dataset is available. Nevertheless, it remains advantageous for reducing computational costs and mitigating carbon emissions due to faster convergence. Note that the initial base model used in our experiments benefits from the multi-center data used to compose the large sparsely annotated dataset. This dataset provides a diverse range of examples and may help the model generalize better to new chal-

lenging tasks. For instance, the results of the lung lobes segmentation task (see Figure 5.7) show how difficult it is to separate lung lobes. While the prostate segmentation task (see Figure 5.6) shows both transfer learning strategies obtained more consistent results than the experiment trained from scratch. Note the predictions of the transfer learning with fine-tuning training strategy obtain better results than the vanilla transfer learning, except on the vertebrae segmentation task where the difference among training strategies is small.

Moreover, we investigated whether transfer learning from a model trained as proposed is more beneficial than transfer learning from a simpler model, trained for a single task and with a smaller dataset. While both transfer learning training strategies generally resulted in improved performance, especially for small training sets, transfer learning from a more generic model trained with many segmentation tasks consistently improved the performance, this was more task-dependent when using a single-task base model (Table 5.4). The results presented in Tables 5.3 and 5.4 provide insight into the performance of different transfer learning training strategies for organ segmentation tasks. Our findings suggest that vanilla transfer learning from the base model (Exp0) performs better than vanilla transfer learning from the simpler base model (ExpVS90), except for two experiments in the prostate segmentation task where the simpler base model slightly outperforms Exp0. Furthermore, the simpler base model consistently performs better than experiments trained from scratch in all tasks except for the lung lobes segmentation task. This comparison suggests that transfer learning can be a useful approach for improving the performance of deep learning models in medical imaging, especially when annotated data is limited. Although we trained the base model on CT scans, transfer learning from the base model (Exp0) was also favorable for the prostate segmentation task on MR images. The simpler base model (ExpVS90) slightly obtained higher results in two experiments than the proposed base model.

While multi-center data is beneficial for generalization, the difference among annotation tools and protocols in datasets may difficult the learning process of the network. For instance, Figure 5.4 and Figure 5.6 show that the annotations (green regions) of the esophagus and prostate are not consistent in the sagittal orientation. The annotation tool was designed to annotate the structures in a single orthogonal orientation, and no corrections were made on the other orthogonal orientations. Moreover, all the networks learned to segment a continuous region. Similarly, the dataset VerSe19 annotation protocol skips a partially present vertebra in the CT scan (see Figure 5.5).

This study has a few limitations. First, some CT scans of the esophagus dataset were present in the large sparsely annotated dataset to train Exp0 but without the an-

notations of the esophagus. This data overlap may bias the transfer learning results for the esophagus segmentation task; other segmentation tasks (lung lobes, prostate, and vertebrae) contain unseen CT scans. Moreover, this paper aims to provide a base model that benefits new medical segmentation tasks and not to obtain high performance on the large sparsely annotated dataset.

Second, although the difference in performance between experiments trained from scratch and using transfer learning was small when training with more than 50 CT scans, transfer learning experiments may be more robust in unseen images due to the knowledge learned from the multi-center dataset; future research should investigate whether this improved generalization is indeed observed.

Compared to our approach, Chen et al.[129] created a binary prediction per each of the eight structures, while our approach uses a single softmax layer to obtain the predictions. Federated learning has gained significant attention in recent years as an approach to facilitate multi-data center learning without the necessity of sharing sensitive patient data[151–153]. It is noteworthy that while both federated learning and our study address sparsely annotated data, our study is not directly compatible with federated learning due to the centralization of data into a single large but sparsely annotated dataset. This centralization approach, while effective for our research goal, differs from the decentralized nature of federated learning.

Overall, our study highlights the potential of transfer learning for organ segmentation in medical imaging, and our results provide valuable insights for researchers and practitioners looking to optimize the performance of deep learning models in this domain. Further investigation may determine the optimal approach for selecting a base model and understanding the factors contributing to the performance differences observed in different tasks.

5.7 Conclusions

In conclusion, this study demonstrates the effectiveness of transfer learning in improving the performance of deep learning models in medical imaging, mainly when annotated data is limited. Our results indicate that the learned features of a network trained on a partially annotated dataset can be transferred to new segmentation tasks, providing significant benefits, particularly on tasks where annotated data is scarce. Our experiments show that transfer learning can be applied successfully to four segmentation tasks (esophagus, lung lobes, vertebrae, and prostate) and can significantly reduce the need for extensive annotation efforts. Additionally, we have demonstrated that cross-modality transfer learning can be effective, as shown by our results in prostate segmentation in MR scans. Furthermore, we found that fine-

tuning the pre-trained base model before transfer learning is more beneficial than using vanilla transfer learning. However, further research is needed to explore the limitations and potential applications of transfer learning in medical imaging and to develop more effective methods for utilizing sparsely annotated datasets.

Appendix

Comparison of training strategies

When comparing results per training strategy, the experiments that used transfer learning with fine-tuning obtained higher results than the other training strategies (scratch and vanilla transfer learning), except for the vertebrae experiments. This improvement gradually reduces when increasing the training set size, see Figure 5.3. Transfer learning with fine-tuning was more beneficial to the esophagus experiments with small training sets, where the difference in Dice score between training strategies reached +0.129 Dice (ExpEF10 0.588 - ExpES10 0.459) when training on ten scans. Similarly, the prostate segmentation task experiments show that transfer learning is beneficial in most cases, except for ExpPF10, which performs similarly to ExpPS10. The largest difference in the prostate segmentation reached +0.043 Dice (ExpPF20 0.863 - ExpPS20 0.819) when training on 20 scans. While the lung lobes and vertebrae segmentation tasks show a slight improvement when using transfer learning with fine-tuning.

We compared two scenarios, networks trained on limited data and networks trained on large training data. Our results show that transfer learning benefits experiments with limited data (usually up to 30 images). Ideally, transfer learning would reduce the need for large training sets. For instance, the difference between the experiment trained with 10 CT scans and the experiment trained from scratch on the full dataset of the esophagus decreased from +0.235 (ExpES158 - ExpES10) to +0.106 Dice (ExpES158 - ExpEF10) after using transfer learning. Moreover, the transfer learning experiment with fine-tuning reached 0.666 Dice when training with 40 images, reducing the difference to +0.027 in Dice score (ExpES158 - ExpEF40). Note the transfer learning experiment reaches 0.027 difference in Dice score with 118 fewer images in the training set than ExpES158. For the lung lobes segmentation task, the performance difference between training with the entire training set and training from scratch using ten images reached +0.052 (ExpLS90 0.969 - ExpLS10 0.917). Transfer learning with fine-tuning reduced that difference to +0.028 (ExpLS90 0.969 - ExpLF10 0.941). Moreover, the transfer learning experiment with fine-tuning reached 0.961 Dice when training with 40 images, reducing the difference to 0.008

Dice (ExpLS90 - ExpLF40). Note the transfer learning experiment reaches a 0.008 difference in Dice score with 50 fewer images in the training set than ExpLS90. Similarly, the difference for the vertebrae segmentation task reached +0.036 (ExpVS90 0.956 - ExpVS10 0.920). The difference is already small; however, transfer learning with fine-tuning slightly reduced that difference to 0.027 (ExpVS90 0.956 - ExpVF10 0.929). For this segmentation task, both transfer learning strategies obtained similar results. The vanilla transfer learning experiment trained on 30 images reached a 0.942 Dice score, 0.014 less than the experiment trained from scratch on the full dataset; these experiments have a difference of 60 images in training sets. Transfer learning did not benefit the prostate segmentation task with the smallest training set, as with previous segmentation tasks, but with more images in the training set. For instance, when using 20 images as training set, the difference with the experiment trained from scratch on the full dataset decreased from +0.033 (ExpPS45 0.852 Dice - ExpPS20 0.819 Dice) to -0.01 (ExpPS45 0.852 Dice - ExpPF20 0.862 Dice) after using transfer learning. This shows that transfer learning with fine-tuning on 20 scans obtained already a higher performance than the experiment trained from scratch on the full dataset. The difference in performance among training strategies when using the full dataset reached +0.032 (ExpPF45 - ExpPS45) for the prostate segmentation task; this shows the benefits of cross-modality transfer learning. Note that the same difference in the CT segmentation tasks is lower (+0.0001 lung lobes, +0.015 esophagus, and -0.0012 vertebrae). These results show that transfer learning was highly beneficial when having limited data, boosting performance and reducing the need for a large annotated dataset to obtain high segmentation performance.

Transfer learning increased the performance of segmentation tasks compared to experiments trained from scratch (see Table 5.3 and Figure 5.3). Moreover, the experiments that use transfer learning with fine-tuning obtained higher results than vanilla transfer learning. These results show that gradual training (fine-tuning) boosts the network's performance of the four new segmentation tasks, including the prostate segmentation (cross-modality transfer learning), which has MR data only. Raghu et al.[115] showed that different domain transfer learning does not improve performance. This paper shows that the same domain/modality transfer learning is beneficial for the medical domain, including different modalities (CT to MR prostate).

Results of experiments trained on 10 scans

This section shows the results of the experiments trained on 10 scans. This figures are 5.8, 5.10, 5.11, and 5.9.

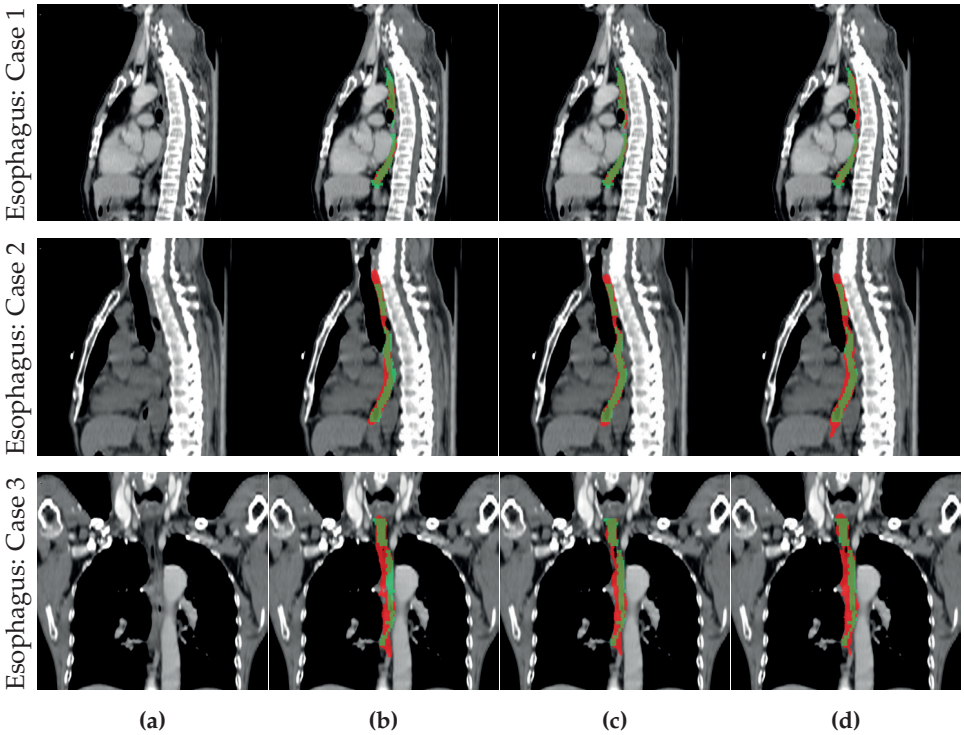


Figure 5.8: Predictions of the experiments of the esophagus segmentation task trained on 10 CT scans. (a) Shows the original slice, and the training strategies (b) scratch ExpES10, (c) vanilla transfer learning ExpET10, and (d) transfer learning with fine-tuning ExpEF10.

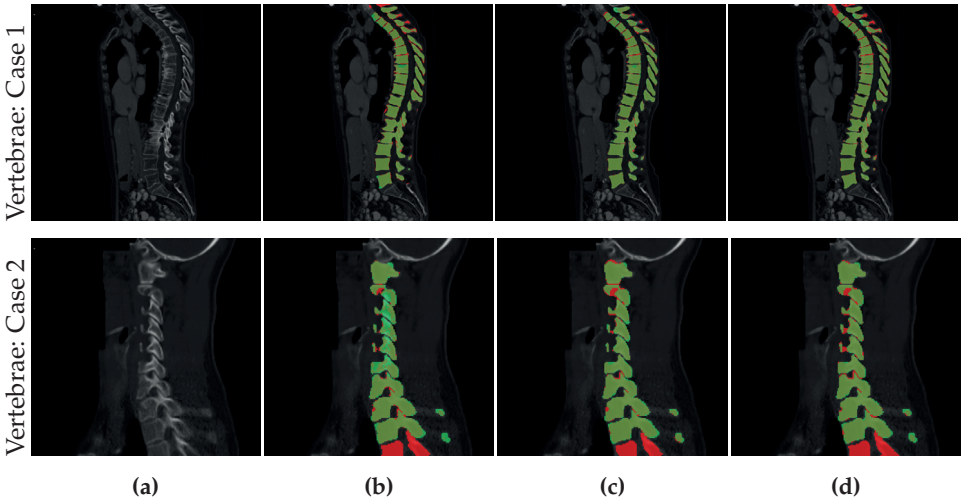


Figure 5.9: Predictions of the experiments of the vertebrae segmentation task trained on 10 CT scans. (a) Shows the original slice, and the training strategies (b) scratch ExpVS10, (c) vanilla transfer learning ExpVT10, and (d) transfer learning with fine-tuning ExpVF10.

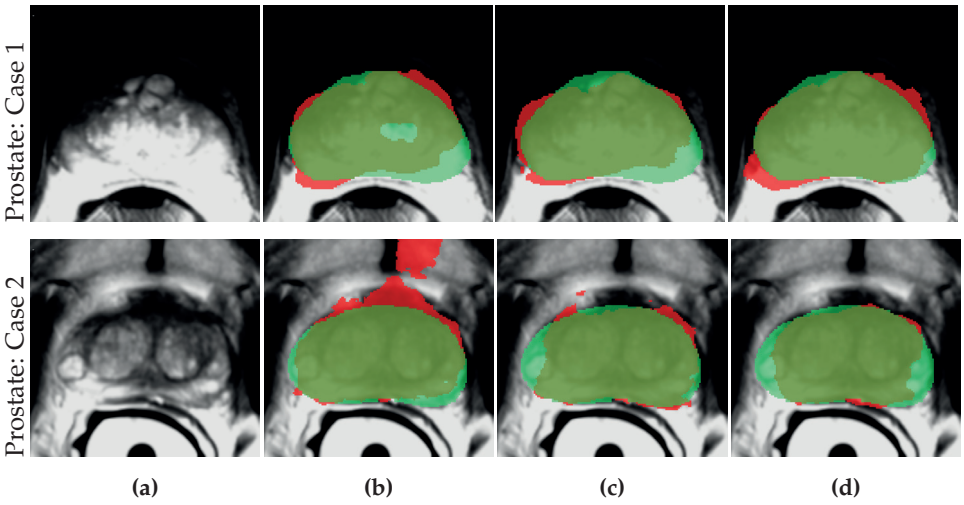


Figure 5.10: Predictions of the experiments of the prostate segmentation task trained on 10 CT scans. (a) Shows the original slice, and the training strategies (b) scratch ExpPS10, (c) vanilla transfer learning ExpPT10, and (d) transfer learning with fine-tuning ExpPF10.

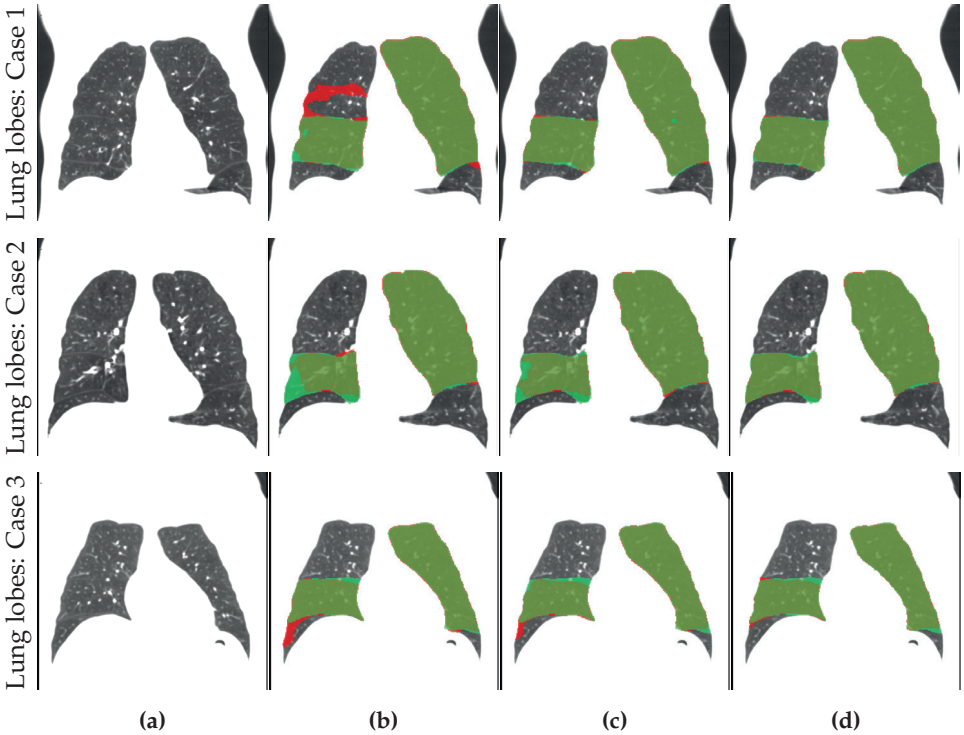


Figure 5.11: Predictions of the experiments of the lung lobes segmentation task trained on 10 CT scans. (a) Shows the original slice, and the training strategies (b) scratch ExpLS10, (c) vanilla transfer learning ExpLT10, and (d) transfer learning with fine-tuning ExpLF10.

General Discussion

6

In this thesis, we described deep learning methods for detection and localization and for segmentation of structures and organs in CT scans from patients referred by the oncology department. These patients typically have complex pathologies in various body regions. Therefore, developing automated algorithms capable of detecting, localizing, and segmenting the affected structures and organs is not an easy task but doing so may create a valuable tool for radiologists and oncologists, providing them with precise and quantitative information to support diagnosis, assess treatment response, and do treatment planning. Thus, the methods developed in this thesis have the potential to make an impactful contribution to clinical practice. In this chapter, we discuss the main contributions of this thesis, describe limitations, and propose potential directions for future research.

Structure localization

Structure localization is particularly challenging in patients with complex anatomy and multiple abnormalities[17,18,30]. To address this challenge, we developed a multilabel deep learning algorithm that simultaneously localizes multiple structures in CT scans in Chapter 2.

The localization of certain organs in medical imaging can be challenging due to various factors. In our study, we found that the sinister wall of the liver was particularly difficult to localize as this region of the liver contains only a small tip; the longer the tip, the more complex it is to provide a precise localization. The gallbladder was also challenging to detect, likely due to its relatively small size. Similarly, the bladder was also a difficult organ to localize due to the variation in its size among individuals. Additionally, using a single slice as input made it difficult to differentiate between the left and right lungs in the sagittal view, while experiments with multiple slices showed improved performance, indicating the importance of providing 3D context to the network.

Our approach enhanced the network's contextual awareness by adding adjacent slices as input. Future research may expand this contextual scope by integrating recurrent neural networks (RNNs) such as long short-term memory (LSTM)[154] networks alongside 2D/3D convolutions. These combined architectures assess a sequence of slices where a prediction of a slice influences the upcoming ones, leveraging a broader context to achieve smoother and more refined predictions.

A limitation of our method is that it requires access to large amounts of CT scans for training and validation. To address this limitation, it is important to explore other methods that can use less data for training and still generate robust results. This could involve investigating the use of transfer learning, where pre-trained models

are adapted to new tasks with limited annotated data. With the recent advancements in Natural Language Processing (NLP), there has been a surge of interest in Large Language Models (LLMs)[155]. Few-shot learning methods[156], where only a few annotated samples would be sufficient to optimize a model for a new downstream task, have also demonstrated promising performance. This approach holds great potential for the detection of lesions and abnormalities within the human body, including rare conditions where only a limited number of samples are accessible. Although few-shot learning performs well for NLP, research indicates that it is a challenging task in computer vision, and while the concepts of burstiness and a large set of rarely occurring classes do occur in the context of few-shot learning with images, they are not as naturally prevalent as in natural language data[157].

Finally, structure localization methods may help CAD systems, as they could assist radiologists in identifying and localizing structures of interest more efficiently. Furthermore, it has the potential to improve the accuracy of CAD systems by accurately indicating to which structure an abnormality belongs.

Organ segmentation

Organ segmentation is important in medical image analysis, particularly for accurate diagnosis, treatment planning, and disease monitoring[131]. Measurements obtained from binary segmentations are essential in many medical tasks, such as determining tumor size, assessing disease progression, and monitoring treatment response. To address organ segmentation, we used deep learning segmentation models in Chapters 3, 4, and 5.

In Chapter 3, our research focused on the spleen, an important organ for assessing disease progression, and the splenic volume change (SVC) which can occur as an effect of chemotherapy, infection, and lymphoma[44–51]. We presented a segmentation-based algorithm that accurately measures splenic volume change over time. Our segmentation method (automatic SVCa) obtained 89% agreement with the reference standard; these findings suggest that precise segmentations can be useful for radiologists in assessing SVC. Our study showed that automatic SVCa, based on spleen segmentation, achieved a more precise splenic volume approximation than the widely used in clinical practice splenic index equation (visual SVCa). Our experiments showed that while deep learning networks require substantial data for optimal performance, data quality is crucial in achieving a robust network. Additionally, the presence of rare conditions, such as the beavertail liver in the test set, had an impact on the performance of our algorithm. Similarly, our method returned false-positive regions in scans from patients who had undergone splenectomy. These

false positive regions occurred in regions where the spleen is usually located. While we made efforts to incorporate challenging cases into the training set, it remains difficult to encompass the full spectrum of abnormal conditions, which would need extensive sampling to achieve a balanced dataset and enhance its representativeness in the training set. Due to the wide range of pathologies that oncology patients may have, future research should expand this dataset to tackle SVC in different subgroups, for example groups with different ethnicities, and increase the robustness of the method to reliably use this method in clinical practice. Moreover, the hard threshold to determine SVC classification is a limitation of our study as it was set empirically to $\pm 25\%$. Future research should analyze the threshold to determine the SVC classification which could lead to more relevant results for radiologists.

To tackle kidney cancer, the 12^{th} most deadly cancer worldwide[3], in Chapter 4 of this thesis, we focused on the kidney parenchyma and kidney abnormalities segmentation in patients referred for CT imaging by the oncology department. While the results are comparable for the kidney parenchyma class across the experiments, notable variations emerge when considering the kidney abnormalities class. The conducted ablation study has provided valuable insights, showing that the experiments with systems with a greater number of modules, Experiment 1 and Experiment 2, obtained the best performance in segmenting the kidney parenchyma and kidney abnormalities. The influence of individual modules on the segmentation of kidney abnormalities was examined in Table 4.1, with Experiment 3, involving data augmentation, showing relatively minimal impact on performance. In contrast, as demonstrated in Experiment 4, the multi-resolution module emerged as the most influential in enhancing segmentation results, followed closely by spatial dropout in Experiment 1 and the top-k module in Experiment 2. In addition to our approach, we also evaluated nnUNet, which performed similarly to our approach when segmenting the parenchyma but was better when segmenting the kidney abnormalities.

The performance of nnUNet improved after adding our dedicated postprocessing(nnUNet = 0.521, nnUNet + postproc = 0.576) which shows that SOTA algorithms can still be benefited by customized postprocessing.

Due to the complexity of the kidney abnormality segmentation task, our experiments, including our best experiment (Dice score=0.585) could not outperform the second observer (Dice score = 0.664). A limitation of our study is the data annotation protocol which excluded abnormalities in the collecting system. However, it is noteworthy that a significant portion of the false positives generated by our methods were tumors located in the collecting system. These abnormalities can exhibit visual characteristics that can be confused with kidney cysts, thereby contributing to false positive segmentations. Figure 4.6c shows how the second observer made

similar errors, showing the difficulty of this issue. For better performance, future research could focus on independently annotating the classes that we grouped as kidney abnormalities. While this chapter contributed to the segmentation of kidney parenchyma and kidney abnormalities, future research could focus on automating nephometry scoring systems[79] to offer a more standardized output to radiologists. Our training data including annotations is available to the research community under a CC-BY 4.0 license on `https://doi.org/10.5281/zenodo.8014289`.

Transfer learning

Publicly available datasets and challenges have facilitated the development and evaluation of automatic organ segmentation methods. They provide large amounts of medical imaging data with annotations and ground truth labels, leading to rapid advances in the field.

In Chapter 5, we trained a segmentation network on a sparsely annotated dataset to use as a base model for transfer learning. We trained a model to learn from only the annotated regions and ignored other regions, which could belong to one of the other classes. Our results showed that transfer learning with fine-tuning was most beneficial to models trained with very small datasets. With larger datasets, transfer learning never hurt performance compared to training from scratch, but the benefits were unfortunately only minor.

We also tried transfer learning from a simpler model (ExpVS90), and the results showed that training from our sparsely annotated dataset was better; still, transfer learning from a simpler model is beneficial compared to training from scratch. Moreover, cross-modality transfer learning was also beneficial, from our sparsely annotated dataset (CT scans) to prostate segmentation task on MR images.

While utilizing multi-center data can enhance the robustness of models, it is important to acknowledge that discrepancies in annotation tools and protocols across datasets can introduce complexities into the network's learning process. Future research should focus on analyzing the robustness of the models after transfer learning; we believe the models trained with our dataset are more robust since they were trained with multi-center data. A limitation of this study was the utilization of a partially annotated validation set; a fully annotated validation set can lead to more robust weights for the base model. Despite nnUNet's status as the State-of-the-Art (SOTA) in medical imaging segmentation tasks, it predominantly relies on fully annotated datasets. Therefore, there is a potential avenue for further improvement through adaptations that enable training with sparsely annotated datasets.

The algorithms introduced in this thesis are part of an extensive list of algorithms

available for testing at `https://grand-challenge.org/algorithms/`. This comprehensive repository of algorithms is dedicated to enhancing the reporting process for CT scans of oncology patients.

Future research

While this thesis has contributed to the field of medical image analysis, there are still limitations and challenges that need to be addressed.

For instance, the algorithms we developed were trained and tested on a specific patient population and scanner type, and their generalizability to other populations and imaging modalities has not been studied in this thesis. Therefore, future research should focus on validation on a wider range of populations and imaging modalities to ensure that we obtain robust and generalizable algorithms that can be applied to a wider range of patients and imaging settings. Nonetheless, the work presented in this thesis is an important step towards using deep learning methods to assist in the clinical assessment of oncology patients and has the potential to improve patient outcomes and quality of life.

This thesis primarily employed data derived from patients referred to the oncology department, and the methodologies developed in this thesis were trained with patients exhibiting oncological abnormalities. One of the limitations is the absence of data containing non-oncological abnormalities within our dataset. Other studies tackled this issue by annotating more data[158,159]; this is not always feasible as data annotation is expensive and time-consuming. Other domains solve the lack of data by applying data augmentation to the training data to obtain a large variety of data. In the 3D medical imaging domain, the typical data augmentation (rotation, scaling, shifting, elastic deformations, etc.) cannot create synthetic abnormalities in the CT scans. Generative adaptive networks (GANs) are a potential solution for this problem by creating realistic synthetic 3D data. Studies show that GANs improved the results of tumor segmentation after adding 2D synthetic patches that added tumors[160]. To the best of our knowledge, this was not done to generate 3D realistic patches yet[161–163].

In recent years, object detection architectures, such as YOLO[164], have gained significant interest and can be applied for organ localization within medical imaging. By using an object detection framework across all orthogonal views within a CT scan, we could yield a coarse 3D segmentation. Furthermore, architectures like Mask R-CNN[41,165] offer the capability to simultaneously derive a 3D bounding box and segmentation, potentially enhancing the precision and comprehensiveness of the organ localization network.

YOLO and Mask R-CNN use a single image as their input. Enhancing the input with additional slices or contextual information could potentially lead to improved outcomes, and RNNs can be advantageous. Furthermore, the adoption of a 3D approach for organ localization has the potential to encompass a broader contextual scope, thereby facilitating more precise localization results.

Recent advancements in the field, such as nnDetection[166], have adopted a self-configuring methodology as nnUNet[76] for localizing structures within medical images. As the availability of publicly available datasets and pre-trained models continues to expand, the demand for fully annotated training datasets for new tasks will reduce significantly. Instead, the focus may shift towards annotating abnormalities within 2D slices rather than fully annotated organs in CT scans. Exploring a synergistic approach that combines nnDetection with few-shot learning holds the potential for substantial enhancements in the domain of medical image analysis.

While nnUNet has established itself as the SOTA solution for medical segmentation tasks, applying these networks requires large computational resources. This slow inference response time poses a significant issue to its clinical adoption. In a recent development, Huang et al.[167] addressed this challenge by devising a more streamlined version of nnUNet, achieving a remarkable 10 times boost in inference speed, at the cost of a slight performance trade-off. Further optimization efforts could enhance nnUNet's speed, rendering it practical for clinical use. Moreover, nnUNet, which is based on 3D U-Net architecture introduced in 2016, could benefit from the integration of newer deep learning techniques, such as transformers[155,168], which have emerged since its inception. These innovations may offer opportunities for performance enhancement without creating complex architectures. The future evolution of nnUNet may involve modest architectural refinements and the utilization of partially annotated data to further boost its capabilities.

To ensure the practical implementation of these automated methods, future research can focus on integrating them into existing clinical workflows and evaluating their impact on radiologists' efficiency[169] and diagnostic accuracy. This can involve conducting studies to assess the benefits of using these tools in reducing interpretation time, improving inter-observer variability, and enhancing clinical decision-making.

In summary, the methods presented in this thesis have the potential to improve the accessibility and applicability of deep learning algorithms for medical image analysis in various clinical settings.

Summary

Both the number of scans acquired and the size of individual scans are growing rapidly. As a result, radiologists face increasing workloads. This is especially the case in CT scans of oncology patients. Analyzing such CT scans is time-consuming and strenuous. These scans often contain many small details, making it easy to miss subtle abnormalities. Automated tools can offer significant advantages in terms of both time and accuracy. By automating specific tasks, such as detecting and segmenting organs and abnormalities, radiologists can save time and focus on more critical aspects of their work, such as diagnosis and treatment planning. Moreover, these tools can help reduce inter-observer variability, especially when tracking changes over time. Therefore, the implementation of automated tools can enhance the efficiency and accuracy of oncological radiological analysis, leading to better patient treatment.

3D structure localization is an essential pre-processing step for computationally intensive tasks, where limiting the analysis to an area of interest may speed up the analysis (e.g., 3D full segmentation and registration). Initial studies on structure localization employed a single network per orthogonal view to localize an organ; a method that is not scalable for localizing multiple organs. To address this problem, **Chapter 2** describes a fully automatic method for multi-structure localization; we propose utilizing a single network for each orthogonal view, designed to localize multiple structures simultaneously. Our method uses a sigmoid activation in the last layer to allow multi-label predictions. We observed that identifying the presence of an organ on a single 2D slice may be challenging for some structures. For instance, radiologists find it challenging to localize the fibrous appendix of the liver (the last tip of the liver) in the sagittal view when limited to a single 2D slice. Moreover, distinguishing between the left and right lung may be difficult in the sagittal view. Typically, radiologists scroll through the CT scan to address this issue. To emulate this approach, we added the surrounding slices to the slice of interest as an additional input to the network. These surrounding slices provide the network with additional contextual information, aiding in the localization of organs in challenging regions such as the fibrous appendix of the liver. We showed that multi-label networks boost their performance when adding the surrounding slices to the slice of interest as input. We used a large dataset (1884 CT scans) to develop this method and applied extensive data augmentation. Our best configuration achieved an average wall distance of 3.20 ± 7.33 mm in the test set, while the human observer obtained 1.23 ± 3.39 mm.

Patients with cancer undergo invasive treatments that can result in various side effects (anemia, fatigue, hair loss, nausea, and organ volume change). Organ volume change can be essential information for the oncological team and can provide

an indication that the treatment plan of a patient should be adjusted. The spleen is one of the organs that may have its volume altered after cancer treatment. In clinical practice, the splenic index equation approximates the volume of the spleen, and isderived from simple length measurements. In **Chapter 3**, we propose a fully automatic spleen segmentation method using deep learning to obtain a precise approximation of the splenic volume. We employ a 3D U-Net architecture to segment the spleen in 3D thorax-abdomen CT scans. In our first experiment, we trained a 3D U-Net on a small set (100 CT scans). The resulting first network helped to identify relevant cases (300 out of 1000 CT scans) where the network failed on a large scale. Our annotators took the predictions of the first network as a starting point to correct the annotations. We trained a second 3D U-Net using the corrected annotations from the relevant cases (300 CT scans) + the initial training set (100 CT scans). In our initial experiment, with the U-Net trained on 100 CT scans, we obtained 0.950 ± 0.040 Dice score, while the second 3D U-Net (trained on 400 relevant scans) obtained 0.962 $\pm$ 0.016 Dice score in a test set of 50 CT scans. An independent observer obtained 0.964 ± 0.012. When comparing the relative absolute volume, the splenic index obtained an error of 16.6%, the first 3D U-Net 5.99%, the second 3D U-Net 4.39%, and the independent observer 3.94%. In a qualitative observer experiment, the observer had an 81% (81 scan pairs of 100) agreement with the reference standard for visual classification of volume change; this agreement increased to 92% when aided by our algorithm. Moreover, the radiologist scored the quality of 94% of the predicted segmentations as ready for clinical use.

The presence of anatomical abnormalities is common in patients with cancer, particularly in advanced stages of the disease. Accurately quantifying these abnormalities and keeping track of them over time is crucial for monitoring the progression of cancer and evaluating the effectiveness of treatment. CT scans are often used to visualize these abnormalities, but interpreting these images can be time-consuming and subjective. This is where automated tools like segmentation networks can be particularly helpful. In **Chapter 4**, we used convolutional neural networks to segment kidney abnormalities and kidney parenchyma in CT scans. To create a balanced dataset, we selected CT scans with kidney abnormalities from radiology reports and added scans without abnormalities. We conducted an ablation study to evaluate the effectiveness of five modules in our proposed segmentation network. The results show that our system using all proposed modules obtained the highest score (0.487 ± 0.314) in segmenting the kidney abnormalities among the ablation study experiments. The system that omits the last module obtained the highest score (0.957 ± 0.006) in segmenting the kidney parenchyma in patients without kidney abnormalities among the ablation study experiments. As an additional experiment, we trained the state-

of-the-art nnUNet on our data and obtained a 0.521±0.303 Dice score for kidney abnormalities, an increase of 0.034 compared to our best ablation study experiments. By applying our dedicated postprocessing, we improved the performance of nnUNet by 0.055, reaching 0.576±0.290 Dice score. After ensembling the nnUNet with our best-performing method, we reached a 0.585±0.293 Dice score for kidney abnormality segmentation. While all our experiments outperformed an independent human observer in segmenting the kidney parenchyma in patients without kidney abnormalities, this human observer obtained higher score in segmenting the kidney abnormalities. Thus, more research is needed to further improve automated segmentation of kidney lesions.

In contrast to traditional machine learning approaches, deep learning performance improves with more training data. Although hospitals have large amounts of medical imaging data, these are typically not annotated. Data annotation is tedious and time-consuming, limiting the opportunity to have large annotated datasets for medical imaging projects. Instead, medical imaging projects typically start with a small number of annotations to train a deep-learning network. Robustness is one of the main problems of methods trained on small training sets; they may fail when applied to data obtained with input settings that differ slightly from those used to obtain the scans in the (small) training set. Transfer learning may be a solution for transferring knowledge from a large to a small dataset. Unfortunately, transfer learning has not yet been widely used in the medical domain. In **Chapter 5**, we performed transfer learning from a model trained on a partially annotated dataset to four new segmentation tasks. We joined annotations of six publicly available datasets and medical segmentation challenges to compose our partially annotated dataset. We trained a 3D U-Net using the partially annotated dataset; this network uses a weight map that forces the network to learn from the annotated regions only. We used the optimal model as a pre-trained model for four additional segmentation tasks. One of these tasks used data from a different domain, MRI scans. The tasks that we addressed were vertebrae, esophagus, lung lobes, and prostate segmentation. We evaluated whether transfer learning benefits the segmentation task. Additionally, we evaluated the influence of the training set size in the new segmentation tasks. Our experiments show that transfer learning benefits the performance of the segmentation task, especially when the available training set is of limited size. For instance, the esophagus segmentation task obtained 0.459±0.245 Dice score when training from scratch and obtained 0.588±0.193, an increase of +0.129, in Dice score when using transfer learning with fine-tuning step on a dataset of 10 CT scans. The improvement for the vertebrae, lung lobes, and prostate segmentation were +0.009 (from 0.920 to 0.929), +0.024 (from 0.917 to 0.941), and +0.002 (from 0.816 to 0.818).

Cross-modality transfer learning from CT to MR data was also beneficial with up to +0.043 (from 0.819 to 0.862) increase in Dice score for the experiment with 20 MR scans in the training set.

In summary, this thesis presented a range of methods for structure localization and organ segmentation that may contribute to automating radiology reports and can be incorporated in computer-aided detection and diagnosis systems.

Samenvatting

Dit proefschrift bevat resultaten van ontwikkelde computersystemen die medische beelden analyseren met behulp van deep learning, met name convolutionele neurale netwerken. We hebben ons vooral gericht op de analyse van CT scans van de borst en buik van mensen met kanker.

Er worden steeds meer scans gemaakt en die scans bevatten gemiddeld steeds meer beelden. Dat betekent dat radiologen het steeds drukker krijgen. Dit geldt vooral voor het verslaan van CT-scans van patiënten met kanker. Het duurt lang om deze scans goed te bekijken, en het is makkelijk om kleine details over het hoofd te zien. Door radiologen te ondersteunen met computerprogramma's hopen we dit proces sneller en nauwkeuriger te maken.

Door bepaalde taken, zoals het herkennen van organen en afwijkingen, aan de computer over te laten, kunnen radiologen zich focussen op belangrijkere dingen zoals het stellen van een diagnose en het bedenken van een behandelplan. Ook zorgen deze computerprogramma's ervoor dat radiologen meer op één lijn zitten, vooral bij het inschatten van veranderingen over tijd. Dus door deze tools te gebruiken, kunnen we de zorg voor mensen met kanker verbeteren.

Voordat de computer een specifieke analyse van een scan maakt, is het handig om eerst te bepalen welk deel van de scan hiervoor moet worden geanalyseerd. Dat noemen we 3D-structuurlokalisatie. Dit is het onderwerp van **Hoofdstuk 2**. In eerdere studies gebruikten onderzoekers hiervoor één computernetwerk per kijkrichting van de scan, maar dat is niet efficiënt als je meerdere organen of gebieden tegelijk wilt vinden. Om dat op te lossen, hebben we een nieuwe methode bedacht. Die gebruikt ook één netwerk per kijkrichting, maar kan dan wel meerdere organen tegelijk vinden. We gebruiken hier een speciale techniek voor, genaamd 'sigmoid activation', waardoor het netwerk meerdere labels kan voorspellen.

We zagen dat het soms lastig is om een orgaan te vinden als je maar één 2D-plaatje in één richting bekijkt. Bijvoorbeeld, het laatste stukje van de lever is moeilijk te zien in zo'n enkel plaatje. En het is ook lastig om het verschil tussen de linker- en rechterlong te zien in een sagittale doorsnede. Normaal gesproken scrollen radiologen door de scan om dit beter te kunnen zien. Daarom hebben we ons programma zo aangepast dat het ook de omliggende plaatjes meeneemt. Zo krijgt het netwerk meer informatie en kan het beter organen vinden die moeilijk te zien zijn.

We hebben dit getest met een grote verzameling scans (1884 stuks) en veel verschillende versies van de netwerken. Onze beste instelling haalde een gemiddelde nauwkeurigheid waarbij het 3.20 ± 7.33 mm van de referentie af zat. Dit kwam in de buurt van menselijke waarnemers; zij maakten een fout van 1.23 ± 3.39 mm.

Patiënten met kanker krijgen behandelingen die nogal wat bijwerkingen kunnen

hebben, zoals bloedarmoede, vermoeidheid, haaruitval, misselijkheid en veranderingen in de grootte van sommige organen. Als een orgaan van grootte verandert, is dat belangrijke informatie voor het behandelteam. Het kan namelijk betekenen dat het behandelplan aangepast moet worden. De milt is zo'n orgaan dat kan veranderen van grootte na een kankerbehandeling.

In de dagelijkse praktijk meten radiologen de milt op in een paar richtingen door een lijntje te trekken op het beeld. Dit levert, met behulp van een formule, de miltindex, een schatting voor het volume van de milt. Echt nauwkeurig is dit niet. In **Hoofdstuk 3** stellen we een nieuwe manier voor om met behulp van computers de grootte van de milt exact te meten. We hebben een computermodel getraind, een 3D U-Net, om dit te doen. Eerst hebben we 100 CT-scans gebruikt om het model te trainen. Daarna hebben we de fouten die het model maakte verbeterd en een tweede model getraind met 400 scans.

In ons eerste experiment, met de U-Net getraind op 100 CT-scans, behaalden we een Dice-score van 0.950 ± 0.040. Het tweede 3D U-Net (getraind op 400 scans) kreeg een Dice-score van 0.962 ± 0.016 in een testsample van 50 CT-scans. Een onafhankelijke menselijke waarnemer behaalde een Dice-score van 0.964 ± 0.012. Als we kijken naar de relatieve fout in het absolute volume, had de miltindex een foutmarge van 16.6%, het eerste 3D U-Net 5.99%, het tweede 3D U-Net 4.39%, en de onafhankelijke waarnemer 3.94%.

In een observerstudie behaalde een radioloog in 81% van de gevallen (81 scanparen van 100) overeenstemming met de referentiestandaard voor visuele classificatie van significante volumeverandering; deze overeenstemming steeg naar 92% toen ons algoritme ter ondersteuning werd gebruikt. Bovendien beoordeelde de radioloog de kwaliteit van 94% van de voorspelde segmentaties als klaar voor klinisch gebruik.

Het komt vaak voor dat er afwijkingen zichtbaar zijn in scans van patiënten met kanker, vooral in gevorderde stadia van de ziekte. Het nauwkeurig kwantificeren van deze afwijkingen en ze monitoren is cruciaal om de voortgang van de ziekte te volgen en de effectiviteit van de behandeling te evalueren. Het interpreteren van scans gemaakt op meerdere tijdspunten kan tijdrovend en subjectief zijn. Hier komen geautomatiseerde tools zoals segmentatienetwerken goed van pas. In **Hoofdstuk 4**, gebruikten we convolutionele neurale netwerken om afwijkingen aan de nieren en nierparenchym in CT-scans te segmenteren.

Om een gebalanceerde dataset te maken, hebben we CT-scans met nierafwijkingen geselecteerd door te zoeken naar trefwoorden in de radiologieverslagen, en we hebben scans zonder afwijkingen toegevoegd. Vervolgens hebben we allerlei net-

werken met verschillende nieuwe elementen getraind en getest. We hebben een ablatiestudie uitgevoerd om de effectiviteit van vijf modules in ons voorgestelde segmentatienetwerk te evalueren. Hierbij laat je steeds één element weg en kijk je wat het effect op de nauwkeurigheid is.

De resultaten laten zien dat ons systeem dat alle voorgestelde modules gebruikt, de hoogste Dice score behaalde (0.487 ± 0.314) bij het segmenteren van de nierafwijkingen in vergelijking met de andere experimenten in de ablatiestudie. Het systeem dat alleen de laatste module weglaat, behaalde de hoogste score (0.957 ± 0.006) bij het segmenteren van het nierparenchym bij patiënten zonder nierafwijkingen.

Als extra experiment hebben we een nnUNet, de state-of-the-art in automatisch segmenteren, getraind op onze data en behaalden we een Dice-score van 0.521 ± 0.303 voor nierafwijkingen, een stijging van 0.034 vergeleken met onze beste resultaten uit de ablatiestudie. Onze method bevat een speciale nabewerking, en door deze toe te passen op de output van de nnUNet verbeterden we de prestaties van nnUNet met 0.055, wat resulteerde in een Dice-score van 0.576 ± 0.290.

Na het combineren van de nnUNet met onze best presterende methode, behaalden we een Dice-score van 0.585 ± 0.293 voor het segmenteren van nierafwijkingen. Al onze experimenten presteerden beter dan een onafhankelijke menselijke waarnemer bij het segmenteren van het nierparenchym bij patiënten zonder nierafwijkingen. Maar deze menselijke waarnemer behaalde wel een hogere score bij het segmenteren van de nierafwijkingen. Er is dus nog ruimte voor verbetering voor geautomatiseerde segmentatie van nierletsels.

In tegenstelling tot traditionele machine learning benaderingen, verbetert de prestatie van deep learning als er meer trainingsdata beschikbaar zijn. Hoewel ziekenhuizen veel medische beelden hebben, zijn deze meestal niet geannoteerd, dat wil zeggen dat de precieze posities van organen en afwijkingen niet worden ingetekend. Dit soort data-annotatie, essentieel om computerprogramma's te trainen, is een langdurig en tijdrovend proces. Daarom beginnen deze projecten meestal met een klein aantal annotaties om een netwerk te trainen. Gebrek aan robuustheid is een van de belangrijkste problemen van methoden die zijn getraind op kleine datasets; ze kunnen falen als ze worden toegepast op data die enigszins afwijken van de scans in de (kleine) trainingsset.

Transfer learning kan een oplossing zijn om kennis over te dragen van een grote naar een kleine dataset. Helaas is transfer learning nog niet veel gebruikt in de medische sector. In **Hoofdstuk 5** hebben we transfer learning toegepast vanuit een model dat is getraind op een gedeeltelijk geannoteerde dataset.

We hebben vier nieuwe segmentatietaken onderzocht. Hiervoor hebben we de

data en annotaties van zes openbaar beschikbare datasets samengevoegd. We hebben een 3D U-Net getraind met deze gedeeltelijk geannoteerde dataset; ons netwerk gebruikt een speciale techniek die het in staat stelt om alleen van de geannoteerde gebieden te leren. We hebben daarna dit model gebruikt als een vooraf getraind model voor vier extra segmentatietaken. Een van deze taken gebruikte data uit een andere domein, namelijk MRI-scans. De taken die we hebben aangepakt waren segmentatie van wervels, slokdarm, longkwabben en prostaat. We hebben geëvalueerd of transfer learning voordelen oplevert voor de segmentatietaak. Daarnaast hebben we de invloed van de grootte van de trainingsset op de nieuwe segmentatietaken geévalueerd.

Onze experimenten laten zien dat transfer learning de prestaties van de segmentatietaak ten goede komt, vooral als de beschikbare trainingsset klein is. Bijvoorbeeld, de segmentatietaak van de slokdarm behaalde een Dice-score van 0.459 ± 0.245 bij training vanaf nul en bereikte een Dice-score van 0.588 ± 0.193, een toename van +0.129, bij het gebruik van transfer learning met een fijnafstemmingsstap op een dataset van 10 CT-scans.

De verbetering voor de segmentatie van wervels, longkwabben en prostaat waren respectievelijk +0.009 (van 0.920 naar 0.929), +0.024 (van 0.917 naar 0.941), en +0.002 (van 0.816 naar 0.818). Ook het overzetten van kennis tussen verschillende soorten scans, van CT naar MRI, was voordelig met een toename van maximaal +0.043 (van 0.819 naar 0.862) in Dice-score voor het experiment met 20 MRI-scans in de trainingsset.

Samenvattend heeft dit proefschrift een reeks methoden gepresenteerd die kunnen worden opgenomen in systemen die artsen helpen bij het stellen van diagnoses op basis van CT scans van de thorax en de onderbuik en die later mogelijk zelfs zouden kunnen bijdragen aan het automatiseren van het opstellen van radiologierapporten.

Publications

Papers in international journals

G.E. Humpire Mamani, A.A.A. Setio, B. van Ginneken and C. Jacobs. "Efficient organ localization using multi-label convolutional neural networks in thorax-abdomen CT scans", In: *Physics in Medicine and Biology*, 2018;63(8):085003.

G.E. Humpire Mamani, J. Bukala, E. Scholten, M. Prokop, B. van Ginneken and C. Jacobs, "Fully Automatic Volume Measurement of the Spleen at CT Using Deep Learning", *Radiology: Artificial Intelligence*, 2020;2(4):e190102.

P. Bilic, P. Christ, H.B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, **G.E. Humpire Mamani**, G. Chartrand, F. Lohfer, J.W. Holch, W. Sommer, F. Hofmann, A. Hostettler, N. Lev-Cohain, M. Drozdzal, M.M. Amitai, R. Vivanti, J. Sosna, I. Ezhov, A. Sekuboyina, F. Navarro, F. Kofler, J.C. Paetzold, S. Shit, X. Hu, J. Lipkov, M. Rempfler, M. Piraud, J. Kirschke, B. Wiestler, Z. Zhang, C. Hlsemeyer, M. Beetz, F. Ettlinger, M. Antonelli, W. Bae, M. Bellver, L. Bi, H. Chen, G. Chlebus, E.B. Dam, Q. Dou, C.W. Fu, B. Georgescu, X.G. Nieto, F. Gruen, X. Han, P.A. Heng, J. Hesser, J.H. Moltz, C. Igel, F. Isensee, P. Jger, F. Jia, K.C. Kaluva, M. Khened, I. Kim, J.H. Kim, S. Kim, S. Kohl, T. Konopczynski, A. Kori, G. Krishnamurthi, F. Li, H. Li, J. Li, X. Li, J. Lowengrub, J. Ma, K. Maier-Hein, K.K. Maninis, H. Meine, D. Merhof, A. Pai, M. Perslev, J. Petersen, J. Pont-Tuset, J. Qi, X. Qi, O. Rippel, K. Roth, I. Sarasua, A. Schenk, Z. Shen, J. Torres, C. Wachinger, C. Wang, L. Weninger, J. Wu, D. Xu, X. Yang, S.C.H. Yu, Y. Yuan, M. Yue, L. Zhang, J. Cardoso, S. Bakas, R. Braren, V. Heinemann, C. Pal, A. Tang, S. Kadoury, L. Soler, B. van Ginneken, H. Greenspan, L. Joskowicz, B. Menze, "The Liver Tumor Segmentation Benchmark (LiTS)". *Medical Image Analysis*, 84:102680, 2023.

Preprints

G.E. Humpire Mamani, N. Lessmann, E. Scholten, M. Prokop, C. Jacobs and B. van Ginneken "Kidney abnormality segmentation in thorax-abdomen CT scans". *arXiv:2309.03383*, 2023

G.E. Humpire Mamani, C. Jacobs, M. Prokop, B. van Ginneken and N. Lessmann, "Transfer learning from a sparsely annotated dataset of 3D medical images". *arXiv:2311.05032*, 2023

Papers in conference proceedings

G.E. Humpire Mamani, A.A.A. Setio, B. van Ginneken and C. Jacobs. "Organ detection in thorax abdomen CT using multi-label convolutional neural networks", In: *Medical Imaging*, volume 10134 of Proceedings of the SPIE, 2017.

Abstracts in conference proceedings

G. Chlebus, **G.E. Humpire Mamani**, A. Schenk, B. van Ginneken and H. Meine, "Mimicking radiologists to improve the robustness of deep-learning based automatic liver segmentation", *Annual Meeting of the Radiological Society of North America*, 2019.

J. Bukala, **G.E. Humpire Mamani**, E. Scholten, M. Prokop, B. van Ginneken and C. Jacobs, "Fully Automatic Measurement of the Splenic Volume in CT with U-Net Convolutional Neural Networks", *Annual Meeting of the Radiological Society of North America*, 2017.

Datasets

G.E. Humpire Mamani, L. Builtjes, C. Jacobs, B. van Ginneken, M. Prokop and E. Scholten. "Dataset for: Kidney abnormality segmentation in thorax-abdomen CT scans". Zenodo 2023. `https://zenodo.org/records/8014290`

Bibliography

[1] Dattani S., Spooner F., Ritchie H., and Roser M. Causes of death. *Our World in Data*, 2023.

[2] RIVM. Medische stralingstoepassingen: Trends en stand van zaken, 2023. URL `https://www.rivm.nl/medische-stralingstoepassingen/trends-en-stand-van-zaken/diagnostiek`.

[3] Siegel R. L., Miller K. D., and Jemal A. Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(1):7–34, 2019.

[4] Dhar R., Seethy A., Singh S., Pethusamy K., Srivastava T., Talukdar J., Rath G. K., and Karmakar S. Cancer immunotherapy: Recent advances and challenges. *Journal of Cancer Research and Therapeutics*, 17(4):834–844, 2021.

[5] Röntgen W. C. Über eine neue Art von Strahlen. *Sitzungsberichte der Physikalisch-Medicinisch Gesellschaft zu Würzburg*, pages 132–141, 1895.

[6] LeCun Y., Bengio Y., and Hinton G. Deep learning. *Nature*, 521(7553):436–444, 2015.

[7] Litjens G., Kooi T., Ehteshami Bejnordi B., Setio A. A. A., Ciompi F., Ghafoorian M., van der Laak J., van Ginneken B., and Sánchez C. I. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[8] Ronneberger O., Fischer P., and Brox T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241, 2015.

[9] Çiçek Ö., Abdulkadir A., Lienkamp S. S., Brox T., and Ronneberger O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9901 of *Lecture Notes in Computer Science*, pages 424–432, 2016.

[10] Kaspar D. Application of directional antennas in rf-based indoor localization systems. Master's thesis, Swiss Federal Institute of Technology Zurich, 2005.

[11] Cuingnet R., Prevost R., Lesage D., Cohen L. D., Mory B., and Ardon R. Automatic detection and segmentation of kidneys in 3D CT images using random forests. In *Medical Image Computing and Computer-Assisted Intervention*, pages 66–74, 2012.

[12] Zhou X., Ito T., Zhou X., Chen H., Hara T., Yokoyama R., Kanematsu M., Hoshi H., and Fujita H. A universal approach for automatic organ segmentations on 3D CT images based on organ localization and 3D Grabcut. In *Medical Imaging*, volume 9035 of *Proceedings of the SPIE*, page 90352V, 2014.

[13] Zhan Y., Zhou X. S., Peng Z., and Krishnan A. Active scheduling of organ detection and segmentation in whole-body medical images. In *MICCAI*, volume 5241 of *Lecture Notes in Computer Science*, pages 313–321. Springer, 2008.

[14] Zheng Y., Georgescu B., Barbu A., Scheuering M., and Comaniciu D. Four-chamber heart modeling and automatic segmentation for 3D cardiac CT volumes. In *Medical Imaging*, volume 6914 of *Proceedings of the SPIE*, pages 691416–691416, 2008.

[15] Zhou X., Wang S., Chen H., Hara T., Yokoyama R., Kanematsu M., and Fujita H. Automatic localization of solid organs on 3D CT images by a collaborative majority voting decision based on ensemble learning. *Computerized Medical Imaging and Graphics*, 36(4):304–313, 2012.

[16] Zhou X., Morita S., Zhou X., Chen H., Hara T., Yokoyama R., Kanematsu M., Hoshi H., and

Fujita H. Automatic anatomy partitioning of the torso region on CT images by using multiple organ localizations with a group-wise calibration technique. In *Medical Imaging*, volume 9414 of *Proceedings of the SPIE*, page 94143K, 2015.

[17] Criminisi A., Shotton J., and Bucciarelli S. Decision forests with long-range spatial context for organ localization in CT volumes. In *MICCAI workshop on Probabilistic Models for Medical Image Analysis (MICCAI-PMMIA)*, January 2009.

[18] Criminisi A., Shotton J., Robertson D., and Konukoglu E. Regression forests for efficient anatomy detection and localization in CT studies. In *Medical Image Computing and Computer-Assisted Intervention*, volume 6533 of *Lecture Notes in Computer Science*, pages 106–117, 2010.

[19] Criminisi A., Robertson D., Konukoglu E., Shotton J., Pathak S., White S., and Siddiqui K. Regression forests for efficient anatomy detection and localization in computed tomography scans. *Medical Image Analysis*, pages 1293–1303, 2013.

[20] Pauly O., Glocker B., Criminisi A., Mateus D., Martinez-Mller A., Nekolla S. G., and Navab N. Fast multiple organ detection and localization in whole-body mr dixon sequences. In *MICCAI (3)*, volume 6893 of *Lecture Notes in Computer Science*, pages 239–247. Springer, 2011.

[21] Donner R., Menze B. H., Bischof H., and Langs G. Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization. *Medical Image Analysis*, 17(8): 1304–1314, 2013.

[22] Gauriau R., Cuingnet R., Lesage D., and Bloch I. Multi-organ localization with cascaded global-to-local regression and shape prior. *Medical Image Analysis*, 23(1):70–83, 2015.

[23] Samarakoon P., Promayon E., and Fouard C. Light random regression forests for automatic, multi-organ localization in CT images. In *IEEE International Symposium on Biomedical Imaging*, 2017.

[24] Zhou X., Watanabe A., Zhou X., Hara T., Yokoyama R., Kanematsu M., and Fujita H. Automatic organ segmentation on torso CT images by using content-based image retrieval. In *Medical Imaging*, volume 8314 of *Proceedings of the SPIE*, 2012.

[25] Gal V., Kerre E., and Tikk D. Organ detection in medical images with discriminately trained deformable part model. In *International conference on Computational Cybernetics*, pages 153–157, 2013.

[26] Shin H.-C., Orton M. R., Collins D. J., Doran S. J., and Leach M. O. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1930–1943, 2013.

[27] Roth H. R., Lee C. T., Shin H.-C., Seff A., Kim L., Yao J., Lu L., and Summers R. M. Anatomy-specific classification of medical images using deep convolutional nets. In *IEEE International Symposium on Biomedical Imaging*, pages 101–104, 2015.

[28] Humpire Mamani G. E., Setio A. A. A., van Ginneken B., and Jacobs C. Organ detection in thorax abdomen CT using multi-label convolutional neural networks. In *Medical Imaging*, volume 10134 of *Proceedings of the SPIE*, 2017.

[29] de Vos B., Wolterink J., de Jong P., Leiner T., Viergever M., and Isgum I. Convnet-based localization of anatomical structures in 3D medical images. *IEEE Transactions on Medical Imaging*, 36 (7):1470–1481, 2017.

*

[30] Hussain M. A., Amir-Khalili A., Hamarneh G., and Abugharbieh R. Segmentation-free kidney localization and volume estimation using aggregated orthogonal decision CNNs. In *Medical Image Computing and Computer-Assisted Intervention*, volume 10435 of *Lecture Notes in Computer Science*, pages 612–620, 2017.

[31] Yan Z., Zhan Y., Peng Z., Liao S., Shinagawa Y., Metaxas D. N., and Zhou X. S. Bodypart recognition using multi-stage deep learning. In *International Conference on Information Processing in Medical Imaging*, pages 449–461. Springer, 2015.

[32] de Vos B. D., Wolterink J. M., de Jong P. A., Viergever M. A., and Išgum I. 2D image classification for 3D anatomy localization: employing deep convolutional neural networks. In *Medical Imaging*, Proceedings of the SPIE, pages 97841Y–97841Y, 2016.

[33] Krizhevsky A., Sutskever I., and Hinton G. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.

[34] van Rikxoort E. M., de Hoop B., Viergever M. A., Prokop M., and van Ginneken B. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Medical Physics*, 36(7):2934–2947, 2009.

[35] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., and Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1): 1929–1958, 2014.

[36] Clare A. and King R. Knowledge Discovery in Multi-label Phenotype Data. In *Principles of Data Mining and Knowledge Discovery*, volume 2168, pages 42–53, 2001.

[37] Tieleman T. and Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4, 2012.

[38] van Grinsven M. J. J. P., van Ginneken B., Hoyng C. B., Theelen T., and Sánchez. C. I. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE Transactions on Medical Imaging*, 35(5):1273–1284, 2016.

[39] Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., and Rabinovich A. Going deeper with convolutions. *arXiv:14094842v1*, 2014.

[40] Milletari F., Navab N., and Ahmadi S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *arXiv:1606.04797*, 2016.

[41] He K., Gkioxari G., Dollár P., and Girshick R. B. Mask R-CNN. In *International Conference on Computer Vision*, pages 2961–2969, Oct 2017.

[42] Bergstra J., Breuleux O., Bastien F., Lamblin P., Pascanu R., Desjardins G., Turian J., Warde-Farley D., and Bengio Y. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, volume 4, page 3, 2010.

[43] Bastien F., Lamblin P., Pascanu R., Bergstra J., Goodfellow I., Bergeron A., Bouchard N., Warde-Farley D., and Bengio Y. Theano: new features and speed improvements. *Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop*, 2012.

[44] Harris A., Kamishima T., Hao H. Y., Kato F., Omatsu T., Onodera Y., Terae S., and Shirato H. Splenic volume measurements on computed tomography utilizing automatically contouring

software and its relationship with age, gender, and anthropometric parameters. *European Journal of Radiology*, 75(1):97–101, 2010.

[45] Robertson F., Leander P., and Ekberg O. Radiology of the spleen. *European Radiology*, 11(1): 80–95, jan 2001.

[46] Bergman R. A., Heidger P. M., and Scott-Conner C. E. H. The anatomy of the spleen. In *The Complete Spleen: Structure, Function, and Clinical Disorders*, pages 3–9. Humana Press, Totowa, NJ, 2002.

[47] Moroz P., Anderson J. E., Van Hazel G., and Gray B. N. Effect of selective internal radiation therapy and hepatic arterial chemotherapy on normal liver volume and spleen volume. *Journal of Surgical Oncology*, 78(4):248–252, 2001.

[48] Jacobs K., Visser B., and Gayer G. Changes in spleen volume after resection of hepatic colorectal metastases. *Clinical Radiology*, 67(10):982–987, 2012.

[49] Odorico I. D., Spaulding K. A., Pretorius D. H., Lev-Toaff A. S., Bailey T. B., and Nelson T. R. Normal splenic volumes estimated using three-dimensional ultrasonography. *Journal of Ultrasound in Medicine*, 18(3):231–236, mar 1999.

[50] Joiner B. J., Simpson A. L., Leal J. N., D'Angelica M. I., and Do R. K. G. Assessing splenic enlargement on CT by unidimensional measurement changes in patients with colorectal liver metastases. *Abdominal Imaging*, 40(7):2338–2344, jun 2015.

[51] Cruz-Romero C., Agarwal S., Abujudeh H. H., Thrall J., and Hahn P. F. Spleen volume on CT and the effect of abdominal trauma. *Emergency Radiology*, 23(4):315–323, 2016.

[52] Simpson A. L., Leal J. N., Pugalenthi A., Allen P. J., DeMatteo R. P., Fong Y., Gonen M., Jarnagin W. R., Kingham T. P., Miga M. I., Shia J., Weiser M. R., and D'Angelica M. I. Chemotherapy-induced splenic volume increase is independently associated with major complications after hepatic resection for metastatic colorectal cancer. *Journal of the American College of Surgeons*, 220 (3):271–280, mar 2015.

[53] Prassopoulos P., Daskalogiannaki M., Raissaki M., Hatjidakis A., and Gourtsoyiannis N. Determination of normal splenic volume on computed tomography in relation to age, gender and body habitus. *European Radiology*, 7(2):246–248, 1997.

[54] Yetter E. M., Acosta K. B., Olson M. C., and Blundell K. Estimating splenic volume: sonographic measurements correlated with helical CT determination. *American Journal of Roentgenology*, 181 (6):1615–1620, 2003.

[55] Linguraru M. G., Sandberg J. K., Li Z., Shah F., and Summers R. M. Automated segmentation and quantification of liver and spleen from CT images using normalized probabilistic atlases and enhancement estimation. *Medical Physics*, 37(2):771–783, 2010.

[56] Tong T., Wolz R., Wang Z., Gao Q., Misawa K., Fujiwara M., Mori K., Hajnal J. V., and Rueckert D. Discriminative dictionary learning for abdominal multi-organ segmentation. *Medical Image Analysis*, 23(1):92–104, 2015.

[57] Wolz R., Chu C., Misawa K., Fujiwara M., Mori K., and Rueckert D. Automated abdominal multi-organ segmentation with subject-specific atlas generation. *IEEE Transactions on Medical Imaging*, 32(9):1723–1730, 2013.

*

[58] Okada T., Linguraru M. G., Hori M., Summers R. M., Tomiyama N., and Sato Y. Abdominal multi-organ segmentation from CT images using conditional shape–location and unsupervised intensity priors. *Medical Image Analysis*, 26(1):1–18, 2015.

[59] Hammon M., Dankerl P., Kramer M., Seifert S., Tsymbal A., Costa M., Janka R., Uder M., and Cavallaro A. Automated detection and volumetric segmentation of the spleen in CT scans. *RoFo: Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin*, 184(8):734–739, 2012.

[60] Wood A., Soroushmehr S. M. R., Farzaneh N., Fessell D., Ward K. R., Gryak J., Kahrobaei D., and Na K. Fully automated spleen localization and segmentation using machine learning and 3D active contours. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 53–56, jul 2018.

[61] Gloger O., Tönnies K., Bülow R., and Völzke H. Automatized spleen segmentation in non-contrast-enhanced MR volume data using subject-specific shape priors. *Physics in Medicine and Biology*, 62(14):5861–5883, 2017.

[62] Gauriau R., Ardori R., Lesage D., and Bloch I. Multiple template deformation application to abdominal organ segmentation. In *IEEE International Symposium on Biomedical Imaging*, pages 359–362, 2015.

[63] Huo Y., Xu Z., Bao S., Bermudez C., Plassard A. J., Yao Y., Liu J., Assad A., Abramson R. G., and Landman B. A. Splenomegaly segmentation using global convolutional kernels and conditional generative adversarial networks. In *Medical Imaging*, volume 10574 of *Proceedings of the SPIE*, page 1057409, mar 2018.

[64] Zhou X., Ito T., Takayama R., Wang S., Hara T., and Fujita H. Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting. In *Deep Learning and Data Labeling for Medical Applications*, pages 111–120. Springer International Publishing, Cham, 2016.

[65] Roth H. R., Oda H., Zhou X., Shimizu N., Yang Y., Hayashi Y., Oda M., Fujiwara M., Misawa K., and Mori K. An application of cascaded 3D fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99, jun 2018.

[66] Gibson E., Giganti F., Hu Y., Bonmati E., Bandula S., Gurusamy K., Davidson B., Pereira S. P., Clarkson M. J., and Barratt D. C. Automatic multi-organ segmentation on abdominal CT with dense V-Networks. *IEEE Transactions on Medical Imaging*, 37(8):1822–1834, aug 2018.

[67] Landman B. A., Bobo M. F., Huo Y., Yao Y., Bao S., Virostko J., Plassard A. J., Lyu I., Assad A., Abramson R. G., and Hilmes M. A. Fully convolutional neural networks improve abdominal organ segmentation. In *Medical Imaging*, volume 10574 of *Proceedings of the SPIE*, page 105742V, mar 2018.

[68] Chlebus G., Schenk A., Moltz J. H., van Ginneken B., Hahn H. K., and Meine H. Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. *Nature Scientific Reports*, 8:15497, October 2018.

[69] Dalmis M. U., Litjens G., Holland K., Setio A., Mann R., Karssemeijer N., and Gubern-Mérida A. Using deep learning to segment breast and fibroglandular tissue in MRI volumes. *Medical Physics*, 44:533–546, February 2017.

[70] Aresta G., Araújo T., Jacobs C., van Ginneken B., Cunha A., Ramos I., and Campilho A. Towards an automatic lung cancer screening system in low dose computed tomography. In *MICCAI Workshop: Thoracic Image Analysis*, volume 11040 of *Lecture Notes in Computer Science*, 2018.

[71] Ehteshami Bejnordi B., Veta M., van Diest P. J., van Ginneken B., Karssemeijer N., Litjens G., van der Laak J. A. W. M., the CAMELYON16 Consortium, Hermsen M., Manson Q. F., Balkenhol M., Geessink O., Stathonikos N., van Dijk M. C., Bult P., Beca F., Beck A. H., Wang D., Khosla A., Gargeya R., Irshad H., Zhong A., Dou Q., Li Q., Chen H., Lin H.-J., Heng P.-A., Haß C., Bruni E., Wong Q., Halici U., Öner M. U., Cetin-Atalay R., Berseth M., Khvatkov V., Vylegzhanin A., Kraus O., Shaban M., Rajpoot N., Awan R., Sirinukunwattana K., Qaiser T., Tsang Y.-W., Tellez D., Annuscheit J., Hufnagl P., Valkonen M., Kartasalo K., Latonen L., Ruusuvuori P., Liimatainen K., Albarqouni S., Mungal B., George A., Demirci S., Navab N., Watanabe S., Seno S., Takenaka Y., Matsuda H., Ahmady Phoulady H., Kovalev V., Kalinovsky A., Liauchuk V., Bueno G., Fernandez-Carrobles M. M., Serrano I., Deniz O., Racoceanu D., and Venâncio R. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Journal of the American Medical Association*, 318:2199–2210, December 2017.

[72] Setio A. A. A., Traverso A., de Bel T., Berens M. S. N., Bogaard C. v. d., Cerello P., Chen H., Dou Q., Fantacci M. E., Geurts B., Gugten R. v. d., Heng P. A., Jansen B., de Kaste M. M. J., Kotov V., Lin J. Y.-H., Manders J. T. M. C., Sonora-Mengana A., Garcia-Naranjo J. C., Papavasileiou E., Prokop M., Saletta M., Schaefer-Prokop C. M., Scholten E. T., Scholten L., Snoeren M. M., Torres E. L., Vandemeulebroucke J., Walasek N., Zuidhof G. C. A., Ginneken B. v., and Jacobs C. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Medical Image Analysis*, 42: 1–13, 2017.

[73] Zhu W., Huang Y., Zeng L., Chen X., Liu Y., Qian Z., Du N., Fan W., and Xie X. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical Physics*, 46(2):576–589, 2019.

[74] Pattanayak P., Turkbey E. B., and Summers R. M. Comparative evaluation of three software packages for liver and spleen segmentation and volumetry. *Academic Radiology*, 24(7):831–839, 2017.

[75] Humpire Mamani G. E., Setio A., van Ginneken B., and Jacobs C. Efficient organ localization using multi-label convolutional neural networks in thorax-abdomen CT scans. *Physics in Medicine and Biology*, 63(8):085003, 2018.

[76] Isensee F., Jaeger P. F., Kohl S. A. A., Petersen J., and Maier-Hein K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 2020.

[77] Balagopal A., Kazemifar S., Nguyen D., Lin M.-H., Hannan R., Owrangi A., and Jiang S. Fully automated organ segmentation in male pelvic CT images. *Physics in Medicine and Biology*, 63 (24):245015, 2018.

[78] Guo Z., Zhang L., Lu L., Bagheri M., Summers R. M., Sonka M., and Yao J. Deep LOGISMOS: deep learning graph-based 3D segmentation of pancreatic tumors on CT scans. In *IEEE International Symposium on Biomedical Imaging*, pages 1230–1233. IEEE, 2018.

[79] Kutikov A. and Uzzo R. G. The RENAL nephrometry score: a comprehensive standardized system for quantitating renal tumor size, location and depth. *Journal of Urology*, 182(3):844–853, 2009.

[80] Lin D.-T., Lei C.-C., and Hung S.-W. Computer-aided kidney segmentation on abdominal CT images. *IEEE Transactions on Information Technology in Biomedicine*, 10(1):59–65, 2006.

[81] Kim D.-Y. and Park J.-W. Computer-aided detection of kidney tumor on abdominal computed tomography scans. *Acta Radiologica*, 45(7):791–795, 2004.

[82] Lu J., Chen J., Zhang J., and Yang W. Segmentation of kidney using CV model and anatomy priors. In *Medical Imaging, Parallel Processing of Images, and Optimization Techniques*, volume 6789 of *Proceedings of the SPIE*, page 678911, 2007.

[83] Linguraru M. G., Yao J., Gautam R., Peterson J., Li Z., Linehan W. M., and Summers R. M. Renal tumor quantification and classification in contrast-enhanced abdominal CT. *Pattern Recognition*, 42(6):1149–1161, 2009.

[84] Skalski A., Heryan K., Jakubowski J., and Drewniak T. Kidney segmentation in CT data using hybrid Level-Set method with ellipsoidal shape constraints. *Metrology and Measurement Systems*, 24(1):101–112, 2017.

[85] Ali A. M., Farag A. A., and El-Baz A. S. Graph cuts framework for kidney segmentation with prior shape constraints. In *Medical Image Computing and Computer-Assisted Intervention*, pages 384–392. Springer, 2007.

[86] Yoruk U., Hargreaves B. A., and Vasanawala S. S. Automatic renal segmentation for MR urography using 3D-GrabCut and random forests. *Magnetic Resonance in Medicine*, 79(3):1696–1707, 2018.

[87] Turco D., Valinoti M., Martin E. M., Tagliaferri C., Scolari F., and Corsi C. Fully automated segmentation of polycystic kidneys from noncontrast computed tomography: A feasibility study and preliminary results. *Academic Radiology*, 25(7):850–855, 2018.

[88] Badura P. and Wieclawek W. Calibrating level set approach by granular computing in computed tomography abdominal organs segmentation. *Applied Soft Computing*, 49:887–900, 2016.

[89] Farmaki C., Marias K., Sakkalis V., and Graf N. Spatially adaptive active contours: a semi-automatic tumor segmentation framework. *International Journal of Computer Assisted Radiology and Surgery*, 5(4):369–384, 2010.

[90] Khalifa F., Soliman A., Elmaghraby A., Gimelfarb G., and El-Baz A. 3D kidney segmentation from abdominal images using spatial-appearance models. *Computational and mathematical methods in medicine*, 2017, 2017.

[91] Wieclawek W. 3D marker-controlled watershed for kidney segmentation in clinical CT exams. *Biomedical Engineering Online*, 17(1):26, 2018.

[92] Chen X., Summers R., and Yao J. FEM based 3D tumor growth prediction for kidney tumor. In *International Workshop on Medical Imaging and Virtual Reality*, pages 159–168. Springer, 2010.

[93] Kaur R., Juneja M., and Mandal A. A hybrid edge-based technique for segmentation of renal lesions in CT images. *Multimedia Tools and Applications*, 78(10):12917–12937, 2019.

[94] Zheng Y., Liu D., Georgescu B., Xu D., and Comaniciu D. Deep learning based automatic

segmentation of pathological kidney in CT: Local versus global image context. In *Advances in Computer Vision and Pattern Recognition*, pages 241–255. Springer, Cham, 2017.

[95] Sharma K., Rupprecht C., Caroli A., Aparicio M. C., Remuzzi A., Baust M., and Navab N. Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease. *Nature Scientific Reports*, 7(1):2049, 2017.

[96] Bilic P., Christ P., Li H. B., Vorontsov E., Ben-Cohen A., Kaissis G., Szeskin A., Jacobs C., Humpire-Mamani G. E., Chartrand G., Lohfer F., Holch J. W., Sommer W., Hofmann F., Hostettler A., Lev-Cohain N., Drozdzal M., Amitai M. M., Vivanti R., Sosna J., Ezhov I., Sekuboyina A., Navarro F., Kofler F., Paetzold J. C., Shit S., Hu X., Lipkov J., Rempfler M., Piraud M., Kirschke J., Wiestler B., Zhang Z., Hlsemeyer C., Beetz M., Ettlinger F., Antonelli M., Bae W., Bellver M., Bi L., Chen H., Chlebus G., Dam E. B., Dou Q., Fu C.-W., Georgescu B., i Nieto X. G., Gruen F., Han X., Heng P.-A., Hesser J., Moltz J. H., Igel C., Isensee F., Jger P., Jia F., Kaluva K. C., Khened M., Kim I., Kim J.-H., Kim S., Kohl S., Konopczynski T., Kori A., Krishnamurthi G., Li F., Li H., Li J., Li X., Lowengrub J., Ma J., Maier-Hein K., Maninis K.-K., Meine H., Merhof D., Pai A., Perslev M., Petersen J., Pont-Tuset J., Qi J., Qi X., Rippel O., Roth K., Sarasua I., Schenk A., Shen Z., Torres J., Wachinger C., Wang C., Weninger L., Wu J., Xu D., Yang X., Yu S. C.-H., Yuan Y., Yue M., Zhang L., Cardoso J., Bakas S., Braren R., Heinemann V., Pal C., Tang A., Kadoury S., Soler L., van Ginneken B., Greenspan H., Joskowicz L., and Menze B. The Liver Tumor Segmentation Benchmark (LiTS). *Medical Image Analysis*, 84:102680, 2023.

[97] Taha A., Lo P., Li J., and Zhao T. Kid-Net: convolution networks for kidney vessels segmentation from CT-volumes. In *Medical Image Computing and Computer-Assisted Intervention*, pages 463–471. Springer, 2018.

[98] Jackson P., Hardcastle N., Dawe N., Kron T., Hofman M., and Hicks R. J. Deep learning renal segmentation for fully automated radiation dose estimation in unsealed source therapy. *Frontiers in oncology*, 8:215, 2018.

[99] Yu Q., Shi Y., Sun J., Gao Y., Zhu J., and Dai Y. Crossbar-Net: A novel convolutional neural network for kidney tumor segmentation in CT images. *IEEE Transactions on Image Processing*, 28(8):4060–4074, 2019.

[100] Yang G., Li G., Pan T., Kong Y., Wu J., Shu H., Luo L., Dillenseger J., Coatrieux J., Tang L., and Zhu X. Automatic segmentation of kidney and renal tumor in CT images based on 3D fully convolutional neural network with pyramid pooling module. In *International Conference on Pattern Recognition*, pages 3790–3795, August 2018.

[101] Han X. Automatic liver lesion segmentation using a deep convolutional neural network method. *arXiv:1704.07239*, 2017.

[102] Blau N., Klang E., Kiryati N., Amitai M., Portnoy O., and Mayer A. Fully automatic detection of renal cysts in abdominal CT scans. *International Journal of Computer Assisted Radiology and Surgery*, 13(7):957–966, 2018.

[103] Haghighi M., Warfield S. K., and Kurugol S. Automatic renal segmentation in DCE-MRI using convolutional neural networks. In *IEEE International Symposium on Biomedical Imaging*, pages 1534–1537, April 2018.

[104] Heller N., Isensee F., Maier-Hein K. H., Hou X., Xie C., Li F., Nan Y., Mu G., Lin Z., Han M., Yao G., Gao Y., Zhang Y., Wang Y., Hou F., Yang J., Xiong G., Tian J., Zhong C., Ma J., Rickman J.,

Dean J., Stai B., Tejpaul R., Oestreich M., Blake P., Kaluzniak H., Raza S., Rosenberg J., Moore K., Walczak E., Rengel Z., Edgerton Z., Vasdev R., Peterson M., McSweeney S., Peterson S., Kalapara A., Sathianathen N., Papanikolopoulos N., and Weight C. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Medical Image Analysis*, 67:101821, 2021.

[105] Heckel F., Schwier M., and Peitgen H.-O. Object-oriented application development with MeVisLab and Python. *Lecture Notes in Informatics (Informatik 2009: Im Focus das Leben)*, 154:1338–1351, 2009.

[106] Tompson J., Goroshin R., Jain A., LeCun Y., and Bregler C. Efficient object localization using convolutional networks. In *Computer Vision and Pattern Recognition*, pages 648–656, 2015.

[107] Berrada L., Zisserman A., and Kumar M. P. Smooth loss functions for deep top-k classification. *arXiv preprint arXiv:1802.07595*, 2018.

[108] Glorot X. and Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256, 2010.

[109] Kingma D. and Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[110] Gibson E., Li W., Sudre C., Fidon L., Shakir D. I., Wang G., Eaton-Rosen Z., Gray R., Doel T., Hu Y., et al. NiftyNet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 158:113–122, 2018.

[111] Heinrich M. P., Oktay O., and Bouteldja N. OBELISK-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions. *Medical Image Analysis*, 54:1–9, May 2019.

[112] Wang Y., Zhou Y., Tang P., Shen W., Fishman E. K., and Yuille A. L. Training multi-organ segmentation networks with sample selection by relaxed upper confident bound. In *Medical Image Computing and Computer-Assisted Intervention*, pages 434–442. Springer, 2018.

[113] Shin H.-C., Roth H. R., Gao M., Lu L., Xu Z., Nogues I., Yao J., Mollura D., and Summers R. M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[114] Deng J., Dong W., Socher R., Li L., Li K., and Fei-Fei L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[115] Raghu M., Zhang C., Kleinberg J., and Bengio S. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, pages 3342–3352, 2019.

[116] Jimenez-del Toro O., Müller H., Krenn M., Gruenberg K., Taha A. A., Winterstein M., Eggel I., Foncubierta-Rodríguez A., Goksel O., Jakab A., et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks. *IEEE Transactions on Medical Imaging*, 35(11):2459–2475, 2016.

[117] Gibson E., Giganti F., Hu Y., Bonmati E., Bandula S., Gurusamy K., Davidson B., Pereira S. P., Clarkson M. J., and Barratt D. C. Multi-organ abdominal CT reference standard segmentations, Feb 2018. URL `https://doi.org/10.5281/zenodo.1169361`.

[118] Roth H. R., Lu L., Farag A., Shin H.-C., Liu J., Turkbey E. B., and Summers R. M. DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 556–564. Springer, 2015.

[119] Landman B., Xu Z., Igelsias J., Styner M., Langerak T., and Klein A. Multi-atlas labeling beyond the cranial vault-workshop and challenge, 2017. URL `https://doi.org/10.7303/syn3193805`.

[120] Shie C., Chuang C., Chou C., Wu M., and Chang E. Y. Transfer representation learning for medical image analysis. In *Conference Proceedings of the IEEE Engineering in Medicine and Biology Society*, pages 711–714, 2015.

[121] Rajpurkar P., Park A., Irvin J., Chute C., Bereket M., Mastrodicasa D., Langlotz C. P., Lungren M. P., Ng A. Y., and Patel B. N. AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. *Nature Scientific Reports*, 10(1):1–7, 2020.

[122] Conze P.-H., Brochard S., Burdin V., Sheehan F. T., and Pons C. Healthy versus pathological learning transferability in shoulder muscle MRI segmentation using deep convolutional encoder-decoders. *Computerized Medical Imaging and Graphics*, 83:101733, 2020.

[123] Yang J., Huang X., He Y., Xu J., Yang C., Xu G., and Ni B. Reinventing 2D convolutions for 3D images. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3009–3018, 2021.

[124] van Opbroek A., Ikram M. A., Vernooij M. W., and De Bruijne M. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Transactions on Medical Imaging*, 34:1018–1030, 2015.

[125] Carneiro G., Nascimento J., and Bradley A. P. Unregistered multiview mammogram analysis with pre-trained deep learning models. In *Medical Image Computing and Computer-Assisted Intervention*, pages 652–660. Springer, 2015.

[126] Ravishankar H., Sudhakar P., Venkataramani R., Thiruvenkadam S., Annangi P., Babu N., and Vaidya V. Understanding the mechanisms of deep transfer learning for medical images. In *Deep Learning and Data Labeling for Medical Applications*, pages 188–196. Springer, 2016.

[127] Cheplygina V. Cats or CAT scans: Transfer learning from natural or medical image source data sets? *Current Opinion in Biomedical Engineering*, 9:21–27, 2019.

[128] Zhou Z., Sodha V., Pang J., Gotway M. B., and Liang J. Models genesis. *Medical Image Analysis*, 67:101840, 2021.

[129] Chen S., Ma K., and Zheng Y. Med3D: Transfer learning for 3D medical image analysis. *arxiv:1904.00625v4*, 2019.

[130] Ji Y., Bai H., GE C., Yang J., Zhu Y., Zhang R., Li Z., Zhanng L., Ma W., Wan X., and Luo P. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Advances in Neural Information Processing Systems*, volume 35, pages 36722–36732, 2022.

[131] Antonelli M., Reinke A., Bakas S., Farahani K., Kopp-Schneider A., Landman B. A., Litjens G., Menze B., Ronneberger O., Summers R. M., van Ginneken B., Bilello M., Bilic P., Christ P. F., Do R. K. G., Gollub M. J., Heckers S. H., Huisman H., Jarnagin W. R., McHugo M. K., Napel S., Pernicka J. S. G., Rhode K., Tobon-Gomez C., Vorontsov E., Meakin J. A., Ourselin S., Wiesenfarth M., Arbeláez P., Bae B., Chen S., Daza L., Feng J., He B., Isensee F., Ji Y., Jia F., Kim I., Maier-Hein K., Merhof D., Pai A., Park B., Perslev M., Rezaiifar R., Rippel O., Sarasua I.,

Shen W., Son J., Wachinger C., Wang L., Wang Y., Xia Y., Xu D., Xu Z., Zheng Y., Simpson A. L., Maier-Hein L., and Cardoso M. J. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[132] Dong N., Kampffmeyer M., Liang X., Xu M., Voiculescu I., and Xing E. P. Towards robust medical image segmentation on small-scale data with incomplete labels. *arxiv:2011.14164*, 2020.

[133] Guo F., Ng M., Kuling G., and Wright G. Cardiac MRI segmentation with sparse annotations: Ensembling deep learning uncertainty and shape priors. *Medical Image Analysis*, 81:102532, 2022.

[134] Lian S., Li L., Luo Z., Zhong Z., Wang B., and Li S. Learning multi-organ segmentation via partial- and mutual-prior from single-organ datasets. *Biomedical Signal Processing and Control*, 80:104339, 2023.

[135] Petit O., Thome N., Charnoz A., Hostettler A., and Soler L. Handling missing annotations for semantic segmentation with deep convnets. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 20–28, 2018.

[136] Tajbakhsh N., Jeyaseelan L., Li Q., Chiang J. N., Wu Z., and Ding X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

[137] Wang G., Luo X., Gu R., Yang S., Qu Y., Zhai S., Zhao Q., Li K., and Zhang S. PyMIC: A deep learning toolkit for annotation-efficient medical image segmentation. *Computer Methods and Programs in Biomedicine*, 231:107398, 2023.

[138] Zhou Y., Li Z., Bai S., Chen X., Han M., Wang C., Fishman E. K., and Yuille A. L. Prior-aware neural network for partially-supervised multi-organ segmentation. In *International Conference on Computer Vision*, pages 10671–10680. IEEE, 2019.

[139] Jin D., Xu Z., Harrison A. P., George K., and Mollura D. J. 3D convolutional neural networks with graph refinement for airway segmentation using incomplete data labels. In *Machine Learning in Medical Imaging*, volume 10541, pages 141–149. Springer, 2017.

[140] Fang X. and Yan P. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020.

[141] Shi G., Xiao L., Chen Y., and Zhou S. K. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70:101979, 2021.

[142] Liu P., Xiao L., and Zhou S. K. Incremental learning for multi-organ segmentation with partially labeled datasets. *arXiv:2103.04526*, 2021.

[143] Zhang J., Xie Y., Xia Y., and Shen C. DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Computer Vision and Pattern Recognition*, pages 1195–1204, 2021.

[144] Yang J., Veeraraghavan H., Armato III S. G., Farahani K., Kirby J. S., Kalpathy-Kramer J., van Elmpt W., Dekker A., Han X., Feng X., et al. Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. *Medical Physics*, 45(10):4568–4581, 2018.

[145] Yao J., Burns J. E., Forsberg D., Seitel A., Rasoulian A., Abolmaesumi P., Hammernik K., Urschler M., Ibragimov B., Korez R., et al. A multi-center milestone study of clinical verte-

bral CT segmentation. *Computerized Medical Imaging and Graphics*, 49:16–28, 2016.

[146] Sekuboyina A., Husseini M. E., Bayat A., Loffler M., Liebl H., Li H., Tetteh G., Kukacka J., Payer C., Stern D., Urschler M., Chen M., Cheng D., Lessmann N., Hu Y., Wang T., Yang D., Xu D., Ambellan F., Amiranashvili T., Ehlke M., Lamecker H., Lehnert S., Lirio M., de Olaguer N. P., Ramm H., Sahu M., Tack A., Zachow S., Jiang T., Ma X., Angerman C., Wang X., Brown K., Kirszenberg A., Puybareau E., Chen D., Bai Y., Rapazzo B. H., Yeah T., Zhang A., Xu S., Hou F., He Z., Zeng C., Xiangshang Z., Liming X., Netherton T. J., Mumme R. P., Court L. E., Huang Z., He C., Wang L.-W., Ling S. H., Huynh L. D., Boutry N., Jakubicek R., Chmelik J., Mulay S., Sivaprakasam M., Paetzold J. C., Shit S., Ezhov I., Wiestler B., Glocker B., Valentinitsch A., Rempfler M., Menze B. H., and Kirschke J. S. VerSe: A Vertebrae Labelling and Segmentation Benchmark for multi-detector CT images. *Medical Image Analysis*, 73:102166, 2021.

[147] Löffler M. T., Sekuboyina A., Jacob A., Grau A.-L., Scharr A., El Husseini M., Kallweit M., Zimmer C., Baum T., and Kirschke J. S. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020.

[148] Regan E. A., Hokanson J. E., Murphy J. R., Make B., Lynch D. A., Beaty T. H., Curran-Everett D., Silverman E. K., and Crapo J. D. Genetic epidemiology of COPD (COPDGene) study design. *COPD*, 7:32–43, 2010.

[149] Litjens G., Toth R., van de Ven W., Hoeks C., Kerkstra S., van Ginneken B., Vincent G., Guillard G., Birbeck N., Zhang J., Strand R., Malmberg F., Ou Y., Davatzikos C., Kirschner M., Jung F., Yuan J., Qiu W., Gao Q., Edwards P. E., Maan B., van der Heijden F., Ghose S., Mitra J., Dowling J., Barratt D., Huisman H., and Madabhushi A. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis*, 18(2):359–373, 2014.

[150] Chen X., Weng J., Luo W., Lu W., Wu H., Xu J., and Tian Q. Sample balancing for deep learning-based visual recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10): 3962–3976, 2020.

[151] Xu X., Deng H. H., Gateno J., and Yan P. Federated multi-organ segmentation with inconsistent labels. *IEEE Transactions on Medical Imaging*, 42(10):2948–2960, 2023.

[152] Dong N., Kampffmeyer M., Voiculescu I., and Xing E. Federated partially supervised learning with limited decentralized medical images. *IEEE Transactions on Medical Imaging*, 42(7):1944–1954, 2023.

[153] Shen C., Wang P., Yang D., Xu D., Oda M., Chen P.-T., Liu K.-L., Liao W.-C., Fuh C.-S., Mori K., Wang W., and Roth H. R. Joint multi organ andtumor segmentation from partial labels using federated learning. In *Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health*, pages 58–67, Cham, 2022. Springer Nature Switzerland.

[154] Hochreiter S. and Schmidhuber J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[155] Zhao W. X., Zhou K., Li J., Tang T., Wang X., Hou Y., Min Y., Zhang B., Zhang J., Dong Z., Du Y., Yang C., Chen Y., Chen Z., Jiang J., Ren R., Li Y., Tang X., Liu Z., Liu P., Nie J.-Y., and Wen J.-R. A survey of large language models. *arXiv:2303.18223*, 2023.

[156] Wang Y., Yao Q., Kwok J. T., and Ni L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3), jun 2020.

*

[157] Chan S., Santoro A., Lampinen A., Wang J., Singh A., Richemond P., McClelland J., and Hill F. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pages 18878–18891, 2022.

[158] Gibson E., Hu Y., Huisman H., and Barratt D. Designing image segmentation studies: statistical power, sample size and reference standard quality. *Medical Image Analysis*, 42:44–59, 2017.

[159] Gibson E., Yipeng, Ghavami H. N., Ahmed H. U., Moore C., Emberton M., Huisman H., and Barratt D. Inter-site variability in prostate segmentation accuracy using deep learning. In *Medical Image Computing and Computer-Assisted Intervention*, volume 11073 of *Lecture Notes in Computer Science*, pages 506–514, 2018.

[160] Frid-Adar M., Diamant I., Klang E., Amitai M., Goldberger J., and Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[161] Chen Y., Yang X.-H., Wei Z., Heidari A. A., Zheng N., Li Z., Chen H., Hu H., Zhou Q., and Guan Q. Generative adversarial networks in medical image augmentation: A review. *Computers in Biology and Medicine*, 144:105382, 2022.

[162] Chlap P., Min H., Vandenberg N., Dowling J., Holloway L., and Haworth A. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021.

[163] Russ T., Goerttler S., Schnurr A.-K., Bauer D. F., Hatamikia S., Schad L. R., Zöllner F. G., and Chung K. Synthesis of CT images from digital body phantoms using CycleGAN. *International Journal of Computer Assisted Radiology and Surgery*, 14:1741–1750, 2019.

[164] Jiang P., Ergu D., Liu F., Cai Y., and Ma B. A review of Yolo algorithm developments. *Procedia Computer Science*, 199:1066–1073, 2022.

[165] Wu Y., Kirillov A., Massa F., Lo W.-Y., and Girshick R. Detectron2, 2019. URL `https://github.com/facebookresearch/detectron2`.

[166] Baumgartner M., Jäger P. F., Isensee F., and Maier-Hein K. H. nnDetection: A self-configuring method for medical object detection. In *Medical Image Computing and Computer-Assisted Intervention*, volume 12905 of *Lecture Notes in Computer Science*, pages 530–539, 2021.

[167] Huang Z., Wang H., Ye J., Niu J., Tu C., Yang Y., Du S., Deng Z., Gu L., and He J. Revisiting nnU-Net foriterative pseudo labeling andefficient sliding window inference. In *MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation*, volume 13816 of *Lecture Notes in Computer Science*, pages 178–189, 2022.

[168] Khan S., Naseer M., Hayat M., Zamir S. W., Khan F. S., and Shah M. Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s):1–41, sep 2022.

[169] van Leeuwen K. G., de Rooij M., Schalekamp S., van Ginneken B., and Rutten M. J. How does artificial intelligence in radiology improve efficiency and health outcomes? *Pediatric Radiology*, 52:1–7, October 2021.

Dankwoord

One day, I was in the middle of the carnival in warm Brazil, at midnight, wearing a T-shirt, and the very next day, I was wearing many layers of clothing in Nijmegen in the middle of winter, 35°C less than what I was used to but still, ready to join DIAG. This Ph.D. journey has undeniably proven to be the most arduous chapter of my life, yet with the support of many people, I am happy to have successfully concluded this exciting chapter. For that, I would like to express my gratitude and thank all the people I met during my years at DIAG.

To my promoters: I would like to thank **Bram van Ginneken**, whose guidance has pushed me forward from my initial day at DIAG. On that day, you gave me my first task: submitting a paper to MICCAI within 30 days. Since then, I have known this journey would be challenging. Your meticulous attention to detail was the main reason for my difficult meetings with you. A simple remark such as "you should know your data..." made me think a lot. Bram, I have learned invaluable lessons from our interactions and am sincerely grateful for your mentorship. **Mathias Prokop**, thanks for your invaluable clinical feedback during our "Pulmo" meetings. Upon learning of my struggles with writing papers, you promptly recommended "The Elements of Style" by William Strunk Jr., a resource that enhanced my writing skills. Your encouraging remarks, such as "I want to see this project working in the clinics," have kept me motivated throughout my research. Your insights have been fundamental to the completion of this thesis.

To my co-promoters: **Colin Jacobs**, you have been a fantastic co-promotor, always guiding me and explaining everything with (a lot of) patience; I appreciate that a lot. The "hola" at the start of our meetings showed that your Spanish was better than my Dutch. I have always been impressed by how you fixed the problems I had with MevisLab in a few minutes. Thanks for changing the name of the Chest CT team to the Body CT team to make me feel more like part of the team. **Nikolas Leßmann**, I am also happy you joined me as my co-promotor. Your sharp knowledge of Deep Learning helped me refine the last two chapters of this thesis. I appreciate your constant support and valuable insights, which have been essential to the successful completion of this thesis. **Ernst Th. Scholten**, although you are not officially my co-promotor, I consider you one; your clinical perspective was critical for this thesis. Thank you very much for your kind way of explaining complex anatomy in simple terms. Your willingness to offer assistance and demonstrate practical clinical applications has been invaluable to this thesis.

To my paranymphs: **Anton Schreuder**, it was fun helping you with some of the scripts you needed for your papers. Luckily, you did not need help for all 20 papers you published during your PhD. Thanks for sharing every board game possible with me. I noticed clearly that I disappointed you when I told you I liked Monopoly. I

am sorry for that. It was an honor to be your "paralele", and I retribute the honor of having you as one of my "paralele". **Weiyi Xie**, we had many fun conversations, from Deep Learning to random topics. I admire your way of coming up with Deep Learning solutions to every problem. Thanks for joining the beer group that Ecem started. I am so pleased to have you as my "paralele".

To my friends who helped me a lot during the first years of my PhD: **Arnaud Setio** (aka Arnoud), you could easily be one of my co-promoters. From you, I have learned Theano (R.I.P.), Mevislab, and many new Pokemon names that I did not know existed. Apparently, we ran out of Pokemon names, and now we talk about stocks. You contributed a lot to my first journal paper, and I am happy to have you as a co-author. **Nadine Kraamwinkel**, I had the pleasure of guiding you through your MSc thesis project at DIAG. Despite our different backgrounds, we enjoyed discussing topics like deep learning, alpacas, and "happy livers". Our friendship deepened further when my daughter arrived just 10 days after your son was born. I'm grateful for all the parenting advice you've shared; now, you've become my mentor in the journey of parenthood. **Carl Shneider**, always with a smile and a, usually Star Wars, related joke. I am sorry for pushing you to create a LinkedIn profile. I would never have imagined that you would become a LinkedIn superstar. Thanks for all the pictures of the Peruvian granadillas you eat from time to time.

To the DIAG Turkish community: **Ecem Sogancioglu** (aka Eceu) , thanks for the gossip, beers, wine, and parties. I am so glad we have helped each other in many ways. I am sorry for the biscuit joke, but I could not resist. I always enjoyed the way you get easily lost with Google Maps but I am happy we found our way to finish this PhD journey, you are up next. **Erdi Çalli**, I admire how you handled everything with patience; I barely saw you stressed during your PhD journey. You are so committed that you finished a Master's degree while you were finishing your PhD. Thanks for the trap during my last days at DIAG and for giving me a taste of my own medicine; really memorable.

To my lunch group: **Kiran Vaidhya Venkadesh**, thanks for taking over my role as "the lunch guy". Your PhD path has been amazing, I will complain to your Manuscript committee if they do not give you cum laude. **Matin Hosseinzadeh** (aka "Ma(r)tin"), I'm sorry for bringing you to the bar regularly, and I appreciate your sharing your tricks with me. Although we come from distant countries, it was very nice to see that our cultures have many things in common. **Luuk Boulogne**, thanks for taking care of Valiente. **Kicky van Leeuwen**, thanks for organizing the videocalls during the lockdown. Big thanks to **Max Argus, Joep Kamps, Saurabh Jain, Joris Bukala, Babak Ehteshami Bejnordi, Mohsen Ghafoorian, Patrick Brand, Nils & Ward Hendrix, Riccardo Samperna, Christiana Balta, Coen de Vente**, and

Zijian Bian for enjoying the tasty food at the hospital's restaurant with me.

To the Pathology guys: **Francesco Ciompi**, you were one of my first contacts before arriving in the Netherlands; thanks for all the conversations we had in Spanish and for providing all the information before my arrival. **Péter Bandi**, thanks for all the well-organized information for newcomers, from mobile operators to how to get a DigiID; your tutorial saved my life.

To the Research Software Engineering team: **Paul Gerke**, you saved my life many times, from DeepR (R.I.P.) to cluster issues, big thanks for that. I never saw someone so happy and excited to fix complex bugs. Although everyone was so happy about getting your help, there is a legend that says keyboards are afraid of you. **James Meakin**, I have learned a lot from you, from Docker tricks to fancy Python coding style, thanks for sharing your wisdom. **Haimasree** thanks for helping me with MevisLab.

To the teams I participated with: Thanks to the WebTeam, **Wouter Bulten, Meyke Hermsen**, and **Mart van Rijthoven** for giving DIAG a nice-looking website. To the DIAG weekend 2017 organization team, **Rashindra Manniesing, Bart Liefers, Arnaud Setio**, and **Dagmar Grob** for buying a lot of beers for the DIAGers. To the Spanish spirit of DIAG, **David Tellez, Cristina González**, and **Clarisa Sánchez**, for the karaoke sessions that lasted until breakfast time during the DIAG weekends. To the 4DCT guys, **Midas Meijs, Ajay Patel**, and **Sil van de Leemput**. I also enjoyed sharing happy moments with a wave of visiting researchers: **Usama Pervaiz, Tajwar Aleef, Yeman Brhane, Pedro Macias, Inês Bagulho, Alberto Rossi**, and **Ruben Kluge**. Thank you all for the fun moments.

Thanks to my Peruvian friends, **Alexinho, Herbert, Christian, Nataly, Marisol, Daniel, William, Laura, Juan, Olenka, Patty, Juan Carlos**, and **César**, for keeping me up-to-date with the new Peruvian slang, gossip, and jokes. I want to also show my gratitude to my Brazilian friends **Newton** and **Maysa**; sometimes, I feel that you guys are more Peruvian than me when sharing photos and videos of tasty Peruvian food. To the WhatsApp group "A llama de Peruano", **Heitor, Natalia**, and **Rafael**, thank you for sharing way too many pictures and videos of llamas; I hope you learned by now the difference between llamas, alpacas, guanacos, and vicuñas. The last bit of this paragraph in Portuguese to the people I also consider family, **Aninha, Elias**, e **Lion** muito obrigado por serem minha segunda família, pelo carinho e pelos gratos momentos, esse mundo precisa de pessoas com corações tão grandes quanto vocês.

This paragraph in Spanish is fully dedicated to my family. Creo que no hay palabras que puedan describir el esfuerzo que mis padres hicieron para poder darme una buena educación, muchas gracias **Bertita** y **Felo** por el esfuerzo que hicieron por

mi. A mis hermanos, **Saúl, Maritza** y **Elsa**, gracias por el ejemplo inculcado que fue fundamental para mi desarrollo professional. A mis cuñados **José** y **Yolanda**, muchas gracias por los sobrinitos y sobrinitas, por la cola de mono y el rico pisco. A mis sobrinitos **Kami, Pepito, Yaguito, Sebitas** y **Alondrita**, muchas gracias por los tiernos momentos, es increíble verlos crecer tan rápido.

To my girls: **Ingridd**, words cannot adequately express your boundless love and kindness. From the beginning of this journey, you were there by me and supported me through every high and low. Your belief in me has been a guiding light, giving me the strength to persevere. This achievement is as much yours as it is mine, and I am deeply grateful for your love and support. **Sofia**, mi preciosura, you brought so much joy into our lives, always lifting my spirits even after a tough day. Please keep waking me up by singing "Dad I love you, I love you..." or "Sofia acordouuuuuu". I love the both of you to the end of the world and beyond.

Curriculum Vitae

Gabriel Efrain Humpire Mamani was born in Arequipa, Peru, on the 3rd of September 1987. He pursued his undergraduate studies in Systems Engineering at the National University of San Agustin, Arequipa, Peru. During his undergraduate years, he embarked on a research endeavor, where he was responsible for developing the feature extraction module for a project aimed at detecting parasites in microscopical images. This project received funding from the Peruvian government and was conducted under the supervision of César Beltrán Castañon. In 2011, he decided to further his academic journey by pursuing a master's degree in Computer Science at the Institute of Mathematical and Computer Science at the University of São Paulo, Brazil. During this time, his research efforts centered on feature extraction and feature selection methods, specifically focusing on their application to medical imaging. His dedicated work in this area was conducted under the supervision of Agma Traina. After completing his master's degree, he dedicated three years to working in the Brazilian industry. In February 2016, he started working as a Ph.D. student at the Diagnostic Image Analysis Group. His doctoral research, supervised by Bram van Ginneken, Mathias Prokop, Colin Jacobs, and Nikolas Lessmann, focused on detecting and segmenting organs in cancer patients. This thesis documents the culmination of these academic and research efforts and encapsulates the results and findings of his work.

PhD Portfolio

Name:	Gabriel Efrain Humpire Mamani
Department:	Radiology and Nuclear Medicine
Graduate school:	Radboud Institute for Health Sciences (RIHS)
PhD period:	11-02-2016 — 10-02-2020
PhD Supervisors:	Prof. dr. Bram van Ginneken
	Prof. dr. Mathias Prokop
PhD Co-supervisors:	Dr. ir. Colin Jacobs
	Dr. Nikolas Lessmann

Training activities (Year)	**Hours**
Courses	
- Deep Learning 101 workshop (2016)	84.00
- RIHS - introduction course for PhD students (2016)	15.00
- RU - Scientific writing for PhD candidates (2017)	84.00
- RU - Presentation skills (2017)	42.00
- Radboudumc - Scientific integrity (2017)	16.00
Seminars	
- Radboud New Frontiers (2016)	28.00
- NFBIA Summer School (2017)	40.00
- NFBIA symposium (2017)$*$	9.00
- Deep learning Nijmegen Meetup (2018)	4.00
- Annual DIAG-FME symposium (2016-2019)$*$	56.00
- Medical Imaging Symposium for PhD students (2016-2019)	14.00
Conferences	
- SPIE medical imaging conference (2017) $\wedge+$	84.00
- Medical Imaging with Deep Learning (2018-2019)	56.00
Other	
- Radboudumc - General Radboudumc introduction for research personnel (2016)	9.00
- Weekly AMI Oncology meeting (2016-2017)	42.00
- Weekly DIAG Discussion Hour (2016-2020)	196.00
- Weekly Chest/Body CT meeting (2016-2020)	196.00
- Radboud Imaging research meetings (2016-2020)	28.00
- Weekly Deep learning journal club (2017-2020)	196.00
Teaching activities	
Supervision of internships/other	
- Supervision of a Master student graduation project (2017)	56.00
- Teaching assistant at Intelligent Systems in Medical Imaging (ISMI) (2017)	39.00
Total	1,294.00

$*$ Poster presentation.

$\wedge$ Oral presentation.

$+$ Demo presentation.

Research Data Management

Primary and secondary data used in Chapters 2-4 are stored in centrally stored and regularly backed-up Radboudumc servers accessible by members of the Diagnostic Image Analysis Group (DIAG). Chapter 5 of this thesis uses publicly available datasets (1-12) that can be accessed online after registration. Algorithms are stored in a private GitHub repository accessible by DIAG members.

The algorithm described in Chapter 3 can be used online at `http://grand-challenge.org/algorithms/spleen-segmentation/` . The dataset used in Chapter 4 is available at `https://doi.org/10.5281/zenodo.8014290`. The source code used for the experiments presented in Chapter 5 is publicly available on GitHub at `https://github.com/DIAGNijmegen/MedicalTransferLearning3D-UNet` .

1. Kidney Tumor Segmentation Challenge (KiTS19)[104]
   `https://kits19.grand-challenge.org`

2. LiTS - Liver and Tumor Segmentation Challenge[96]
   `https://competitions.codalab.org/competitions/17094`

3. Multi-atlas Labeling Beyond the Cranial Vault challenge[119]
   `https://www.synapse.org/#!Synapse:syn3193805/wiki/217789`

4. The Cancer Image Archive Pancreas-CT dataset[118]
   `https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT`

5. Multi-organ Abdominal CT Reference Standard Segmentations[117]
   `https://zenodo.org/record/1169361`

6. Automatic Structure Segm. for Radiotherapy Planning challenge (StructSeg2019)
   `https://structseg2019.grand-challenge.org/`

7. AAPM Thoracic Auto-Segmentation Challenge[144]
   `http://aapmchallenges.cloudapp.net/competitions/3`

8. Visceral dataset[116]
   `http://www.visceral.eu/benchmarks/anatomy3-open/`

9. Computational Methods and Clinical Applications for Spine Imaging (CSI) workshop[145]
   `https://csi-workshop.weebly.com/challenges.html`.

10. Large Scale Vertebrae Segmentation challenge (VerSe19)[146,147]
    `https://verse2019.grand-challenge.org/`

11. Regan E. A., Hokanson J. E., Murphy J. R., Make B., Lynch D. A., Beaty T. H., Curran-Everett D., Silverman E. K., and Crapo J. D. Genetic epidemiology of COPD (COPDGene) study design. *COPD*, 7:32–43, 2010

12. PROMISE12 challenge[149]
`https://promise12.grand-challenge.org/`

Radboudumc
university medical center

Radboud University