KU Leuven
Biomedical Sciences Group
Faculty of Medicine
Department of Imaging and Pathology
Division of Nuclear Medicine & Molecular Imaging

KU LEUVEN

DOCTORAL SCHOOL
BIOMEDICAL SCIENCES

# SEMI-AUTOMATED IMAGE SEGMENTATION AND REGISTRATION, FACILITATING MULTI-MODAL IMAGE ANALYSIS FOR SELECTIVE INTERNAL RADIATION THERAPY (SIRT)

Xikai Tang

Jury:

Supervisor:
Professor Johan Nuyts
Co-supervisor:
Professor Christophe M. Deroose
Professor Kristof Baete
Chair examining committee:
Professor Steven Dymarkowski
Chair public defence:
Professor Tania Roskams
Jury members:
Professor Arthur J.A.T. Braat
Professor Mark Konijnenberg
Professor Geert Maleux
Professor Patrick Dupont

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Biomedical Sciences

June 2022

# List of abbreviations

**Bq** becquerel

**CBCT** cone beam CT

**CCA** intrahepatic cholangiocarcinoma

**cDVH** cumulative dose-volume histogram

**CNN** convolutional neural network

**CT** computed tomography

**DSC** dice similarity coefficient

**DVH** dose-volume histogram

**[$^{18}$F]FDG** fluorine-18-fluorodeoxyglucose

**FOV** field-of-view

**[$^{68}$Ga]Ga-DOTA-TATE** [$^{68}$Ga]Ga-DOTA-Tyr$^{3}$-octreotate

**HCC** Hepatocellular carcinoma

**HD** Hausdorff distance

**$^{166}$Ho** holmium 166

**HU** Hounsfield Unit

**$^{131}$I** iodine-131

**IVC** inferior vena cava

**LOR** line-of-response

**LPT** liver perfusion territory

**LSF** lung shunt fraction

**MI** mutual information

**MR** magnetic resonance

**MRI** magnetic resonance imaging

**MSD** mean surface distance

**NTCP** normal liver tissue complication probability

**PET** positron emission tomography

**PV** portal vein

**RFA** radiofrequency ablation

**ROI** region of interest

**SIRT** selective internal radiation therapy

**SNR** signal-to-noise ratio

**SPECT** single photon emission computed tomography

**SSQD** sum of squared difference

**TACE** transarterial chemoembolization

**TARE** transarterial radioembolization

**$^{99m}$Tc-MAA** Technetium-99m-macro-aggregated albumin

**TCP** tumor control probability

**TE** echo time

**T/N ratio** tumor to normal tissue activity concentration ratio

**TOF** time-of-flight

**TPR** true positive rate

**TR** repetition time

**VOI** volume of interest

**$^{90}$Y** yttrium-90

# Table of Contents

# 1. Introduction

## 1.1 Overview

Selective internal radiation therapy (SIRT) aims at treating surgically unresectable liver tumors, including the primary ones and metastases. It deposits therapeutic microspheres loaded with β-emitting radionuclides like yttrium-90 ($^{90}$Y) and holmium-166 ($^{166}$Ho) selectively, by injecting the microspheres through a catheter into the arteries supplying the target tumors. The aim is to destroy the tumors with the ionizing radiation emitted by these radionuclides. The emitted β-particles have a short tissue penetration range, which prevents the radiation from reaching the surrounding normal cells. Since 80% to 100% of the blood supplying the tumors are from the hepatic arteries and the normal liver parenchyma acquires 80% of its blood supply from the portal vein, the deposition of the microspheres near the tumors can be achieved by selecting the correct hepatic arteries for injection while avoiding the deposition of microspheres in the normal liver tissue.

The SIRT workflow relies heavily on medical imaging studies. Before the treatment, anatomic imaging like contrast-enhanced CT and MR and nuclear medicine imaging like [$^{18}$F]fluorodeoxyglucose (FDG) or [$^{68}$Ga]Ga-DOTATyr$^{3}$-octreotate ([$^{68}$Ga]Ga-DOTA-TATE) PET/CT are performed for tumor diagnosis. Once SIRT is determined as a potential therapy for a patient, a pre-treatment workup is performed. Angiography is used to select the target branches of the hepatic arteries for injection and cone beam CT (CBCT) is performed to reveal the perfusion territory of each catheter position. After that, technetium-99m-macro-aggregated albumin ($^{99m}$Tc-MAA) particles are injected into the arteries leading to the tumors and a SPECT/CT scan is performed to simulate the activity distribution of the therapeutic microspheres. If the activity is found to be deposited as intended, the patient is selected for the treatment. The therapeutic microspheres are injected via the same arteries selected in the pre-treatment workup and a PET/CT or PET/MR scan is performed to verify the actual activity distribution.

The absorbed dose serves as a radiotoxicity indicator in SIRT; it is used to determine the injected activity and can be used to verify the treatment performance. Its calculation requires the activity distribution map and the volume of interests (VOI) including the liver and tumors. The VOI segmentation is important for the treatment precision. Manual VOI delineation for each patient is a tedious and labor-intensive task, which impeded the application of VOIs for precise and patient-specific SIRT planning and verification. Besides, the activity distribution and VOIs are acquired from multi-modal images in different studies performed at different time points. The lack of good co-registration of multi-modal images increases the difficulty of performing multi-modal image analysis for a precise treatment. Therefore, this project aims at developing (semi-)automatic image segmentation and registration methods to assist the nuclear medicine physicians to perform personalized and precise SIRT, which is encouraged by article 56 of the Council Directive 2013/59/Euratom for the improvement of radionuclide therapy [1].

## 1.2 The liver

### 1.2.1 Liver anatomy

The liver is one of the largest organ in the body, which weighs approximately 1500 grams for an adult [2]. The liver is an essential organ for many biochemical functions, such as bile production for fat breakdown and detoxification of the organism. It plays an important role in the metabolism by regulating glycogen storage and producing hormones.

The liver occupies the upper right quarter of the abdomen under the diaphragm. It is protected by the rib cage and fixed by ligamentous attachments [3], including the falciform ligament, the round ligament, and the right triangular ligament (Figure 1.1).

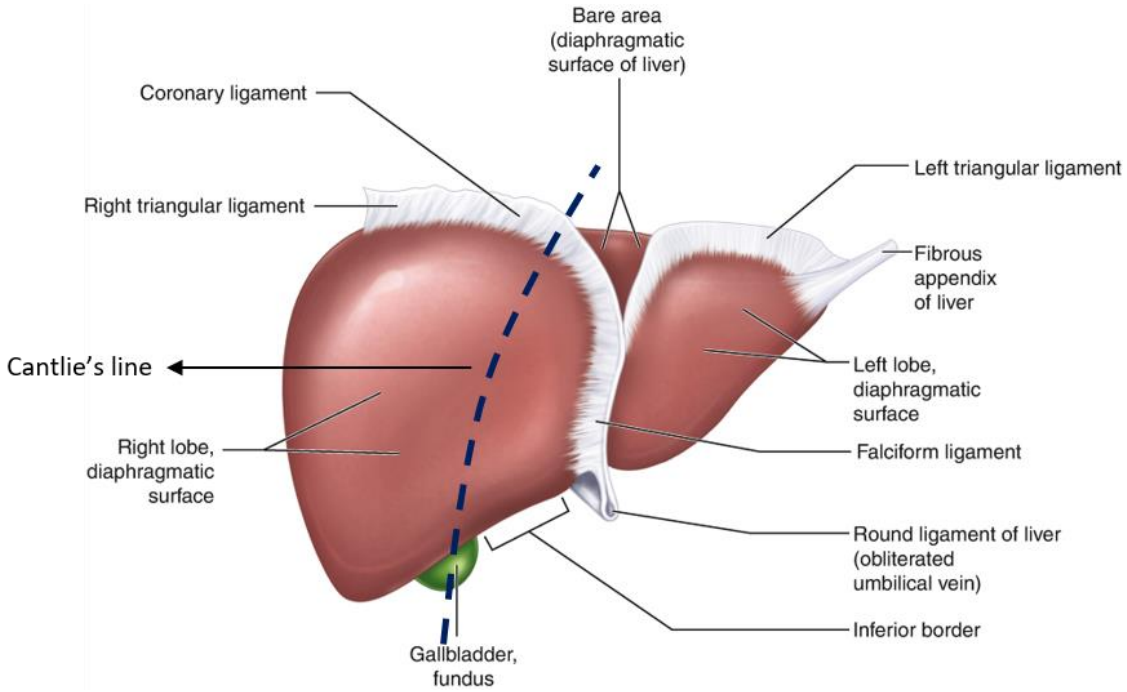



Figure 1.1 Ligaments of the liver. Adapted from Kwon, J.K. *et al.* (2021). Liver. In: Paltiel, H.J., Lee, E.Y. (eds) Pediatric Ultrasound. Springer, Cham. https://doi.org/10.1007/978-3-030-56802-3_11; with permission.

Historically, the liver was separated into the right and left lobes based on the falciform ligament. The anatomical and functional understanding of the liver has evolved in the past centuries. James Cantlie observed a line lateral to the falciform ligament from an atrophic right liver of a patient at autopsy [4]. This line, known as Cantlie's line, starts from the fundus of the gallbladder and ends at the inferior vena cava (IVC) [4]. It demarcates the distributions of the right and left portal veins (PVs) and indicates the true anatomic division between the left and right liver [4]. Later Claude Couinaud proposed the principle of segment division of the liver, which is widely adopted by surgeons. According to the Couinaud classification, the liver segment must have independent vascular inflow, outflow, and biliary drainage [5]. The liver is divided into eight segments based on the bifurcation of the portal vein. The segments are numbered clockwise on an axial plane with the caudate lobe as segment I (Figure 1.2).

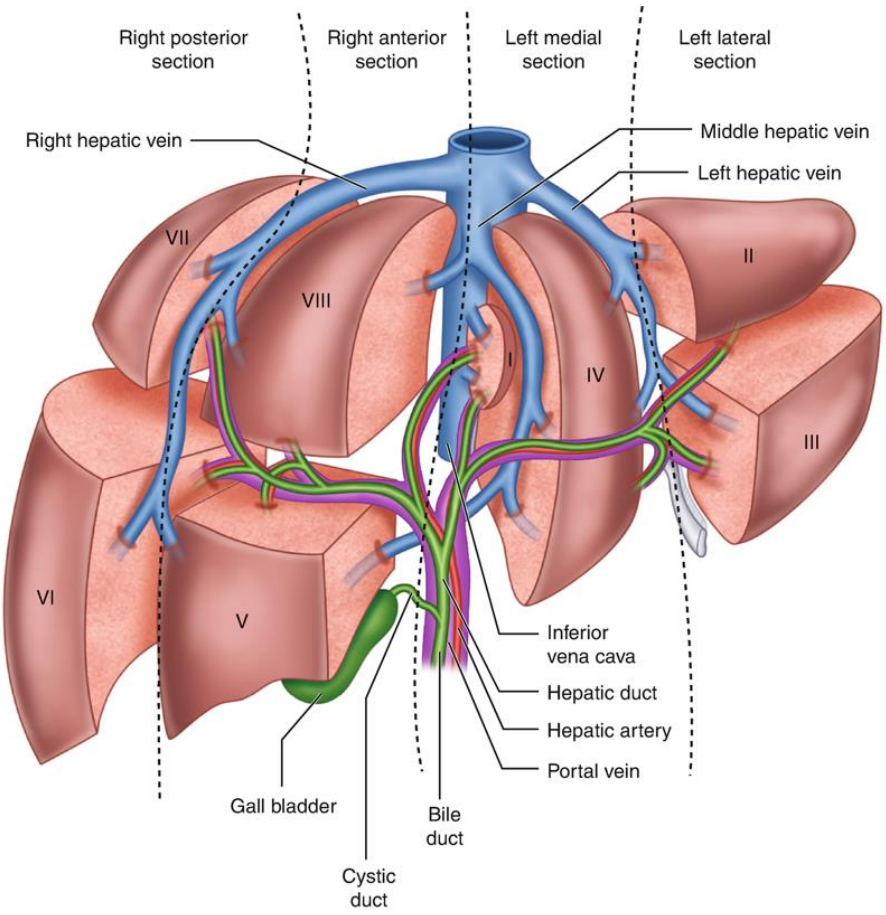



Figure 1.2 Couinaud classification. From Kwon, J.K. *et al.* (2021). Liver. In: Paltiel, H.J., Lee, E.Y. (eds) Pediatric Ultrasound. Springer, Cham. https://doi.org/10.1007/978-3-030-56802-3_11; with permission.

The liver consists of five types of cells, which are parenchymal and nonparenchymal. There is only one parenchymal cell type within the liver, hepatocytes. They occupy the major part of the liver and are responsible for all kinds of biochemical activities of the liver, including synthesis and storage of glycogen, bile production, etc. The other four types of cells in the liver are nonparenchymal. Sinusoidal endothelial cells form a permeable barrier that allows the communication of blood cells with hepatocytes. Kupffer cells are specialized phagocytes in the liver, which have an immunological function [6]. Hepatic stellate cells store vitamin A and are related to the pathological liver remodelling process [7]. Cholangiocytes transport bile and secrete bicarbonate for maintaining the pH in the liver [8].

## 1.2.2 Liver vasculature

The liver is highly vascular, and consumes the largest portion (25%) of the cardiac output [3]. It has a dual blood supply from the hepatic artery and the portal vein. The hepatic artery supplies 25% to 30% of the blood and the portal vein supplies 70% to 75% of the blood [3]. The blood from the hepatic artery and the portal vein blends into the hepatic sinusoids and drains into the hepatic veins [9].

The liver has a highly variable arterial vasculature for different people. The most common arterial configuration is presented in Figure 1.3A. The celiac artery, arising from the abdominal aorta, branches into the left gastric artery, splenic artery, and the common hepatic artery. The proper hepatic artery originates from the common hepatic artery together with the gastroduodenal artery [3]. The proper hepatic artery proceeds towards the liver and divides into the left and right hepatic arteries to supply the left and right liver lobes [3]. For approximately 15% of patients, the left hepatic artery originates from the left gastric artery [3]. The abnormal right hepatic artery originates from the superior mesenteric artery for 20% of patients [3].

The portal vein supplies a large quantity of blood rich in nutrition from the stomach, intestines, pancreas, and spleen to the liver. The portal vein forms from the conflux of the superior and inferior mesenteric veins, the splenic vein, the left and right gastric veins, and the pancreatic veins (Figure 1.3B). It divides into the left and right portal veins near the liver hilum [3]. Before or after the main branching of the portal vein, a small branch to the caudate commonly occurs.

The low-oxygen blood from the liver drains into the intrahepatic veins, which ultimately flow together into three (left, middle, right) hepatic veins (Figure 1.3B). These three hepatic veins drain into the IVC, which brings the de-oxygenated blood back to the heart.

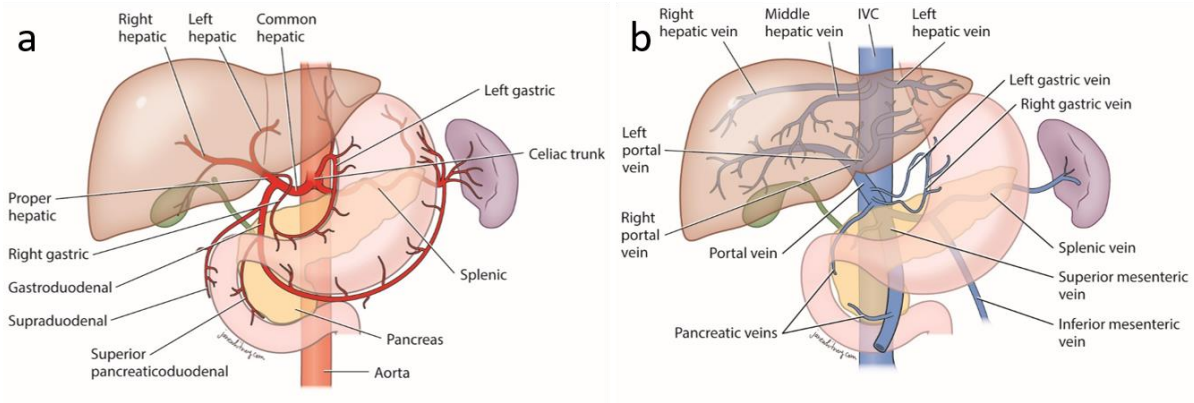



Figure 1.3 (a) Common hepatic arterial vasculature. © Jane Whitney CMI, used with permission. (b) Portal vein and hepatic venous vasculature. © Jane Whitney CMI, used with permission.

## 1.3 Liver cancer

Liver cancer is the sixth-most frequent cancer and the fourth cause of death from cancer in the world [10]. It can originate from liver cells, which is referred to as primary hepatic malignancy. Cancer can also start in other areas of the body, including the colon, lungs, etc. Cancer that spreads to the liver are liver metastases, which are also named secondary liver cancer. Liver metastases are 20 times more common than primary liver cancer [11].

### 1.3.1 Primary hepatic malignancy

Hepatocellular carcinoma (HCC) is the most frequent primary hepatic cancer, which contributes to approximately 75% of primary liver cancer. HCC is formed from hepatocytes, the main type of liver cell. Besides, intrahepatic cholangiocarcinoma (CCA), occurring in the bile duct, accounts for around 6% of primary liver cancers [12]. It is frequently seen in men between 60 and 70 years old [13]. Another type of less common liver cancer is hepatoblastoma, which forms from immature liver cells [12]. It is a very rare hepatic malignancy, which mainly occurs in children. Most hepatoblastomas develop in the right lobe [14].

Multiple causes exist for primary liver cancer, including chronic infection with hepatitis B virus and hepatitis C virus, cirrhosis, diabetes, nonalcoholic fatty liver disease, aflatoxin, etc. There are usually no symptoms in the early stage of primary hepatic cancer. Once the symptoms occur, people might experience weight loss without trying, loss of appetite, upper abdominal pain, nausea, vomiting, fatigue, jaundice, etc.

The diagnosis of primary liver cancer highly relies on different imaging modalities. Diagnosis of HCC can make use of computed tomography (CT), magnetic resonance imaging (MRI), and medical ultrasound. Large lesions have a high probability of being HCC in ultrasound images. HCC is best visible in the arterial phase of MR, where the contrast is present in the liver arteries. A liver biopsy is used when doctors are not certain about the diagnosis based on the imaging studies. For diagnosis of

CCA, endoscopic retrograde cholangiopancreatography and magnetic resonance cholangiopancreatography are used to detect bile duct obstruction [15].

The treatment for primary liver cancer depends on the tumor type and the stage of cancer. The common method for cancer staging is the TNM staging system [16]. Common treatments include medications, ablation, and surgery.

Surgical resection is recommended for HCC if patients retain sufficient functional hepatic parenchyma [17]. CCA usually has multiple focal tumors in the liver. Therefore, over 70% of CCA cases are not resectable at diagnosis. Surgical resection can also be used in treating hepatoblastoma to remove as much of the tumor as possible.

Liver transplantation is used for HCC treatment if the treatment is tolerable and the tumor meets criteria, like the Milan criteria [18]. Patients undergoing liver transplantation usually have severe liver dysfunction or multiple tumors.

Ablation is the other type of curative care for liver cancer besides surgery. The ablation of tumors can be achieved by chemical injection or the creation of high temperatures via microwaves, radiofrequency wave, lasers, etc. Radiofrequency ablation (RFA) and microwave ablation are most commonly used for HCC ablation, which are recommended for small and localized tumors. But it is unable to eliminate tumors near other organs or vessels because of heat generation and the heat sink effect, respectively [19]. RFA can also treat CCA, which may provide successful local control of tumors of intermediate (3-5 cm) or small (< 3 cm) diameter [20].

Transarterial chemoembolization (TACE) is used as a non-invasive treatment of HCC or downstaging of HCC for liver transplantation. It is also used to treat a regional recurrence of tumors after surgical resection. As a minimally invasive procedure, small particles loaded with chemotherapeutic drugs are injected through a catheter into selected arteries that supply the tumors [21]. These particles obstruct the blood supply and deliver cytotoxicity to the tumors. TACE can also be applied for the treatment of CCA and hepatoblastoma.

The radiotherapeutic analogue of TACE is selective internal radiation therapy (SIRT), also named transarterial radioembolization (TARE). Patients with surgically unresectable tumors, such as HCC or metastasis, can be selected for SIRT. Microspheres loaded with radionuclides are injected into the arteries supplying target tumors. These microspheres are deposited in small vessels surrounding the tumors and emit ionizing radiation that can kill the tumors.

### 1.3.2 Liver metastasis

Liver metastases constitute secondary liver cancer. Since the liver has a dual blood supply from the hepatic artery and portal vein, it is a common organ for metastasis. Colorectal cancer (CRC) is the most frequent tumor metastasizing to the liver [22]. Other organs causing liver metastasis include the pancreas, lung, breast, kidney, etc. [22].

Multiple imaging modalities including ultrasound, CT, and MRI are widely used for the diagnosis of liver metastasis. Liver metastasis exhibits a high-echoic core surrounded by a wide hypoechoic band in ultrasound images [23]. In CT imaging, liver metastasis usually displays ring enhancement in the arterial phase, a central avascular region in the portal phase, and delayed enhancement in the equilibrium phase [22]. MRI with contrast agents shows high diagnostic accuracy for early detection of liver metastasis in combination with diffusion-weighted imaging [24]. Besides anatomic imaging, evaluation of tumor metabolism can help estimate the underlying biological progress and therapy response. Positron emission tomography (PET) scan has the advantage of revealing the metabolic activity of metastasis by evaluating the uptake of radio-labeled molecules [22]. [$^{18}$F]fluorodeoxyglucose ([$^{18}$F]FDG) is a widely used tracer in PET-CT for the evaluation of neoplastic tissue [25]. [$^{18}$F]FDG PET exhibits high accuracy compared to CT [25].

Liver transplantation is usually not used for patients with liver metastasis. The efficacy of liver transplantation for the treatment of liver metastasis remains controversial [26]. When surgical resection is not possible, percutaneous ablation, including RFA and microwave ablation, is widely recommended for the treatment of liver metastasis. RFA is recommended for colorectal liver metastasis if the number of tumors is smaller than 5 and the tumor size is smaller than 3 cm according to the Cardiovascular and Interventional Radiological Society of Europe (CIRSE) criteria [27]. Since most liver metastases receive the blood supply from the hepatic artery, it makes transarterial embolization a possible option. Transarterial embolization is divided into TACE and SIRT, depending on whether the spheres contain chemotherapeutics or radionuclides.

## 1.4 Nuclear medicine and radionuclide therapy

Nuclear medicine is a medical speciality closely related to radiology and radiotherapy. It is characterized by the use of radioactive substances or tracers inside the body for diagnosis and therapy. Nuclear medicine imaging is used to reveal functional information within the human body via the uptake of radiopharmaceuticals. It involves multiple disciplines, such as physics, mathematics, computer technology, chemistry, and medicine.

### 1.4.1 Nuclear medicine imaging

Unlike computed tomography (CT), nuclear medicine imaging forms images by using internal radiation sources instead of external sources like X-rays. A part of the radiation escapes from the body and is detected by an imaging system, and from these data an image is computed. After administering a radiopharmaceutical, its uptake by a certain organ is revealed as hyperintense or hypointense areas in nuclear medicine images. By detecting the radiation emitted from a certain region, its local concentration can be estimated, which can help probe disease-related tissues and the response to treatment.

The radioactive radionuclides used in nuclear medicine can emit gamma photons (a form of electromagnetic radiation) and/or positrons during their radioactive decay, which can be used for imaging.

**Gamma camera**. The gamma camera was developed to detect gamma photons. It contains a scintillation crystal. When the gamma photon collides with the atoms in the scintillation crystal, a short and localized pulse of electromagnetic radiation in the range of visible light is produced. The gamma camera can count the optical photons generated by the scintillation crystal interacting with gamma photons. In this way, the incident number of gamma photons can be measured. A collimator is used to constrain the acceptance angle of the gamma photons, to ensure that positional information is provided for the detected photons. In almost all cases, the collimator imposes a very small acceptance angle, such that the gamma camera measurement can be well modelled as a set of line integrals, usually called projections.

**Planar imaging**. Planar imaging uses the gamma camera to detect the gamma photons emitted by the radioactive decays. After the radionuclides are injected into the body, the gamma camera keeps stationary to form 2D images (projections) for planar imaging. The images are usually generated from the front or the back of the patient.

**SPECT**. Like planar imaging, SPECT also measures gammy rays emitted by the injected radionuclides, making use of scintillation in the gamma camera. However, SPECT can produce true 3D images. The gamma camera is rotated during the SPECT scan to obtain multiple 2D projections. Each projection is obtained at a different angle. Then 3D images are formed from the 2D projections by using a tomographic reconstruction algorithm, which can give an estimate of the 3D radioactivity distribution [28]. SPECT is not able to localize the radiation event precisely, because the finite

collimator acceptance angle creates a distance-dependent blurring. A high resolution collimator only accepts a small number of gamma photons, resulting in a low sensitivity. The sensitivity can be increased by allowing more gamma photons to pass the collimator (larger acceptance angle), but this adversely affects the spatial resolution [28].

**PET**. The radionuclides used in PET, such as $^{90}$Y, can emit positrons during the radioactive decay. Each positron annihilates with an electron in a range of a few millimeters. This annihilation results in the emission of two gamma rays (511 KeV) in opposite directions. If the two annihilation photons are detected by the PET scanner within a short time interval, it is recorded as a coincidence event. The annihilation point must be located on the straight line between the spots where the two coincident photons are detected, named line-of-response (LOR). The traditional PET records all LORs seen by the detectors in the projections acquired at different angles and assumes a uniform probability of the positron emission points along the LORs [29]. Thus, similar to SPECT, the traditional PET system acquires line integrals of the activity distribution in the field of view. From the measured counts along all LORs, a 3D image of the tracer distribution is reconstructed. The PET data are corrupted by Poisson noise, which can propagate into the final reconstruction. Recording the time difference in the detection of two coincident gamma photons makes it possible to estimate the location of the positron annihilation along the LOR [29]. The PET using this technology is named time-of-flight (TOF) PET. The TOF PET has higher SNR than the traditional one.

Nuclear medicine imaging can provide molecular information on physiology and metabolism, but cannot provide much information on anatomy [30]. If the radioactive tracer is non-specific and binds with the tissues distributed in the whole body (e.g., muscles, fat), some anatomical information can be extracted from nuclear medicine imaging. If more specific tracers are used for the study of specific cell functions, anatomical imaging is needed to provide anatomical information. Therefore, multi-modality imaging systems have been developed to deal with this problem. In SPECT/CT, PET/CT, or PET/MR systems, nuclear medicine imaging is combined with anatomic imaging in the same equipment so that the two scans can be performed almost simultaneously. This simplifies the registration of nuclear medicine and anatomic images by decreasing as many variables as possible, like the anatomy change over the time interval between two imaging studies. Besides, the anatomic imaging can be used for attenuation correction in nuclear medicine imaging.

Nuclear medicine imaging is used for diagnosing tumors and revealing the tumor response to treatment, monitoring the blood flow and metabolism of cells, and detecting the disease-related states of tissues. For example, fluorine-18-fluorodeoxyglucose (FDG) PET can differentiate most tumors from normal tissues based on the different uptake of glucose.

### 1.4.2 Radionuclide therapy

Radionuclide therapy aims at treating diseases (especially cancer) by using radiopharmaceuticals. It delivers radiopharmaceuticals to the target region by injection or ingestion. The radiopharmaceuticals emit ionizing radiation to kill cells in the target region. Radionuclide therapy can be used independently as an alternative to surgery or other medical treatments. It can also be incorporated into a therapeutic strategy with other treatment methods for curative or palliative intent. Radionuclide therapy has the potential to be highly selective and provide limited toxicity.

Radiopharmaceuticals selected for radionuclide therapy can emit β-particles, α-particles, and Auger electrons, which have a short penetration range [31]. A successful radionuclide therapy should result in sufficient uptake of radioactivity by the targets by remaining in the target tissues for a prolonged period. The calculation of the administered activity remains challenging because the radiopharmaceutical distribution in the target tissues is non-uniform [31].

Radionuclide therapy is widely used for the treatment of various diseases. One of its oldest applications is thyrotoxicosis, in which excess thyroid hormones circulate in the body. The radionuclide,

iodine-131 ($^{131}$I), is used for the treatment, which is proven to have high efficacy [31]. $^{131}$I is also applied to the treatment of thyroid carcinoma in combination with total or near-total thyroidectomy to remove any residual thyroid tissue. Lutetium-177 ($^{177}$Lu) is widely used for targeted radionuclide therapy. The $^{177}$Lu-based agent, $^{177}$Lu-DOTA-TATE, is used for peptide receptor radionuclide therapy [32]. The $^{177}$Lu-DOTA-TATE can target peptide receptors to deliver ionizing radiation that can kill neuroendocrine tumors. Actinium-225 ($^{225}$Ac) emits α particles during decay, which are used for targeted alpha therapy. $^{225}$Ac is conjugated to antibodies to treat diseases like bladder cancer, ovarian cancer, and melanoma [33]. Radionuclide therapy is also applied to locoregional treatment, where the inoperable tumors are localized in an organ (e.g., the liver). The arteries usually supply a larger amount of blood to these tumors than to normal tissue. This makes it possible to deposit the radiopharmaceuticals selectively in the tumor's arterioles and capillaries.

## 1.5 SIRT

### 1.5.1 SIRT overview

SIRT is used as an alternative treatment when surgery is impossible for patients with liver cancer, including primary hepatic malignancies and metastases. It is also used as a bridging therapy [34] to downstage the tumor size for a later liver transplantation. Studies and trials on the use of SIRT to treat HCC [35, 36] and colorectal cancer liver metastases [37, 38] were conducted in the past decade. Their results did not show the superiority of SIRT over chemotherapy in metrics like overall survival and progression-free survival rate. However, SIRT does exhibit less severe adverse events, which makes it better tolerable for patients. Meanwhile, Salem demonstrated in a 15-year study that SIRT shows promising results as an earlier stage treatment, especially when conducting high-dose radiation segmentectomy [39].

Currently, three vendors are providing commercially available radiopharmaceuticals or radioactive microspheres for SIRT (see Table 1.1). Two of the vendors provide microspheres that are loaded with $^{90}$Y while the other one is loaded with holmium-166 ($^{166}$Ho). The $^{90}$Y microspheres are made of either resin (SIR-Spheres) or glass (TheraSphere). The $^{166}$Ho microspheres are made of poly-L-lactic acid (QuiremSpheres). All three microspheres are permanently deposited in the target area and emit high-energy β-particles to kill tumors. Both $^{90}$Y and $^{166}$Ho have a short mean penetration range of a few millimeters, which allows them to administer radiation locally without damaging the normal tissue nearby. Compared with $^{90}$Y, a pure β-emitter, the ionizing radiation of $^{166}$Ho contains both β-particles and gamma rays. The interaction of β-particles from $^{90}$Y decay with tissue generates photons which are used for imaging of bremsstrahlung SPECT. Decay of $^{90}$Y also produces a very small portion of positrons which can be used for PET imaging. The low yield of bremsstrahlung photons and positrons from β-decay of $^{90}$Y makes quantitative imaging difficult [40]. The gamma rays emitted by $^{166}$Ho decay give it superiority over $^{90}$Y in terms of imaging by the gamma camera. Besides, the paramagnetic property of holmium makes $^{166}$Ho visible and quantifiable in MRI even after decay [41]. The microspheres used for SIRT in our hospital are loaded with $^{90}$Y.

Table 1.1 Specifications of the three radioactive microspheres for SIRT

| Manufacturer | Sirtex Medical | Boston Scientific | Quirem Medical (Terumo) |
|---|---|---|---|
| Trade Name | SIR-Spheres | TheraSphere | QuiremSpheres |
| Material | Resin | Glass | poly-L-lactic acid |

| | | | |
|---|---|---|---|
| Activity per particle | 40-70 Bq | 1250-2500 Bq | 330-450 Bq |
| Number of microspheres per 3 GBq vial | 40-80 million | 1.2 million | 40-80 million |
| Radionuclide | Yttrium-90 | Yttrium-90 | Holmium-166 |
| Mean diameter | 32 μm | 25 μm | 30 μm |

The liver is featured by its dual blood supply system of the hepatic artery and portal vein. Around 70% to 75% of the blood is derived from the portal vein [3]. New arteries and sinusoidal capillaries are formed in the tumor during tumor growth, which results in a gradual shift in blood supply for tumors from the portal vein to the hepatic artery [42]. Therefore, the hepatic artery supplies around 90% of the blood to the liver tumors while the normal liver parenchyma is mainly supplied by the portal vein [42]. Tumors in different liver segments are supplied by different branches of the hepatic artery. Selective deposition of radioactivity in the target tumors becomes feasible by injecting the radioactive microspheres into the selected branches of the hepatic artery (see **Error! Reference source not found.**). These microspheres are permanently deposited in the small capillaries and deliver ionizing radiation to the tumors.

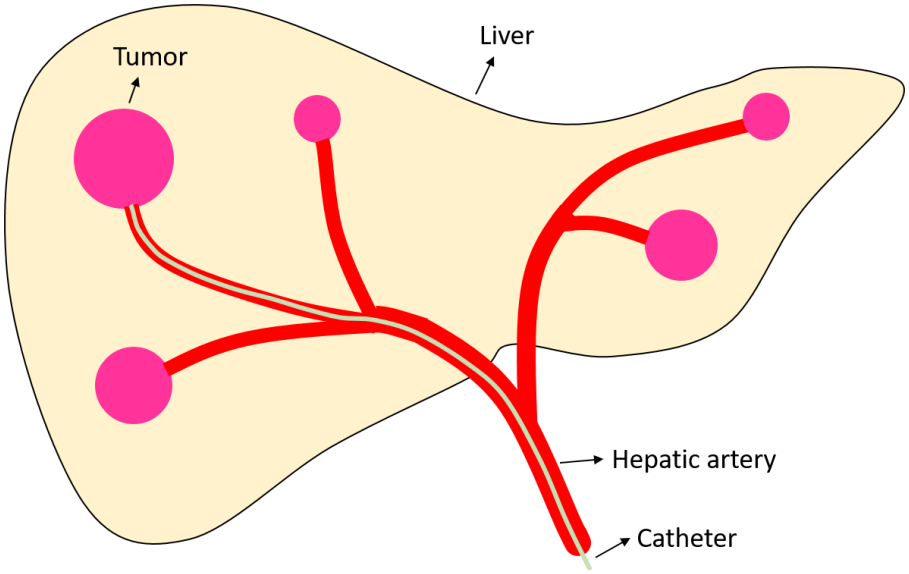



Figure 1.4 Demonstration of transarterial embolization and the blood supply to liver tumors. The catheter is positioned in the selected branch of the hepatic artery so that the injected microspheres can follow the catheter to reach the branch. Most microspheres are deposited in the target tumor because the arterial branch mainly supplies the target tumor.

### 1.5.2 Absorbed dose

The absorbed dose is defined as the amount of energy from ionizing radiation deposited per unit mass. It is commonly used to calculate the dose absorbed by the living tissue in radiation protection and radiotherapy. The SI unit of the absorbed dose is the gray (Gy), which is equal to J/kg. In SIRT, the absorbed dose serves as a radiotoxicity indicator, which is used for the prescription of the injected activity and evaluation of the treatment targeting performance.

One current schema for absorbed dose calculation in SIRT is proposed by the Medical Internal Radiation Dose (MIRD) Committee of the Society of Nuclear Medicine [43]. It is used to measure the mean absorbed dose in the target region from radiation emitted by the source regions. The spatial level

of the estimated regions can be whole organs, sub organs, or voxels [43]. The calculation of MIRD requires the definition of the source and target regions, the activity distribution as a function of time in the source regions, and the radionuclide S values. An S value is defined as the mean absorbed dose to a target region per radionuclide decay in a source region. The computation of the mean absorbed dose in the target region based on the S value is shown in Equation 1.1.

$$D(r_t) = \sum_{r_s} \tilde{A}(r_s) \cdot S(r_t \leftarrow r_s) \tag{1.1}$$

In Equation 1.1, $r_t$ is the target region, $r_s$ is the source regions, $\tilde{A}(r_s)$ is the cumulated activity in the source region, and $D(r_t)$ is the mean absorbed dose in the target region. The activity is defined as the number of radionuclide decays per unit of time. Its SI derived unit of activity is becquerel (Bq). The cumulated activity is used to measure the total number of radionuclide decays over a certain period. The use of SPECT and PET imaging allows in vivo measurement of activity distribution at the voxel level [44]. The radionuclide decay can emit different types of radiation particles, such as β-particles and photons. By determining the mean energy of each particle and its absorbed fraction to the target region, it becomes possible to estimate the S value. In Equation 1.2, $E_i$ is the mean energy of the particle $i$, $N_i$ is the number of particle $i$ emissions per radionuclide decay, $F_i$ is the energy absorbed fraction to the target region $r_t$ per emission of the particle $i$ in the source region $r_s$, and $m(r_t)$ is the mass of the target region.

$$S(r_t \leftarrow r_s) = \sum_i E_i \cdot N_i \cdot \frac{F_i(r_t \leftarrow r_s)}{m(r_t)} \tag{1.2}$$

The current radioactive microspheres used in SIRT are loaded with $^{90}$Y. The β-particles emitted by $^{90}$Y have a mean energy of 0.9267 MeV, a mean penetration range of 2.5 mm, and a maximum penetration range of 10 mm in water equivalent tissue. The half-life of $^{90}$Y is 64.2 hours. Most energy from $^{90}$Y decay within the voxel is trapped in the same voxel. Besides, these microspheres are designed to have a mean diameter of around 30 μm, which allows most of them to be permanently lodged in the capillaries near tumors. Based on the above two facts, the MIRD schema can be simplified for $^{90}$Y-SIRT by using the following assumptions: (1) the radioactive microspheres are permanently trapped in the tumor regions for the convenience of cumulated activity calculation; (2) the energy emitted by the decay of radionuclides in a source region is locally deposited within the same region, which means that $S(r_t \leftarrow r_s) = 1$ for $\mathrm{t} = \mathrm{s}$ and $S(r_t \leftarrow r_s) = 0$ for $\mathrm{t} \neq \mathrm{s}$. Therefore, the mean absorbed dose by 1 kg of mass from 1 GBq activity can be computed as follows:

$$\begin{aligned} \bar{D} = \frac{\text{absorbed energy}}{\text{mass}} = \frac{\int_0^\infty A_t \bar{E} dt}{\text{mass}} = \frac{\bar{E} \cdot \int_0^\infty (A_0 \cdot e^{-(\ln(2)/T_{1/2}) \cdot t}) dt}{\text{mass}} = \frac{\bar{E} \cdot A_0 \cdot T_{1/2}}{\text{mass} \cdot \ln(2)} \\ = \frac{(926.7\text{keV}) \cdot (1.602 \cdot 10^{-16} J/\text{keV}) \cdot (10^9 \text{Bq}) \cdot (64.0416 \cdot 3600\text{sec})}{1\text{kg} \cdot \ln(2)} \\ = 49.38\text{J/kg} = 49.38\text{Gy} \end{aligned} \tag{1.3}$$

In Equation 1.3, $\bar{D}$ is the mean absorbed dose in the mass, $\bar{E}$ is the average energy emitted per disintegration, $A_t$ is the activity in the mass at time $t$ from the injection, $A_0$ is the activity at the time of injection, and $T_{1/2}$ is the half life of $^{90}$Y. According to the above calculation, the mean absorbed dose of a volume of interest (VOI) is derived as $49.38\ [\text{Gy} \cdot \text{g/MBq}] \cdot \frac{\text{activity of VOI [MBq]}}{\text{mass of VOI [g]}}$.

## 1.5.3 Dose calculation methods

The MIRD schema can be performed on VOIs of different spatial levels, including the whole organ, sub organs and voxels. The computed absorbed dose can better reflect the underlying biological effects as the spatial resolution increases.

### 1.5.3.1 Mono-compartment method

The mono-compartment method assumes a uniform distribution of activity in a target volume, which can be the whole liver, different lobes, or different segments. It does not distinguish the activity distribution in the tumors and normal tissue. The calculation of the mean absorbed dose in the target volume is shown in Equation 1.4, where $A_{target}$ is the activity administered to the target volume and $M_{target}$ is the mass of the target volume. In SIRT, the liver density is assumed to be a constant, which is close to the water density. Therefore, the injected activity can be prescribed by measuring the target volume and determining the target mean absorbed dose. The recommended absorbed dose for the mono-compartment method ranges from 80-150 Gy [45]. The maximum limit of the activity shunting to the lung (lung shunt) should be 610MBq for $^{90}$Y, which means approximately 30 Gy in 1 kg lung tissue [46].

$$\overline{D} = 49.38 \cdot \frac{A_{target}}{M_{target}} = 49.38 \cdot \frac{A_{target}}{(\rho_{liver} \cdot V_{target})} \quad 1.4$$

The mono-compartment method ignores the large difference in activity concentration between the tumors and the normal tissue. It is prone to delivering the dose tolerable by the normal tissue so that the occurrence of radiotoxicity-related complications can be minimized. This might result in under-dosing of the tumors for some patients when the mean absorbed dose is dominated by the activity in the normal tissue.

### 1.5.3.2 Multi-compartment method

The target volume is divided into two compartments, tumors and normal tissue [46]. The multi-compartment method assumes uniform activity distribution in the tumors and the normal tissue, respectively. It also takes into account the lung shunt fraction (LSF) when calculating the absorbed dose. In this way, the mean absorbed dose in the tumors and the normal tissue can be calculated separately, which ensures maximizing the absorbed dose to the tumors while delivering a radiotoxicity tolerable by the normal tissue. The mean absorbed dose of each compartment is calculated as follows:

$$\overline{D}_i = 49.38 \cdot \frac{(1 - LSF) \cdot A \cdot FU_i}{\rho_{\text{liver}} \cdot V_i} \quad 1.5$$

$$FU_i = \frac{A_i}{A} \quad 1.6$$

In Equation 1.5, $\overline{D}_i$ is the mean absorbed dose of compartment $i$, $A$ is the total activity injected into the liver, $V_i$ is the volume of compartment $i$, and $FU_i$ is the fractional uptake of compartment $i$. $A_i$ in Equation 1.6 is the activity in compartment $i$. In practice, the fractional uptake of each compartment is estimated from the SPECT scan using Technetium-99m-macro-aggregated albumin ($^{99m}$Tc-MAA) as the tracer, which is described in detail in Selection 1.5.4. The ratio of the activity concentrations in the tumor and in the normal tissue, named tumor to normal tissue activity concentration ratio (T/N ratio), is calculated based on $^{99m}$Tc-MAA SPECT. T/N ratio is used to derive the fractional uptake in SIRT. The computation of the fractional uptake is shown in Equation 1.7, 1.8, and 1.9.

$$(T/N)_i = \frac{A(TUM_i)/V(TUM_i)}{A(NT)/V(NT)} \Rightarrow \frac{A(TUM_i)}{A(NT)} = (T/N)_i \cdot \frac{V(TUM_i)}{V(NT)} \quad 1.7$$

$$FU(TUM_i) = \frac{A(TUM_i)}{\sum_j A(TUM_j) + A(NT)} = \frac{A(TUM_i)/A(NT)}{\sum_j \frac{A(TUM_j)}{A(NT)} + 1} = \frac{(T/N)_i \cdot \frac{V(TUM_i)}{V(NT)}}{\sum_j [(T/N)_j \cdot \frac{V(TUM_j)}{V(NT)}] + 1}$$
$$= \frac{(T/N)_i \cdot V(TUM_i)}{\sum_j [(T/N)_j \cdot V(TUM_j)] + V(NT)} \quad 1.8$$

$$FU(NT) = 1 - \sum_j FU(TUM_j) = 1 - \sum_j \frac{(T/N)_j \cdot V(TUM_j)}{\sum_k [(T/N)_k \cdot V(TUM_k)] + V(NT)}$$
$$= 1 - \frac{\sum_j [(T/N)_j \cdot V(TUM_j)]}{\sum_k [(T/N)_k \cdot V(TUM_k)] + V(NT)}$$
$$= \frac{V(NT)}{\sum_k [(T/N)_k \cdot V(TUM_k)] + V(NT)} \quad 1.9$$

In the above equations, $(T/N)_i$ is the T/N ratio of tumor $i$, $TUM_i$ and $NT$ represents tumor $i$ and normal tissue, $FU$ denotes the fractional uptake.

Although the multi-compartment method provides better precision for dose calculation, it introduces a more labor-intensive workload into the clinical workflow, due to the delineation of compartments. This delineation can be very time consuming, especially when there are multiple tumors. The delineation of compartments can be avoided by computing the T/N ratio based on a small sampled volume from the compartments [47]. However, this may lead to large uncertainties on the estimation of the T/N ratio when the activity distribution is heterogeneous in the tumors and/or the normal tissue [48].

### 1.5.3.3 Voxel-level method

The voxel-based method assumes homogeneous activity distribution within the voxel, taking into account the heterogeneity of activity distribution on a small spatial scale. This allows a much more precise dose calculation than the mono- and multi-compartment methods. The target volume for dose calculation changes from the large-scale volume to the voxel. Dose planning and verification can be performed based on the mean absorbed dose in each voxel. A dose-volume histogram (DVH) is used to evaluate the distribution of the mean absorbed dose in each voxel over a volume (tumors, normal tissue, etc). The DVH is often converted to a cumulative DVH (cDVH) [49]. The cDVH displays the fraction of the total VOI that receives an absorbed dose higher than a certain value. One example of the cDVH is shown in **Error! Reference source not found.**.

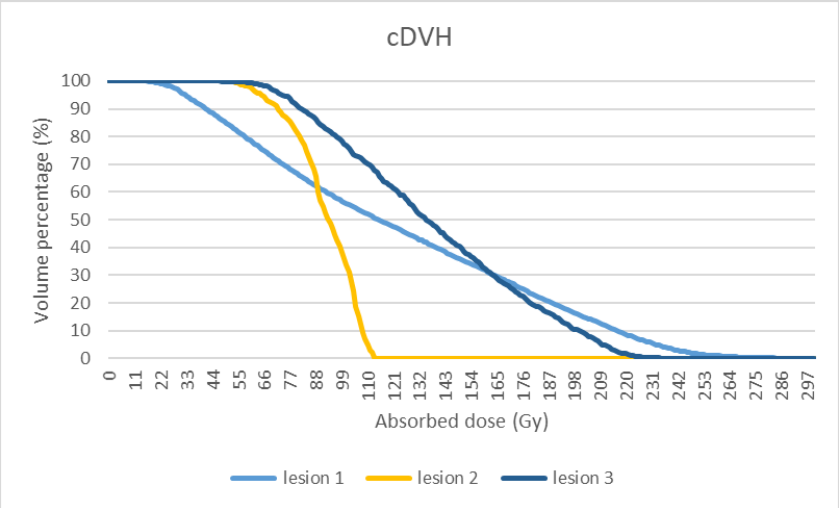



Figure 1.5 Example of the cDVH for three lesions.

The treatment performance is usually evaluated by some DVH parameters, including $D_V$ (the minimum absorbed dose that is received by $V\%$ of the VOI voxels, e.g. $D_{70}$) and $V_D$ (the volume percentage that receives at least $D$ Gy, e.g. $V_{100}$). These parameters are used to quantify the biological effect, which is related to the clinical outcomes, such as tumor control probability (TCP) and normal tissue complication probability (NTCP).

### 1.5.4 SIRT workflow

As is shown in Equation 1.5, the prescription of the injected activity based on different dosimetry methods and the evaluation of the actual dose distribution require the delineation of VOIs (the liver and tumors) and the activity distribution for the computation of the fractional uptake. The SIRT workflow is a complex procedure, consisting of pre-therapy tumor evaluation, patient selection, pre-treatment study, therapy planning, microsphere injection, and post-treatment study.

Before the therapy for the patient is selected, different medical imaging procedures are performed for tumor evaluation. Anatomical imaging like contrast-enhanced CT and nuclear medicine imaging like whole-body PET/CT with [$^{18}$F]FDG or [$^{68}$Ga]Ga-DOTA-Tyr$^3$-octreotate ([$^{68}$Ga]Ga-DOTA-TATE) are commonly performed for liver metastases. Contrast-enhanced MR is performed for diagnosis of HCC. These multi-modal images are used for delineation of the liver and tumors in SIRT planning and verification.

In the process of tumor evaluation, the patient's general condition, laboratory results, imaging study, and physical examination are considered to decide whether the patient should be further evaluated by pre-treatment workup. Only after pre-treatment workup, it can be determined if SIRT is an option for the patient.

During pre-treatment workup, angiography is performed as the first step. A catheter is guided under fluoroscopy by the interventional radiologist from the patient's femoral artery to the hepatic artery [50]. Angiography can be used to find the proper delivery catheter positioning, evaluate the blood flow, and determine the deposition site of the microspheres [50]. After angiography, contrast-enhanced cone beam CT (CBCT) imaging is used to show the perfusion territory of each catheter position. When the catheter positions are determined, $^{99m}$Tc-MAA particles are injected through the catheter into the designated positions. Shortly after the injection, planar imaging using the gamma camera is performed to estimate the LSF. The scan covers the trunk, starting from the top of the neck to the bottom of the hip. Regions of interest (ROIs) are drawn over the whole lung and the whole liver. The LSF is calculated based on the counts in the two ROIs. After that, a SPECT/CT scan is acquired to reveal the activity distribution of $^{99m}$Tc-MAA particles in the liver. Based on the SPECT/CT image, the medical team decides if the patient is to be treated with SIRT.

In therapy planning, the $^{99m}$Tc-MAA SPECT/CT scan is used to simulate the activity distribution of $^{90}$Y microspheres. The fractional uptake of each compartment or voxel is obtained by calculating the ratio of the counts in the target VOI to the counts in the treated segment or lobe. The injected activity for each treated segment or lobe is prescribed based on the target absorbed dose to the target VOI and the fractional uptake of the target VOI.

The designated amount of activity is prepared based on the prescription and injected by the catheter positioned in the selected artery as determined during pre-treatment workup. On the same day of injection or the day after it, a PET/CT or PET/MR scan is performed to verify the actual activity distribution of $^{90}$Y. The actual fractional uptake is calculated based on the $^{90}$Y PET. The absorbed dose in each target VOI can be computed based on the actual fractional uptake and the injected activity. By analyzing the absorbed dose, the biological effect and the radiotoxicity of the target region can be assessed to evaluate the treatment performance and the potential risk.

The physical properties of the radioisotopes used in SIRT are presented in

Table 1.2.

Table 1.2 Characteristics of the radioisotopes used in SIRT

| Radioisotopes | Half life | Emission | Mean energy | Branching ratio (%) |
|---|---|---|---|---|
| Yttrium-90 | 64.1 hours | $\beta^-_{0,0}$ particles | 926.7 keV | 99.983 |
| | | $\beta^-_{0,1}$ particles | 163.7 keV | 0.017 |
| | | $e^+/e^-$ pairs | 768.7 keV | 0.0032 |
| Technetium-99m | 6 hours | Gamma ray | 140.5 keV | 99.0 |
| Fluorine-18 | 1.83 hours | Positrons | 249.5 keV | 96.86 |
| Gallium-68 | 1.13 hours | Positrons | 836 keV | 89.1 |

## 1.5.5 Anatomic imaging for SIRT

As is mentioned in Section 1.5.4, the SIRT workflow includes a series of multi-modal imaging studies to obtain VOIs of the liver and tumors and the activity distribution. Anatomic imaging is briefly introduced in this section. The nuclear medicine imaging was introduced in Section 1.4.1.

### 1.5.5.1 CT

CT uses x-rays to acquire a depiction of the body anatomy noninvasively for diagnosis. An x-ray beam is partly absorbed or scattered when passing through body tissues. The remaining x-rays reach the detector, where a quantitative measurement of the energy is made. The linear attenuation coefficient defines the fraction of the attenuated x-rays in a monoenergetic beam per unit length of a mass. It is calculated in Equation 1.10, where $I_0$ and $I_x$ represent the initial intensity of the x-ray beam and the intensity at the distance of $x$ (cm). The Hounsfield unit (HU) is used as a quantitative scale of the attenuation coefficient in CT, defined in Equation 1.11.

$$\mu = \ln\left(\frac{I_0}{I_x}\right)/x \tag{1.10}$$

$$HU = 1000 \cdot \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{1.11}$$

CT is designed to measure the attenuation coefficients and reconstruct the spatial distribution of the attenuation coefficients of different tissues in a 3D volume. The main components of CT equipment are an x-ray source, a detector, a gantry, and a patient bed. The gantry is a donut-shaped structure, where the x-ray source and the detector are installed at opposite sides of the subject that is imaged in the filed-of-view (FOV). During a scan, the gantry is motorized to rotate around the body and the source generates a fan beam of x-rays through the body (see Figure 1.4). The patient, located on a bed, moves through the gantry at a constant speed until the whole FOV is scanned. Multiple 2D projections at different angles are acquired by the detector for each rotation. These 2D projections are reconstructed to form a 3D image through sophisticated mathematical algorithms. CT is widely used to visualize internal organs, soft tissue, blood vessels, and bones.

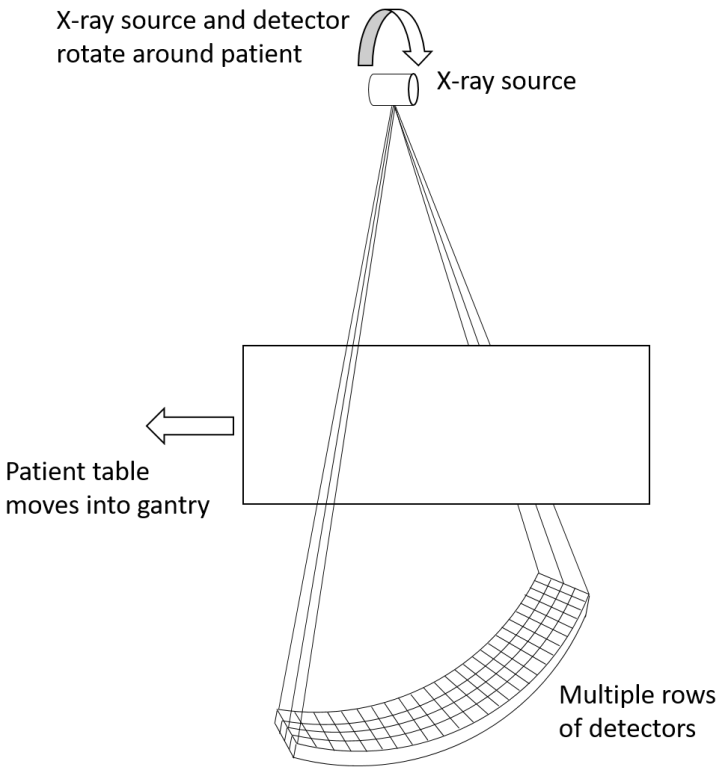



Figure 1.4 Demonstration of a CT scan. The x-ray source and the detector are assembled in a gantry. The gantry rotates around the body during a scan and the source generates a fan beam of x-rays around the body. The body moves through the gantry at a constant speed until the whole FOV is scanned.

The patient is exposed to ionizing radiation during the CT scan. The radiation amount is larger than that of a normal x-ray. This might result in damage to body cells. Doctors try to use the lowest radiation dose, which can produce an image with acceptable quality for clinical use.

### 1.5.5.2 CBCT

The CBCT system mainly consists of a rotating gantry with a fixed x-ray source and a detector [51]. Similar to CT, x-ray pass through the human body and form projections on the detector. These projections contain the information of x-ray attenuations by different body tissues and are reconstructed to obtain a 3D volume containing the spatial position of the attenuation coefficients of different tissues. Instead of a fan beam of x-ray, the source in CBCT generates an x-ray beam with a cone shape (see Figure 1.5). This allows the beam to cover a 3D volume of an object instead of a thin slice. Therefore, the machine only needs to rotate once to obtain all projections needed for a 3D ROI reconstruction. The number of projections of the ROI during the rotation ranges from 150 to more than 600 [51]. CBCT uses 2D detector arrays to collect the attenuated x-ray. The area of the detector limits the FOV for CBCT. The FOV of CBCT is larger than CT. This results in much more detection of scatter during imaging. The image quality of CBCT is degraded due to noise caused by scattering [51].

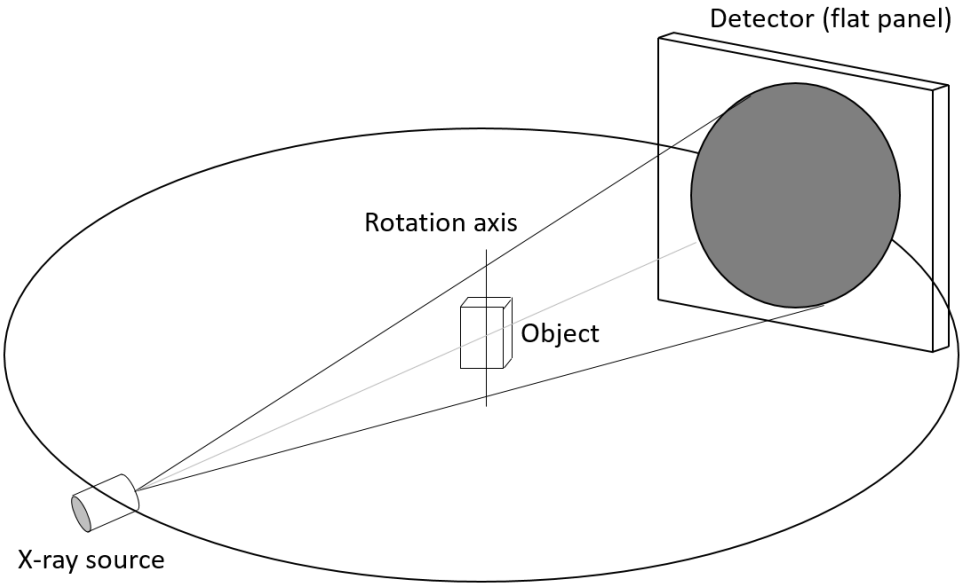



Figure 1.5 Demonstration of a CBCT scan. The x-ray source and the detector are rotated around the object to acquire the projections at different angles.

#### 1.5.5.3 MR

MR uses magnetic fields and radiofrequencies to reveal information about the anatomy and physiological processes. Unlike CT, CBCT, PET, and SPECT, MR scan does not involve x-rays or other types of ionizing radiation. MR imaging makes use of the magnetic properties of atomic nuclei inside the human body. Hydrogen nuclei, with only one proton, are abundant in the human body. They are most often used in MR imaging. The spinning of an atomic nucleus generates a magnetic moment. In absence of an external magnetic field, the magnetic moment of each atomic nucleus in the human body is randomly oriented, which results in no overall magnetic field. When a primary magnetic field is given, the nucleus precesses either parallel or antiparallel to the primary magnetic field. The frequency of precession of the magnetic moment of the nucleus around the primary magnetic field is defined as the Larmor frequency, which for protons equals 42.6 MHz/T. The portion of the parallel aligned nuclei (low energy state) is greater than that of the antiparallel aligned nuclei (high energy state) and the nuclei precess in different phases, which results in a net magnetic field in the same direction as the primary magnetic field. When a radiofrequency pulse with the Larmor frequency of the hydrogen nuclei is given, the nuclei are excited to change the energy state. Some low-energy nuclei flip the orientation of their magnetic moment, which decreases longitudinal magnetization. Meanwhile, the nuclei start precessing around the longitudinal axis (the orientation of the primary magnetic field) in the same phase. This results in a transverse magnetization component, which is perpendicular to the longitudinal axis. When stopping the radiofrequency pulse, the excited nuclei start returning to the equilibrium state and lose energy by emitting radio waves. This process is called relaxation. Most nuclei return to the orientation parallel to the longitudinal axis, which is named T1 relaxation. The nuclei precessing in the same phase begin dephasing in the transverse plane, which results in decreasing transverse magnetization. This process is called T2 relaxation.

The MR system consists of the primary magnet, radiofrequency coils, gradient coils, and the computer system. The primary magnet is used to produce the primary magnetic field. The radiofrequency coils transmit a bundle of different radiofrequency pulses for the excitation of nuclei and receive the radio waves emitted by the nuclei relaxation. The gradient coils generate the magnetic field in the x, y, and z-axis to alter the primary magnetic field. As a result, the magnetic field varies across different spatial locations associated with different lamor frequencies. The nuclei in a certain location can only be excited when the radiofrequency matches the lamor frequency at the location.

During imaging, the field gradients sweep across the FOV so that radio waves emitted from different regions can be detected. The computer system extracts the location information from the detected radio waves to reconstruct the spatial relation of the human body in a 3D image.

The repetition time (TR) refers to the time interval between successive radiofrequency pulses applied to the same slice. The echo time (TE) is the time interval between the delivery of the radiofrequency pulse and the peak of the signal induced in the coil. By adjusting TR and TE, different magnetic properties of tissue are detected by the MR system.

**T1-weighted MR**. T1-weighted MR images are generated by using short TR and TE. T1 relaxation properties predominately determine the brightness and contrast of the image. Fatty tissue appears with high intensity in T1-weighted MR. Tissue containing more water (e.g., tumor, cyst, edema, inflammation, and infection) have low intensity in T1-weighted MR. The T1-weighted MR can be used to detect abnormal liver parenchyma, but it is difficult to distinguish malignant liver tumors and other lesions in this modality.

**T2-weighted MR**. T2-weighted MR images are generated by using longer TR and TE. T2 relaxation properties predominately determine the brightness and contrast of the image. Tissue containing more water has a high intensity in this modality. Fatty tissue appears with low intensity in T2-weighted MR. When suppressing the fat signal, the contrast between lesions and liver parenchyma will increase. In this way, this modality can be used to detect tumors. Due to the high signal of water, cysts become very bright in T2-weighted MR.

**Diffusion-weighted MR**. Diffusion-weighted MR images are used to measure the random Brownian motion of water molecules in different tissues. Water molecules can flow with little restriction to various living tissues in the human body. In certain situations, such as pathological changes and cell membrane boundaries in the brain tissue, the movement of water molecules will be restricted. Diffusion-weighted MR can be used to evaluate movement restriction. When water movement becomes restricted, a high signal appears in this region. The movement of water molecules is restricted in tightly packed tumor cells, which can be detected in diffusion-weighted MR. It helps identify small lesions that are sometimes challenging for T2-weighted MR.

## 1.6 Image processing challenges in SIRT

As is shown in Equation 1.4, the computation of the absorbed dose requires the injected activity, the activity distribution, and the VOI. When performing dose calculation using the multi-compartment and voxel-level methods for dose planning or verification, the activity distribution inside a compartment or a voxel needs to be estimated or measured. Meanwhile, the segmentation of the liver and tumors are critical for computing the mean absorbed dose or obtaining the cDVH in each compartment.

Currently, the delineation of the liver and tumors is manually performed in the hospital due to a lack of automatic delineation methods. When performing the multi-compartment method, the liver and tumor volumes estimated at a coarse scale can be used for dose calculation. A coarse liver delineation can be performed by an expert highly skilled in this task within a few minutes. However, this tedious task still consumes time that can be spent on more valuable work. Tumor delineation requires much more expertise in radiology, oncology, and pathology than liver delineation. The task needs to be performed by more experienced radiologists, oncologists, or nuclear medicine physicians, who usually lack time for tedious and repetitive tasks in a daily clinical routine. Therefore, the acquisition of tumor delineations is much more difficult than that of liver delineations. For the voxel-level method, more precise liver and tumor delineations are required compared with the multi-compartment method. Manual delineation of higher precision costs much more time for the expert, which impedes the frequent use of voxel-level dosimetry in SIRT. Therefore, a (semi-)automated method for liver and tumor segmentation is necessary for precise dose planning and verification in SIRT.

The liver and tumor segmentations are mainly obtained from anatomical images like contrast-enhanced CT and MR or nuclear medicine images like [$^{18}$F]FDG or [$^{68}$Ga]Ga-DOTA-TATE PET. These images are performed for diagnosis. The estimation and measurement of $^{90}$Y activity distribution are obtained from $^{99m}$Tc-MAA SPECT and $^{90}$Y TOF PET. Dose calculation requires mapping the liver and tumor contours into the space of $^{99m}$Tc-MAA SPECT and $^{90}$Y TOF PET. To obtain the anatomic information during the $^{99m}$Tc-MAA SPECT scan or the [$^{18}$F]FDG or [$^{68}$Ga]Ga-DOTA-TATE PET scan, a non-contrast-enhanced CT scan is performed in the same imaging session. Similarly, a non-contrast-enhanced CT or MR scan is performed in the same imaging session of the $^{90}$Y TOF PET scan. It is assumed that the patient does not move during the imaging session. Therefore, the anatomic images can be automatically aligned to the nuclear medicine images without using an image registration method. The segmentation can be mapped to the $^{99m}$Tc-MAA SPECT and $^{90}$Y TOF PET by registering the anatomic images for diagnosis to the anatomic images accompanied with the nuclear medicine images. The registration of multi-modal images including contrast-enhanced CT, non-contrast-enhanced CT, and MR is very challenging. These multi-modal images are acquired at different time points with an interval from 1 day to 2 months between each other. The patient positioning might inherently vary in different imaging studies. The liver is a deformable organ, which exhibits various shapes in different body positioning. Besides, the liver shape and volume can have a large change due to tumor growth and proceeding treatment (resection, ablation, chemotherapy) over several months. All the above factors make liver registration of multi-modal images a very challenging topic.

Due to the lack of good registration methods, nuclear medicine physicians perform treatment planning and verification mainly based on $^{99m}$Tc-MAA SPECT and $^{90}$Y TOF PET, respectively. The liver and tumor segmentations were obtained by thresholding the SPECT or PET image. The threshold value is adjusted so that the liver or tumor segmentation has a volume close to the volume coarsely estimated on the diagnostic image. This might result in over-estimation of dose to tumors and under-estimation of dose to normal liver tissue when some high-activity regions in the healthy parenchyma are included in the thresholded tumor contours. The above under-estimation and over-estimation are risky for patients since an insufficient dose is delivered to tumors and an excessive dose is delivered to normal liver tissue. CT and MR images are used to reveal the anatomic information. They provide segmentations of the liver and tumors closer to the ground truth than the $^{99m}$Tc-MAA SPECT and $^{90}$Y PET which are used to show the activity distribution. A good registration allows the contours defined on the anatomic images to be combined with the activity information from the $^{99m}$Tc-MAA SPECT and $^{90}$Y PET. Therefore, a (semi-)automated liver registration method for multi-modal images is critical for precise dose planning and verification, which conforms to the actual dose delivery in the liver as much as possible.

Precise SIRT planning and verification requires (semi-)automated methods for segmentation of the liver and tumor and liver registration of multi-modal images. The above two challenges need to be solved by using cutting-edge image processing techniques, which are going to be discussed in Section 1.7.

## 1.7 Medical image processing methods

In the past years, many methods for image segmentation and registration were developed for general image processing, and not for medical imaging purposes alone. A brief overview of the methods relevant to the liver is provided and the feasible methods for solving image processing challenges in SIRT are discussed in this section.

### 1.7.1 Liver segmentation

Automatic liver segmentation is a challenging task due to the high variability of the liver shape and texture. The liver appears with a similar intensity to the surrounding organs and tissue in CT and MR

images. For patients with liver cancer, the tumour growth and the proceeding treatments create more variability of the liver shape and texture. To cope with this challenge, a number of advanced methods were developed by utilizing intensity, shape, and spatial information.

**Intensity-based methods**. The most intuitive method is to threshold the image by the liver intensity. This is only possible in CT, where the liver intensity has an expected range of HU. However, the liver has a HU intensity close to the intensity of the surrounding organs, like the heart, stomach, spleen, and kidney. It is very difficult to separate the liver from other organs by thresholding. Region growing is a common region-based segmentation method. It starts from several seed points, and gradually grows the segmentation to its neighbourhood based on similarity criteria (e.g., HU intensity). Region growing has a wide application to liver segmentation [52, 53, 54]. However, region-growing methods often fail when two organs with similar intensities are connected, even if the connection is very subtle. In this case, the liver segmentation will leak into the surrounding organs.

**Statistical shape models**. Statistical shape models were introduced by *Cootes* [55] to extract prior information about the shape. They generalize an average shape and the shape variation from many semantically similar objects. Learning the shape characteristics requires a large training dataset. The liver usually has high shape variation, especially for the pathological liver. It is difficult for statistical shape models to adapt to the abnormal liver shapes due to their lack of flexibility. Therefore, statistical shape models usually are followed by deformation methods to match the template shape to the abnormal ones [56, 57, 58].

**Probabilistic atlas methods**. Probabilistic atlas methods make use of the prior models of the shape and the spatial location for segmentation [59]. They usually start with the registration of training images to a template image. After that, the segmentations of these training images are warped onto the template segmentation. These segmentations are averaged voxel-by-voxel to generate a probabilistic atlas. For segmentation of a new image, the template image is mapped to the new image with a point-to-point spatial correspondence via a deformable registration. This mapping is then applied to the template segmentation so that the new image is segmented. Therefore, a good registration method is critical for the atlas-based segmentation. Probabilistic atlas methods are widely applied to liver segmentation [60, 61, 62, 63, 64].

**Level-set methods**. The basic idea of level-set methods is to represent a curve or a surface as the zero level set of a higher dimensional hyper-surface. The curve changes as the hyper-surface changes with time. Take a problem of 2D segmentation as an example. The hyper-surface is two-dimensional and time-variant, represented by $z = \Phi(x, y, t)$. The segmentation is regarded as a 2D curve, which is a cross section between the hyper-surface and the plane $z = 0$. Therefore, the curve is represented by $\Phi(x, y, t) = 0$, which is the zero level set. In this way, the evolution of the curve is transformed into the evolution of a 3D level set function, $\Phi(x, y, t) = 0$ [65]. The evolution over time is controlled by using a speed function, which specifies the speed at the direction normal to the surface. *Pan* and *Dawant* proposed a liver-specific speed function, which can stop the contour at weak boundaries [66]. Wang *et al* obtained initial liver segmentations by probabilistic atlas methods and then used level-set methods for final liver segmentations [67].

**Graph cut methods**. Graph cut was proposed by Boykov *et al* [68] to solve the segmentation problem as an alternative to boundary-based segmentation methods. A graph consists of vertices (also called nodes) connected by edges. When graph cut is applied to image segmentation, the image pixels are regarded as nodes connected with weighted edges representing the adjacency between the pixels. In graph theory, the cut represents a partition of the vertices into two disjoint subsets. The image is partitioned into the object and the background by minimizing a cost function between all possible cuts. It is a semi-automatic method that needs the user to specify the seeds in the background and the object to be segmented. Therefore, graph cut methods require a lot of manual interactions for seed point selection [69].

**Convolutional neural network (CNN)**. CNN is a type of deep learning method widely applied to image analysis challenges, such as image segmentation, image classification, image recognition, etc. CNN is inspired by the connected patterns of neurons for information processing in the brain. It consists of stacked layers with multiple elements in each layer to mimic the neurons. CNN is featured by using convolutions to connect the elements in the current layer with the elements in the previous layer. The information from the image is passed on layer by layer to extract high-level features for segmentation or classification. CNN has the advantage of full automation, little pre-processing, little requirement of prior knowledge, high-level feature extraction without manual design, and good extrapolation performance compared with the classical methods. As a supervised learning method, CNN needs to learn from a large amount of training data so that it can extrapolate the segmentation or classification of new data with good performance. In the past years, different CNN models have been increasingly applied to medical image segmentation problems, like liver or kidney [70, 71]. Many grand challenges were organized by the scientific and technical communities to evaluate the methods with the best performance for liver segmentation of CT or MR. These challenges include the Segmentation of the Liver Competition 2007 (SLIVER07) for liver segmentation of CT, the Liver Tumor Segmentation Challenge 2017 (LiTS17) for liver and tumor segmentation of CT, Combined (CT-MR) Healthy Abdominal Organ Segmentation (CHAOS) for CT and MR segmentation of the liver, kidneys, and spleen. These challenges allowed the participants to train and test their algorithms on the same data so that the only factor affecting the segmentation result was the algorithm. The results for the test data were evaluated by the specified metrics and the algorithms were ranked according to these metrics. The LiTS17 challenge contains 131 training CT images and 70 test CT images. The best algorithm achieved a dice similarity coefficient (DSC)[1] of 0.96 [72]. The best performing algorithms in the LiTS17 challenge used CNN-based approaches. Most CNN approaches adopted a U-net structure [73] and used transaxial slices (2D) or a stack of a few transaxial slices (2.5D) as the input to reduce the training time and the computation resource [72]. Chelbus *et al* implemented three orthogonal 2D U-net models using the axial, coronal, and sagittal slices from the MR images of SIRT patients [74]. They reported liver segmentation results with a mean DSC of 0.95 [74].

### 1.7.2 Liver tumor segmentation

Compared with liver segmentation, liver tumor segmentation is more challenging due to the large variation in the shape, size, contrast, texture, and position of tumors. The liver is usually located in the upper right quarter of the abdomen under the diaphragm. Liver tumors can occur in multiple spots of the liver. There is no prior knowledge of the tumor shape and intensity. Different types of liver tumors can appear in diverse shapes and sizes. A large HCC can be highly diffuse and heterogeneous, which makes tumor delineation very challenging for the expert. Contrast agents can also influence the tumor's appearance. The tumor is hyper-intense in the arterial phase and hypo-intense in the wash-out phase. And the tumor boundary becomes unclear when the contrast is relatively low. Besides, the same tumor can appear different in CT and MR images, since each modality reveals a different physiological property of the tumor. It is difficult to find a uniform standard for tumor delineation. Therefore, liver tumor segmentation requires more advanced methods which can exploit tumour features in a higher dimension.

**Thresholding methods**. Like liver segmentation, thresholding is the most intuitive method for liver tumor segmentation. It makes use of the intensity difference between tumors and normal tissue. Park *et al* proposed a method to find the statistical optimal threshold for liver tumor segmentation [75]. Nugroho *et al* used histogram manipulation to find a threshold that maximizes the variance between classes [76]. Choudhary *et al* segmented tumors using a minimum cross-entropy multi-thresholding

[1] DSC measures the volume similarity between two segmentations. The value of DSC ranges from 0 to 1. The two segmentations are completely overlapped when the DSC is 1 and have no overlap when the DSC is 0.

algorithm [77]. The tumor segmentation was further refined by region-growing and level-set methods [77]. Moltz *et al* developed a hybrid method that combines adaptive thresholding based on grey values with model-based morphological analysis [78]. Abdel-massieh *et al* filtered the images by Gaussian smoothing and segmented tumors by an Isodata threshold approach [79].

**Region growing methods**. Region growing is often used for tumor segmentation since it is easy to implement. It is a semi-automatic method, which usually requires manual seed selection. Many variations of region growing in 2D or 3D images have been developed for liver tumor segmentation. Wong *et al* proposed a 2D region growing method with knowledge-based constraints [80] for liver tumor segmentation. The constraints are used to ensure an acceptable size and shape of the segmentation [80]. Qi *et al* modelled the probability density function (PDF) of the tumor intensity as a bag of Gaussians and then employed a 3D seeded region growing method [81]. The Gaussians are initialized by manually selected seeds and updated iteratively during the segmentation growing [81].

**Level-set methods**. Level-set methods can also be applied to liver tumor segmentation. Smeets *et al* proposed a level-set method by using a speed function derived from a statistical pixel classification algorithm [82]. Jimenez *et al* implemented a multi-resolution level set method with local curvature constraints for liver and tumor segmentation [83]. They used a curvature restriction to limit the contour expansion and leaking to other organs. The curvature restriction was relaxed as the resolution increased. Choudhary *et al* implemented a minimum cross-entropy multi-thresholding algorithm to segment tumors and smoothened the segmentation by morphological closing and the level-set method [84].

**Convolutional neural network (CNN) for CT**. CNN has the advantage of learning high-level features of tumors with various shapes, sizes, and textures automatically when given enough training data with good tumor delineations. The LiTS17 challenge provided 131 training CT images and 70 test CT images with good liver and tumor delineation. This allows different algorithms for liver tumor segmentation to be compared on the same data. The highest DSC achieved in the challenge was 0.70 for liver tumor segmentation [72]. The best performing methods adopted cascaded U-net approaches and used a stack of 2D slices as the input [72]. They used a first U-net model to output the liver segmentation and then imported the image and the liver segmentation to the second U-net model for tumor segmentation. The cascaded structure can help the CNN focus on the tissue inside the liver. Some pre-processing approaches such as HU-value clipping are used to help the CNN concentrate on the soft tissue. And some post-processing approaches are used to refine the liver segmentation from the CNN. The LiTS17 data are widely used in other publications. Both Bi *et al* [85] and Xi *et al* [86] adopted the cascaded structure and added the residual connection to the CNN. Residual connection means that the input of the previous layer is directly added to the output of the following layer. It can facilitate the flow of information between layers to help the optimization process of the CNN to converge better. Chlebus *et al* implemented a U-net structure with residual connections for liver tumor segmentation and post-processed the tumor segmentation using a model trained by hand-crafted features to reduce the false positives [87].

**Convolutional neural network (CNN) for MR**. Liver tumor segmentation for MR is a more challenging task than that for CT due to the higher intensity variation and heterogeneity of tumor appearance and image artefacts (e.g., motion) for MR. Currently, there are no publicly available MR images from liver cancer patients, which makes it difficult to compare the performance of different segmentation algorithms and create a benchmark for liver tumor segmentation of MR. Jansen *et al* proposed a fully convolutional network by using a set of dynamic contrast-enhanced MR and diffusion-weighted MR registered to each other for liver metastases detection [88]. They reported a sensitivity of 99.8% for metastases detection [88]. Hatamizadeh *et al* proposed a U-net structure with a post-processing framework using a level-set based active contour method for CT and MR segmentation of lesions in different organs including the liver [89].

### 1.7.3 Registration of multi-modal images

Good co-registration of multi-modal images is critical for integrating the information of VOIs and activity distribution. CT and MR are the two main modalities for anatomical information, which is crucial for the registration of images scanned at different time points and under different imaging conditions. In the SIRT workflow, CT and MR are acquired either from the diagnostic imaging procedure or from the $^{99m}$Tc-MAA and $^{90}$Y imaging procedures. The physical principles of CT and MR imaging are very different, which result in completely different intensities and landmark appearances. It makes the registration of these two modalities very challenging.

**Liver-surface-based and vessel-based methods**. The liver surface is an intuitive feature to match the livers of different images. And most tumors are located near vessels to obtain blood supply, which can facilitate the registration of tumors in CT and MR. Therefore, the liver surface and vessels are often used independently or combined for liver registration. Lee *et al* used a multi-scale surface fitting method based on the head-and-hat algorithm [90]. The head-and-hat algorithm uses two surfaces. The head surface remains stationery and the hat surface can be transformed. The registration is performed by iteratively transforming the hat surface so that the closest fit of the hat surface onto the head surface is found. Foruzan *et al* proposed a hybrid registration method for CT and MR registration [91]. It started with a rigid transformation using maximum intensity projections and principal component analysis to align CT and MR at a global scale. After that, a point-based non-rigid algorithm was used to deal with local liver deformations by using liver surface points and the PV branching points as landmarks. Cazoulat *et al* proposed a deformable registration method for CT images by using a biomechanical liver model driven by boundary conditions on the liver surface and the segmentation of the vasculature [92]. Charnoz *et al* developed a deformable method for CT registration by using a tree matching approach for vessels [93]. The segmentation of the vascular systems was modelled as trees and the common bifurcations and vessels from CT images were matched [93].

**Intensity-based methods**. The intensity is the most basic feature for image registration. By measuring intensity similarity, the corresponding regions in two images can be matched with each other. Mutual information (MI) is the most common intensity similarity metric used for the intensity-based registration. MI reaches its maximum value when one uniform region is matched to another one. Gunay *et al* proposed an intensity-based deformable registration method using MI as the similarity metric, extended with a point-to-surface penalty [94]. Manual point annotation on the liver surface was performed in the regions where the intensity-based registration is inaccurate. These points help improve the registration accuracy by using the point-to-surface penalty. Huang *et al* implemented an intensity-based rigid registration for CT and MR [95]. It used MI as the similarity measure and maximized MI via an exhaustive search over the entire image by using Powell's direction method [95]. De Moor *et al* proposed a non-rigid registration method by using either MI or the sum of squared difference (SSQD) as the similarity measure [96]. The voxel deformation was regularized by a spring model, where each pair of neighbouring voxels are assumed to be connected by a spring. The registration rigidity was controlled by the spring strength. Besides the image intensities, the image gradients derived from the intensities are also used to measure image similarity. Spahr *et al* implemented a deformable registration by measuring the image gradient similarity and prevented unrealistic deformation by a curvature regularizer [97].

**Segmentation-based methods**. Image registration using an intensity similarity measure at a global scale might ignore the registration accuracy of VOIs, which are critical for SIRT. VOI segmentations can be used to guide the registration so that the VOIs in two images are well matched. Rieder *et al* proposed a workflow by masking the image with the liver and lesion segmentation and registering the masked images rigidly [98]. The masked images excluded lesions and the surrounding organs to prevent the matching of the lesion shapes instead of the surrounding liver anatomy [98]. The liver segmentation was coarsely segmented by using a histogram analysis and the lesion segmentation was obtained semi-automatically by using region growing. Hasenstab *et al* masked the MR images by using CNN liver

segmentations and then performed affine registration of the masked images [99]. The optimized affine parameters were used to map the MR images.

# 2. Objectives of the research

The objective of this research project is to use advanced image analysis techniques to assist nuclear medicine physicians to perform more precise treatment planning and verification of SIRT. The ultimate goal is to achieve a voxel-level dosimetry for SIRT planning and verification, which is automated as much as possible. However, this goal is difficult to be achieved at the current stage due to the complexity of automatic lesion segmentation and registration of multi-modal images (mainly CT and MR). Therefore, (semi-)automatic approaches for segmentation of the liver and lesions and registration of multi-modal images are to be developed, which minimize the need for manual interactions by the physician. To achieve it, the following research tasks need to be fulfilled:

(1) (semi-)automatic liver segmentation of CT and MR images,

(2) co-registration of multi-modality images (CT and MR),

(3) (semi-)automatic liver lesion segmentation of CT and MR images.

The above tasks are required to achieve multi-modal image analysis for SIRT without introducing much extra manual interaction and workload. They can facilitate the automation and standardization of the SIRT workflow.

The methods for (semi-)automatic liver segmentation for CT and MR images are described in chapter 3 and 4. Chapter 5 and 6 demonstrate the methods for registration of multi-modal images. The methods for (semi-automatic) liver lesion segmentation are introduced in chapter 7. In chapter 8, the clinical workflows for SIRT planning and verification are presented. A discussion of the research project and future perspectives is presented in chapter 9 and a summary of the project is provided in chapter 10.

# 3. Whole liver segmentation based on deep learning and manual adjustment for clinical use in SIRT

This chapter has been published as:

X. Tang, E. Jafargholi Rangraz, W. Coudyzer *et al.* "Whole liver segmentation based on deep learning and manual adjustment for clinical use in SIRT," *Eur J Nucl Med Mol Imaging*, vol. 47, pp. 2742–2752, 2020, Springer Nature. https://doi.org/10.1007/s00259-020-04800-3.

## 3.1 Abstract

**Purpose:** In selective internal radiation therapy (SIRT), an accurate total liver segmentation is required for activity prescription and absorbed dose calculation. Our goal was to investigate the feasibility of using automatic liver segmentation based on a convolutional neural network (CNN) for CT imaging in SIRT, and the ability of CNN to reduce inter-observer variability of the segmentation. **Methods:** A multi-scale CNN was modified for liver segmentation for SIRT patients. The CNN model was trained with 139 datasets from three liver segmentation challenges and 12 SIRT patient datasets from our hospital. Validation was performed on 13 SIRT datasets and 12 challenge datasets. The model was tested on 40 SIRT datasets. One expert manually delineated the livers and adjusted the liver segmentations from CNN for 40 test SIRT datasets. Another expert performed the same tasks for 20 datasets randomly selected from the 40 SIRT datasets. The CNN segmentations were compared with the manual and adjusted segmentations from the experts. The difference between the manual segmentations was compared with the difference between the adjusted segmentations to investigate the inter-observer variability. Segmentation difference was evaluated through Dice Similarity Coefficient (DSC), Volume Ratio (RV), Mean Surface Distance (MSD), and Hausdorff Distance (HD). **Results:** The CNN segmentation achieved a median DSC of 0.94 with the manual segmentation and of 0.98 with the manually corrected CNN segmentation, respectively. The DSC between the adjusted segmentations is 0.98, which is 0.04 higher than the DSC between the manual segmentations. **Conclusion:** The CNN model achieved good liver segmentations on CT images of good image quality, with relatively normal liver shapes and low tumor burden. 87.5% of the 40 CNN segmentations only needed slight adjustments for clinical use. However, the trained model failed on SIRT data with low dose or contrast, lesions with large density difference from their surroundings, and abnormal liver position and shape. The abovementioned scenarios were not adequately represented in the training data. Despite this limitation, the current CNN is already a useful clinical tool which improves inter-observer agreement and therefore contributes to the standardisation of the dosimetry. A further improvement is expected when the CNN will be trained with more data from SIRT patients.

**Key Words:** selective internal radionuclide therapy (SIRT); liver segmentation; convolutional neural network (CNN); internal dosimetry; inter-observer variability

## 3.2 Introduction

Selective internal radionuclide therapy (SIRT) or radioembolization aims at treating surgically non-resectable primary or metastatic liver tumors. In SIRT, yttrium-90 ($^{90}$Y) microspheres are injected into the hepatic artery [50], which is the predominant vessel for blood supply to liver tumors [100]. By selecting the appropriate branch of the hepatic artery, the radioactivity can be selectively administered to the targeted tumors, which results in high dose tumoral irradiation, while keeping the dose to the healthy liver below the tolerance level [101, 102, 103]. In the pre-treatment study, macro-aggregated albumin particles labeled with technetium-99m ($^{99m}$Tc-MAA) are injected and whole-body planar

imaging and SPECT/CT are performed within the hour after injection [104] to estimate the lung shunt fraction and predict the intra-hepatic distribution of $^{90}$Y microspheres inside the liver. Following the injection of $^{90}$Y microspheres, a post-treatment study is performed to obtain $^{90}$Y-PET images (PET/MR or PET/CT). The actual distribution of $^{90}$Y microspheres can be determined by these images [105, 106] to verify the treatment.

In both pre- and post-treatment studies, the absorbed dose to the tumor(s) and the normal liver parenchyma is estimated or measured to predict or verify the treatment's result on the tumor and on the healthy liver [107]. Several approaches are being used for this purpose, including mono-compartment [45] and multi-compartment [108] methods and voxel-based approaches [45]. For all these methods, accurate contours of liver and tumors are required for dosimetric analysis. Therefore, good liver and tumor delineation plays an important role in these dosimetric activity prescription methods [109].

Manual liver segmentation is tedious, time-consuming and suffers from inter-observer variability [74]. This segmentation variability may increase the variability of the absorbed dose computed by dosimetric methods. In recent years, CNNs have been increasingly used in the medical field for segmenting different organs, such as liver or kidney [70, 71]. Many grand challenges for automatic liver segmentation, organized by the scientific and technical communities, made available CT or MRI datasets with reference liver delineations on the internet. The participants trained their algorithms on the training datasets and tested them on the test datasets. The test results were evaluated through several metrics and the final ranking of the algorithms was published. The Liver Tumor Segmentation Challenge (LiTS) held in 2017 involves the tasks of liver and tumor segmentation on CT data. The best algorithm, trained on 131 datasets and tested on 70, achieved a dice similarity coefficient (DSC) of 0.96 [72]. Most algorithms adopted U-net derived architectures and used 2D or so-called 2.5D images due to the long training time and high resource requirements [72]. Chelbus *et al*. implemented three orthogonal 2D U-net CNN models trained with axial, coronal, and sagittal image patches from the MR data of SIRT patients [74]. They reported good liver segmentation results with a mean DSC of 0.95. According to their experimental results, manual correction of the CNN segmentation resulted in a much lower inter-observer variability than manual routine segmentations [74].

The aim of our study is to develop an automatic method of liver segmentation on CTs for SIRT patients, including both contrast-enhanced CTs and non-contrast-enhanced CTs, the latter obtained during PET/CT. This method should be generic, robust, and applicable to CT images with various contrast and irregular liver shapes. For this purpose, we modified a 3D CNN structure named DeepMedic [110] for the task of liver segmentation. Our hypotheses were that the liver segmentations from the CNN can be good enough for clinical application with limited adjustments and will reduce the inter-observer variability of liver segmentation for CT. SIRT patients usually underwent a variety of preceding treatments (e.g., liver resection, chemotherapy, tumor ablation) and have abnormal liver shapes and high disease burden. Therefore, applying the CNN model to liver segmentation of CTs for these patients is more challenging than for those data from public challenges. Besides, automatic liver segmentation using the CNN has the potential of speeding up the segmentation process by minimizing manual interaction from the medical doctors and technologists and standardizing the clinical workflow. Its practical value remains to be investigated in a more clinical context.

## 3.3 Materials and methods

### 3.3.1 Data

In our study, most training datasets were from several international challenges of liver and tumor segmentation for CT, including SLIVER07 challenge (20 datasets), Liver Tumor Segmentation (LiTS17) challenge (131 datasets), and Medical Segmentation Decathlon (MSD) challenge (131 datasets). In the LiTS17 challenge, 20 datasets were discarded due to errors in their image headers. The

image in-plane resolution of the challenge datasets ranged from 0.56 to 1.0 mm and the slice thickness ranged from 0.7 to 5.0 mm. During the process of our research, 65 CT datasets from the SIRT patients with liver delineation were collected in our hospital. Their image in-plane resolution ranged from 0.65 to 1.37 mm and their slice thickness ranged from 1.0 to 5.0 mm.

When carrying out the experiments, the datasets from the MSD challenge were found to be identical to those from the LiTS17 challenge, which as far as we could see is not clearly indicated on the respective websites. As a result, 91 of these datasets from the two challenges had twice the weight of the other ones during training. Since the same network trained with uniform weights did not perform better (the difference was within the variation caused by the random CNN initialization), we continued using the original model.

The number of the training, validation, and test data is shown in Table 3.1. The characteristics of the 38 SIRT patients from whom the 40 test SIRT datasets come and the 25 SIRT patients for training and validation are presented in Table 3.2. The challenge datasets for training and validation are anonymous and publicly available. Their patient characteristics are not available.

Table 3.1 Number of the training, validation, and test datasets

| | Training | Validation | Test | Total |
|---|---|---|---|---|
| LiTS17/MSD | 121 | 10 | 0 | 131 |
| SLIVER07 | 18 | 2 | 0 | 20 |
| SIRT data | 12 | 13 | 40 | 65 |
| Total | 151 | 25 | 40 | 216 |

Table 3.2 Characteristics of the SIRT patients for training, validation, and test

| Characteristics | Training | Validation | Test |
|---|---|---|---|
| No. of SIRT patients | 12 | 13 | 38 |
| Age (y), median [range] | 58 [32,72] | 61 [42,70] | 67 [25,87] |
| Sex (female/male) | 5/7 | 3/10 | 13/25 |
| Weight (kg), median [range] | 76 [61,119] | 82 [58,125] | 74 [46,129] |

Two clinical experts manually segmented the livers in CT images from the test set. In addition, they also performed manual corrections to the segmentations produced by the CNN. All SIRT datasets for this research were evaluated at the KU Leuven after approval by the Ethics Committee Research of UZ / KU Leuven.

### 3.3.2 CNN development

The CNN model used in the paper is a modified version of the dual pathway, 11-layer deep, three-dimensional structure (named DeepMedic) designed for the task of brain lesion segmentation [110]. The network adopts a hybrid scheme between the common patch-wise training (the CNN model only

predicts the central voxel of the input image patch) and the so-called dense training on the full image (the network outputs the prediction for all voxels in the input image) [111]. If the input of CNN is the whole 3D image, the dense training setting is mainly constrained by the limited GPU memory. In the patch-wise setting, the same voxels in the overlapping patches are repeatedly involved in the convolutional computations for the prediction of different central voxels, which is inefficient in making full use of the computational power and memory of GPU. The DeepMedic structure overcomes these problems by using image segments with a size larger than the receptive field as the CNN input. This scheme enables the network to output the prediction for multiple voxels in the image segment in one forward pass.

Furthermore, the DeepMedic network introduces the multi-scale processing technique by using parallel convolutional pathways at different resolutions. The contextual information inside the CNN's receptive field plays an important role in the CNN inference. The more spatial context is incorporated in the inference process, the more comprehensive understanding of the detected object the network can obtain. However, more incorporated contextual information means increasing computation and memory demands if the images with the normal resolution are used. The DeepMedic structure employs a clever way to incorporate both the local and global contexts by adding a low-resolution pathway operating on down-sampled images. In this way, the receptive field of the low-resolution pathway is enlarged greatly at the cost of resolution. But this cost can be compensated by combining the low- and high-resolution pathways, since the local information is preserved in the high-resolution pathway.

**Architecture.** In the modified CNN structure, a third pathway with lower resolution than the second is added (see Figure 3.1). Considering that the liver is much larger than a brain lesion, this third pathway is introduced to help the CNN learn the context information from the whole abdominal region, which is essential for reducing errors. The down-sample rates of the three pathways are 1, 5, and 15, respectively. The kernel size used by the convolutional layers in the three pathways are $3 \times 3 \times 3$. To give more weight on the context information from the second and the third pathways, the number of features is increased in the deeper layers. The outputs of the second and third pathways are up-sampled by 5 and 15, respectively. Then the outputs of the three pathways are treated as three features which are combined by the next two layers with a $1 \times 1 \times 1$ convolutional kernel. Through one classification layer, the CNN outputs the probability map, where each voxel represents its probability of belonging to the liver.

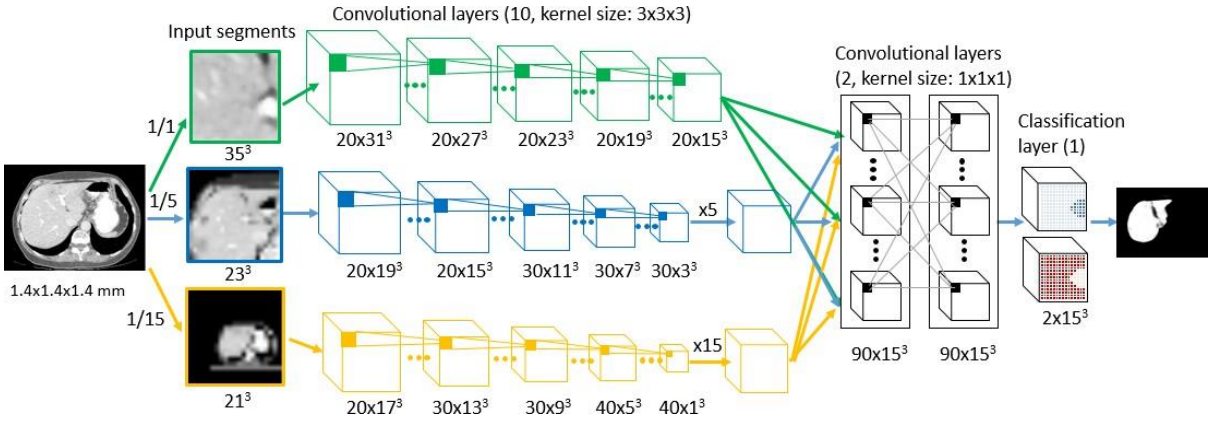



Figure 3.1 Overview of the modified DeepMedic structure. The model consists of three pathways (10 convolutional layers in each pathway) followed by a common pathway with two fully connected layers and one classification layer. The input image has the voxel size of $1.4 \times 1.4 \times 1.4$ mm$^3$. The input image segments of three pathways are randomly sampled from the images down-sampled by 1, 5, and 15 during the training process. The output segment has the size of $15 \times 15 \times 15$.

**Training**. The CNN model was trained on 3D image segments randomly sampled from the 3D image with a batch size of 16. The model used binary cross entropy as the loss function with the stochastic gradient descent optimizer. The initial learning rate is 0.007 and decreased every 32 epochs. The model

quality was evaluated every 8 epochs on the full segmentation of the validation set using the DSC. The training process took 26.75 hours using a GPU of NVIDIA P100 with 16 GB DRAM. The time for the CNN prediction of 40 test SIRT datasets ranged from 11 to 55 seconds using the GPU. When using a CPU of Intel Xeon E5-2699, the time for the CNN prediction of 40 test SIRT datasets was between 3 and 13 minutes.

**Data preprocessing.** The 3D CT images were cropped so that they included the whole abdominal region in each transaxial slice. In an earlier version of the network, we used images containing only the liver. However, we found that, when the images were enlarged such as to contain the entire abdomen in every transaxial slice, the liver segmentation performance of the CNN increased substantially. The cropped images were resampled to 1.4 mm isotropic voxel size so that the CNN could learn about the size of the liver and the surrounding organs. After that, the resampled images were median filtered and normalized by a linear mapping of the Hounsfield units (HU) of the CT images between -200 and 200 to the range of [-0.2, 0.2].

**Data augmentation**. The voxel intensities in the lower-contrast CT images from SIRT patients are often lower than in the contrast enhanced CT images from the challenges. To ensure the robustness of the CNN model to variations in the amount of contrast enhancement, a random intensity shift was applied to modify (and usually decrease) the intensity of the training images. This was done by adding a single random value, drawn from a Gaussian distribution with a mean of -40 HU and a standard deviation of 40 HU, to all the voxel values of a particular training image. Additionally, a random flipping with probability of 0.5 along the x and y axis and random elastic deformations were applied.

**Data postprocessing.** The output of the CNN model was a probability map. It was transformed into a binary mask of liver with the threshold of 0.5. To verify our threshold choice, a simple experiment was done to find an optimal threshold which maximizes the DSCs of the training datasets. After that, the optimal threshold of 0.32 was applied to the validation datasets. The median DSCs of the challenge datasets for validation were around 0.97 for both thresholds and the median DSC of the SIRT datasets for validation using the threshold of 0.32 was 0.6% higher than using the threshold of 0.5. Because the improvement using the optimal threshold is small and the network output is supposed to be a probability map, we prefer to use a threshold of 0.5, which is the natural choice because it selects the voxels which are more likely to belong to the liver than not. The binary mask was eroded to disconnect the regions with weak connection. Then the largest connected region in the binary mask was selected while other small islands were not included in the liver volume of interest. The largest connected region was dilated back to its original size and then was taken as the final result of liver segmentation.

### 3.3.3 Experiments

**Comparison between the CNN segmentation, manual segmentation, and adjusted segmentation.** To evaluate the liver segmentation quality of our CNN model, an experienced radiographer (WC) was asked to delineate the liver segmentation for 40 test datasets from SIRT patients in our hospital, with his choice of appropriate software available to them in the clinic at the time. These segmentations were performed semi-automatically using Siemens Syngo MMWP Volume software (Siemens Healthcare, Erlangen, Germany). After that, the CNN segmentation and the manual segmentation were compared with each other through several metrics. To analyze the errors of the CNN segmentation and its possibility of being used in clinical application, the expert was also asked to adjust the liver segmentation from CNN for all 40 test SIRT datasets. The adjustment was done using MIM software (MIM Software Inc., Cleveland, OH). When the expert did the adjustment, he was asked to score the CNN segmentation from 1 to 5 with a minimum interval of 0.5. The criteria used by the first expert for scoring are listed in Table 3.3. By comparing the CNN segmentation and adjusted segmentation, the areas where the errors of CNN segmentation often occur can be found, which is helpful for the further improvement of the CNN model and the selection of training datasets.

Table 3.3 Criteria for scoring the liver segmentation from CNN (experts 1 & 2)

| | Criteria (expert 1) | Criteria (expert 2) |
|---|---|---|
| 1, 1.5 | The CNN segmentation is very bad. A very large amount of adjustment is needed. It is better to segment the liver manually from scratch. | Major corrections are needed. Starting from the CNN segmentation is counterproductive and perceived more time consuming than manual segmentation. |
| 2 | | Major corrections are needed. Using the CNN seems to result in no benefit. |
| 2.5 | | Major corrections are needed. The CNN segmentation seems to have minor benefit. |
| 3, 3.5 | The CNN segmentation needs limited adjustment and can be used for the clinical application after adjustment. | The CNN segmentation requires moderate corrections, but is deemed a good starting point. Manual adjustment is perceived clearly faster than manual segmentation. |
| 4 | The CNN segmentation needs slight or no adjustment and is ready for the clinical use. | Minor corrections (more than details or glitches) are required, which has limited impact on the volume and consumes substantially less time than manual segmentation. |
| 4.5 | | Minor glitches are changed, which has no expected impact. |
| 5 | | No corrections are made. The CNN segmentation is ready for clinical use. |

**Inter-observer variability.** To evaluate the influence of the CNN segmentation on the inter-observer variability of liver segmentation, a nuclear medicine physician (CMD) with over 10-year experience in SIRT also provided manual liver segmentations and manual adjustments to the segmentations from the CNN. To shorten the processing time and reduce the expert's workload, 20 datasets were randomly selected from 40 test SIRT datasets for de novo segmentation and adjustment. For both tasks, MIM software was used. Out of 40 test SIRT datasets, there were 2 SIRT datasets where the CNN model had a very poor liver segmentation (several large parts of the liver were missing). These two segmentations were excluded intentionally when picking out the 20 datasets. After that, the difference between the 20 manual segmentations from two experts were compared with the difference between their adjusted segmentations through several metrics. The criteria used by the second expert for scoring the CNN segmentation are similar to the criteria used by the first expert but more detailed for each single score (see Table 3.3).

**Analysis of manual adjustment.** The adjusted segmentations from the two experts were compared with the CNN segmentations for the 20 test SIRT datasets through visual inspection to look into the regions most frequently corrected by the experts.

### 3.3.4 Evaluation metrics

In our experiment, the difference between segmentations was measured through several metrics calculated in 3D, including Dice Similarity Coefficient (DSC), Volume Ratio (RV), Mean Surface Distance (MSD), and Hausdorff Distance (HD).

**Dice Similarity Coefficient.** DSC is used to measure the volume-based similarity between two segmentations [112]. The more overlap the two segmentations have, the larger DSC is. The value of DSC is always between 0 and 1.

**Volume Ratio.** RV computes the ratio of the liver volumes from two segmentations, defined as: $RV(seg_1, seg_2) = V_1/V_2$, where $V_1$ and $V_2$ are the volumes of two segmentations.

**Mean Surface Distance and Hausdorff Distance.** MSD and HD are designed to measure the surface-based difference between two segmentations [113]. MSD computes the average distance between the two segmentation surfaces, whereas HD computes the largest distance between them.

## 3.4 Results

### 3.4.1 Comparison between the CNN segmentation, manual segmentation, and adjusted segmentation

The median DSC, RV, MSD, and HD between the CNN segmentation and manual segmentation were 0.94, 0.93, 2.1 mm, and 29.2 mm (see Figure 3.2). The median DSC, RV, MSD, and HD between the CNN segmentation and adjusted segmentation were 0.98, 0.98, 1.0 mm, and 30.1 mm (see Figure 3.2). The median DSC, RV, MSD, and HD between the manual segmentation and adjusted segmentation were 0.95, 1.04, 1.7 mm, and 23.5 mm (see Figure 3.2). From the results of DSC and RV, it is evident that the liver volume from the adjusted segmentation agrees more with the liver volume from the CNN than that from the manual segmentation. According to Figure 3.2(c), the liver surfaces from most adjusted segmentations are more similar to the liver surfaces from the CNN than those from the manual segmentations. The Hausdorff distance between the CNN segmentation and adjusted segmentation is slightly larger than that between the manual segmentation and adjusted segmentation in Figure 3.2(d). This is explainable because the CNN model has some errors in its liver segmentations, due to the inclusion of other tissues or to the exclusion of some parts of the liver.

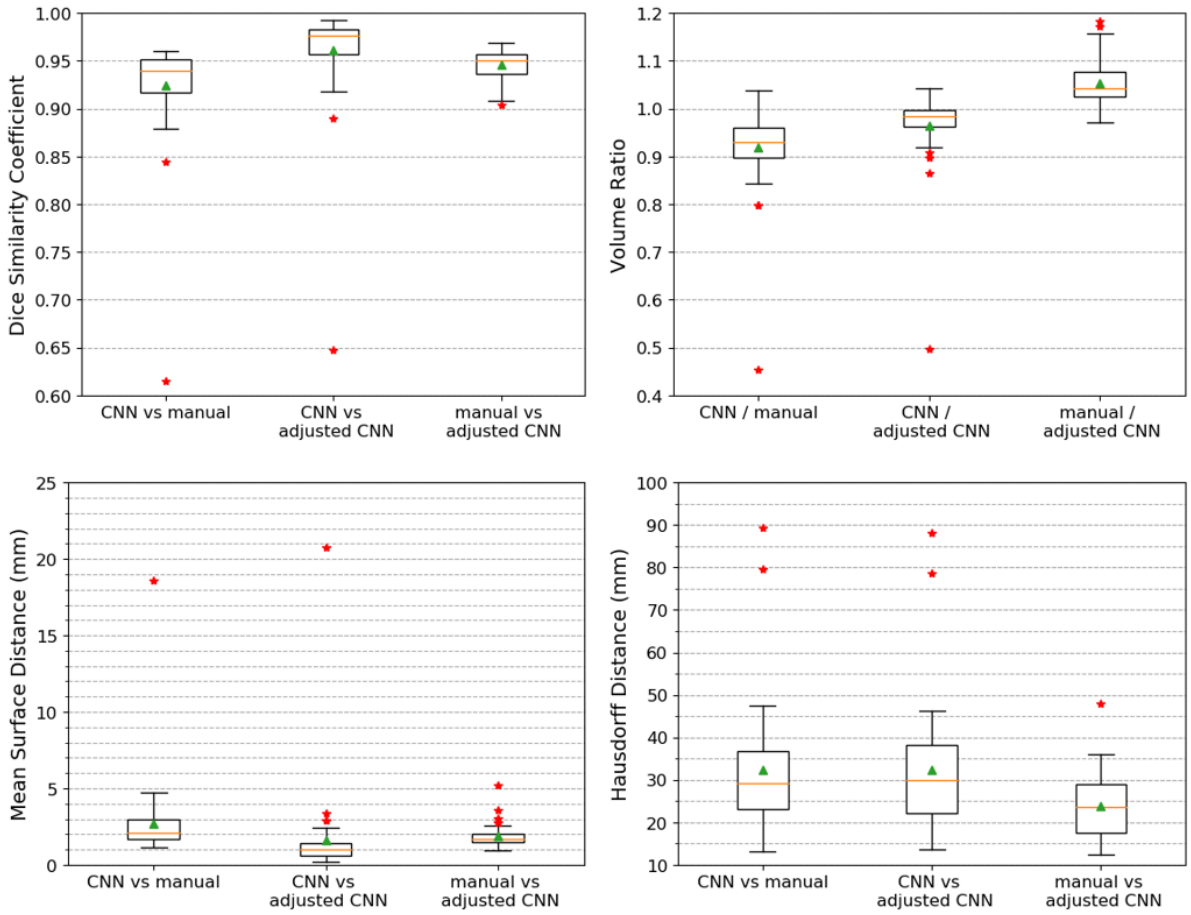



Figure 3.2 Comparison between the CNN segmentation (CNN), manual segmentation (manual), and adjusted segmentation (adjusted CNN) for 40 test SIRT datasets using (top left) dice similarity coefficient, (top right) volume ratio, (bottom left) mean surface distance, and (bottom right) Hausdorff distance.

The scores assigned to the 40 CNN segmentations are presented in Table 3.4. According to the scores given by the first expert, 40% (16/40) of liver segmentations from the CNN are very good and can be used for clinical application with slight or no adjustment from the expert. The CNN segmentations of 47.5% (19/40) SIRT datasets require limited adjustment and are then ready for clinical use. There are 12.5% (5/40) poor liver segmentations from the CNN which should not be applied in clinical use. The scores from the expert verify that 87.5% (35/40) of liver segmentations from CNN are good enough for clinical use with some additional adjustment. Some examples of the liver segmentations from CNN with different scores are presented in Figure 3.3.

Table 3.4 The scores given by the first expert to the liver segmentations from CNN for 40 test SIRT datasets.

| Score | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 |
|---|---|---|---|---|---|---|---|---|
| # datasets | 2 | 0 | 1 | 2 | 7 | 12 | 14 | 2 |

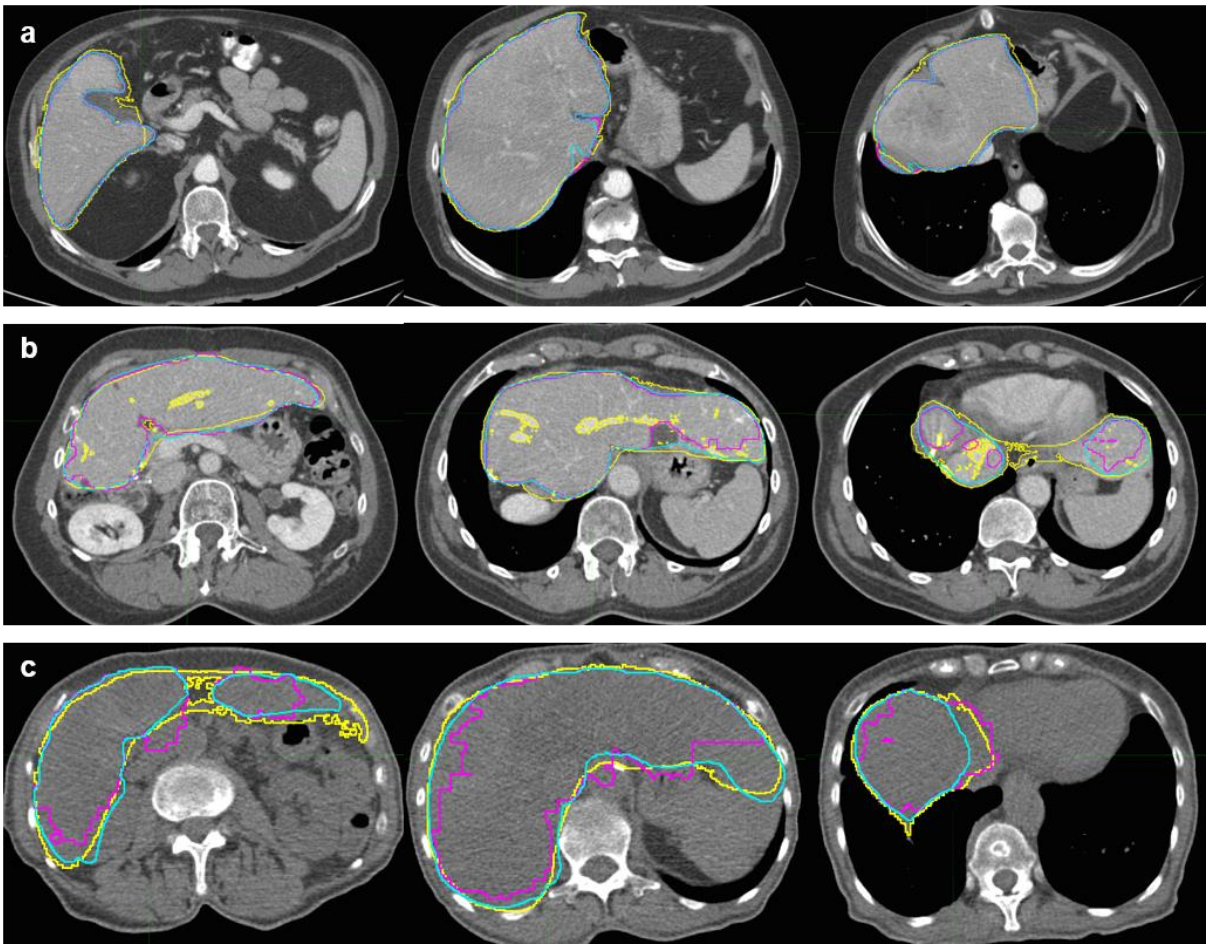



Figure 3.3 Examples of the CNN segmentation (magenta) compared to the manual segmentation (yellow) and adjusted segmentation (cyan) from the first expert. In the following, ”CNN”, ”manual”, and ”adjust” represent the CNN segmentation, the manual segmentation, and the adjusted segmentation, respectively. (a) Case where the CNN segmentation was scored 4.5: the DSCs (CNN vs manual, CNN vs adjust, manual vs adjust) were 0.92, 0.99, and 0.92. (b) Case where the CNN segmentation was scored 3: the DSCs (CNN vs manual, CNN vs adjust, manual vs adjust) were 0.91, 0.95, and 0.93. (c) Case where the CNN segmentation was scored 1: the DSCs (CNN vs manual, CNN vs adjust, manual vs adjust) were 0.84, 0.89, and 0.92.

When looking into the reasons why the CNN model produced poor segmentations on some datasets, we identified the following scenarios which were present in the SIRT datasets but very infrequent in the training datasets: low contrast or low dose, lesions with large density difference from their surroundings, extreme liver position and shape. Some examples of these cases are presented in Figure 3.3. In Figure 3.3(b), one round lesion with low density is seen in the second image and part of the left lobe is located in the extreme left lateral position within the abdomen. The CT shown in Figure 3.3(c) has very low dose and low contrast.

### 3.4.2 Inter-observer variability

The median DSC, RV, MSD, and HD between the 20 manual segmentations were 0.94, 1.08, 2.0 mm, and 25.0 mm (see Figure 3.4). The median DSC, RV, MSD, and HD between the 20 adjusted segmentations were 0.98, 1.01, 0.6 mm, and 21.0 mm (see Figure 3.4). According to the results of DSC and RV, the volume difference between the adjusted segmentations was much smaller than that between the manual segmentations. Similarly, the mean surface distance between the two adjusted liver contours was reduced to a large extent compared with the manual contours from the two experts. The relative decrease of HD was not as large as that of the other three metrics after adjustment. It is mainly because a large discrepancy of delineation between two experts exists in the regions of vessels or ligaments,

where the delineation criteria are not clearly defined. This discrepancy cannot be eliminated using the CNN segmentation as a baseline.

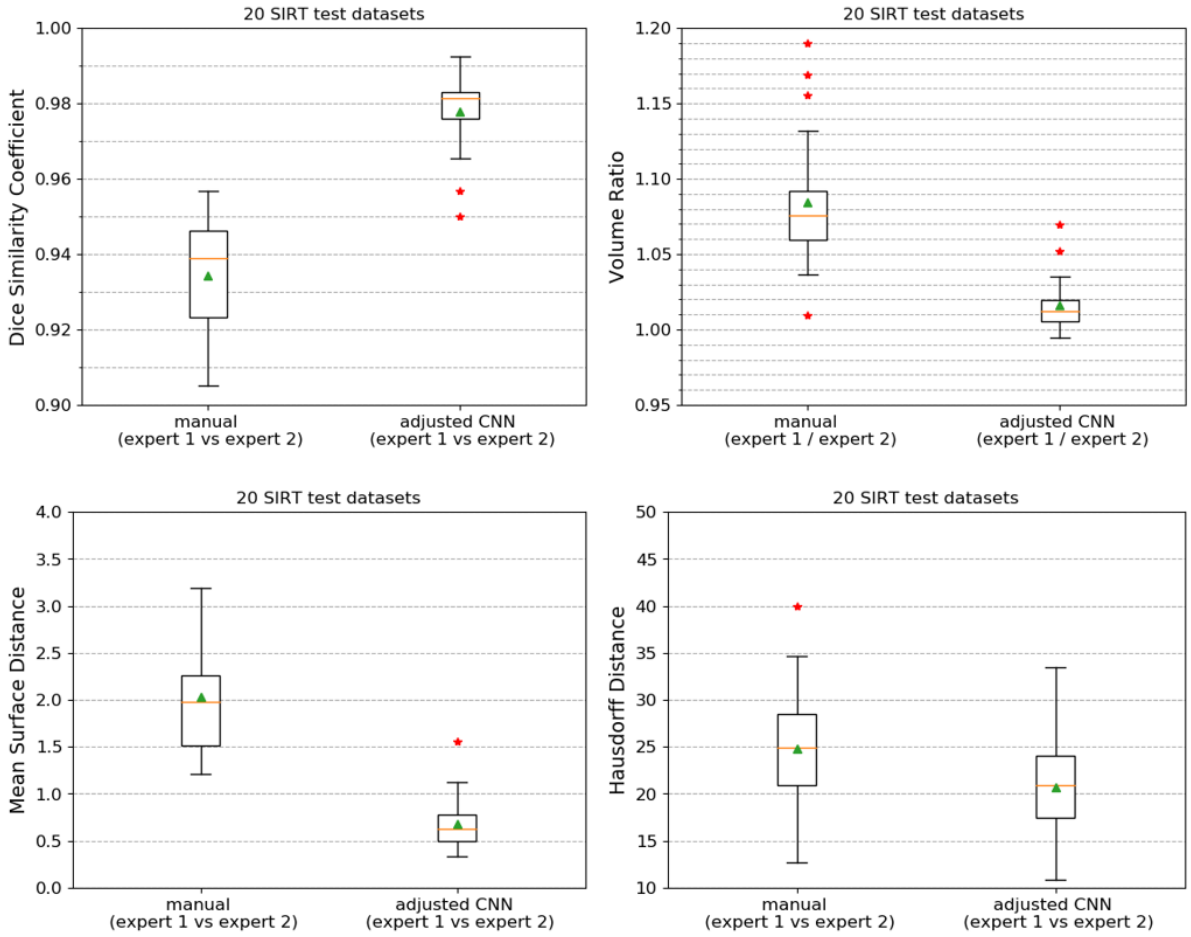



Figure 3.4 Comparison of the agreement between the two manual expert delineations and between the two manually adjusted segmentations using 4 metrics: (top left) dice similarity coefficient, (top right) volume ratio, (bottom left) mean surface distance, and (bottom right) Hausdorff distance.

Besides, the scores of the 20 test SIRT datasets from the two experts are presented in Figure 3.5. The score difference remains within 0.5 for 16 patients. However, a large score difference of over 0.5 exists for the other 4 patients, although the two experts used similar scoring criteria. It is caused by the subjectivity existing in the criteria and in the judgement from the experts.

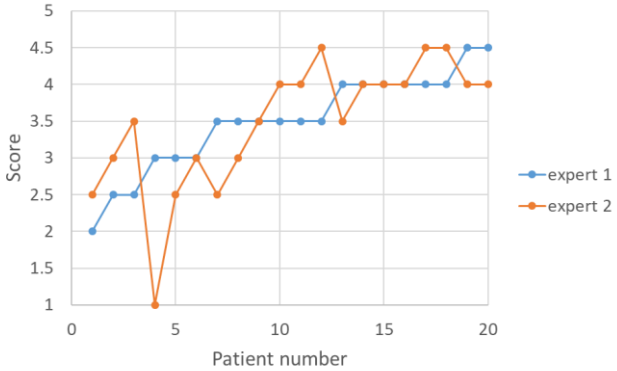



Figure 3.5 The scores of the 20 test SIRT datasets given by experts 1 and 2. The patients are sorted in an ascending order according to the scores given by expert 1.

### 3.4.3 Analysis of manual adjustment

The frequency of every corrected region for each expert was recorded for the 20 test SIRT datasets (see Figure 3.6). From the figure, it is evident that inferior vena cava (IVC) is the region corrected by both experts most frequently. In the training datasets, a part of IVC adjacent to the liver was included in the liver delineation in some datasets while not in the other datasets. As a result, the CNN segmentation appears random and irregular in the IVC region. For the portal vein, expert 1 tended to include it in the liver delineation while expert 2 agreed more with the CNN segmentation to exclude the portal vein from the liver segmentation. Besides, CNN segmentation errors in some regions required additional adjustment from the experts. For example, the left tip of liver is the third most frequently adjusted region since shape abnormality often occurs in this region. The lesions with large density difference from their surroundings are the fourth most frequently corrected regions. The regions between the liver and the surrounding organs (e.g. heart, stomach, duodenum, colon, and so on) are frequently corrected due to CNN segmentation errors caused by their small density difference in low contrast CTs.

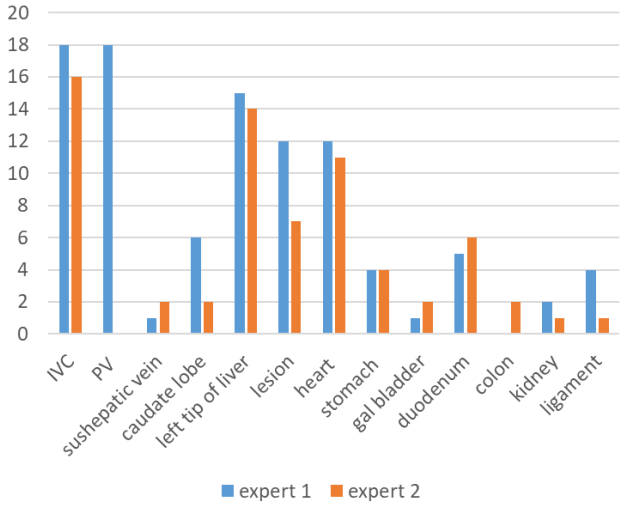



Figure 3.6 The frequency of each region corrected by every expert for the 20 test SIRT datasets. PV - Portal Vein. IVC - inferior vena cava.

### 3.4.4 Time used for manual segmentation and adjustment

The time spent on manual segmentation and adjustment of the CNN segmentation for expert 1 (40 test SIRT datasets) and expert 2 (20 test SIRT datasets) is presented in Figure 3.7. For expert 1, the time for manual segmentation is always within 5 minutes, which is shorter than the time for adjustment. The time for adjustment ranges from 3.17 to 32.75 minutes with a median of 9.18 minutes. Expert 2 spent much less time on adjustment than on manual segmentation. The time for manual segmentation ranges from 22.32 to 64.82 minutes with a median of 28.53 minutes and the time for adjustment ranges from 2.15 to 20.45 minutes with a median of 6.72 minutes.

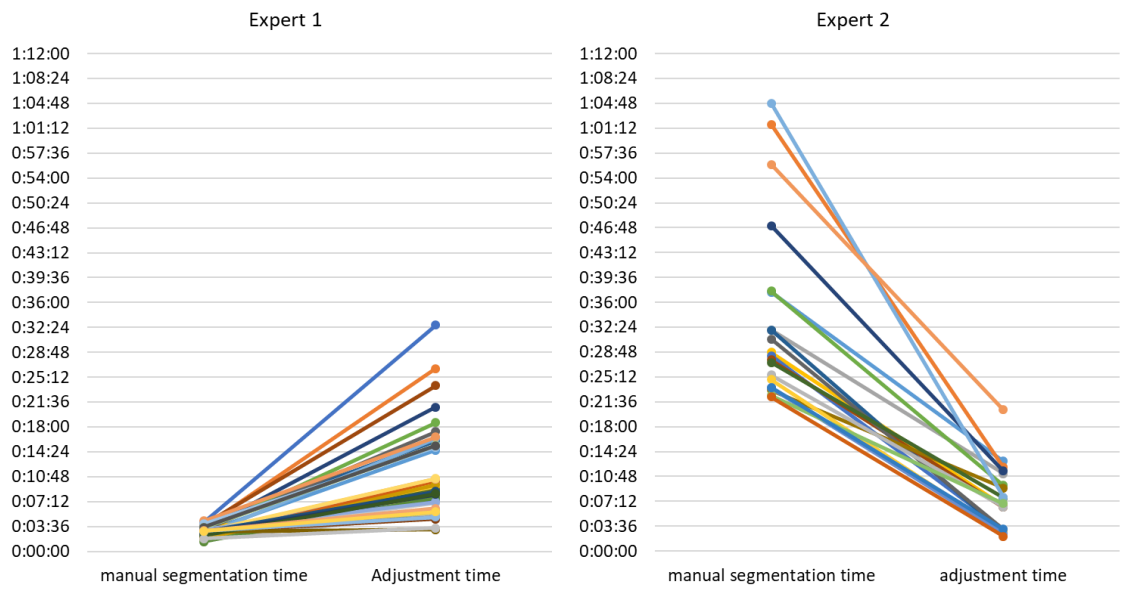



Figure 3.7 Time (hours:minutes:seconds) spent on manual segmentation and adjustment of the CNN segmentation for expert 1 (40 test SIRT datasets) and expert 2 (20 test SIRT datasets).

## 3.5 Discussion

Our modified CNN model mainly trained on public datasets of liver cancer achieved good results on the SIRT CT images with good image quality, relatively normal liver shapes, and low disease burden. The CNN segmentation achieved a median DSC of 0.94 with the manual segmentation and of 0.98 with the adjusted segmentation, respectively. Only 2 out of the 40 test SIRT datasets had a RV outside the range from 0.9 to 1.1 between the CNN segmentation and adjusted segmentation. It indicates that the difference of injected activity caused by CNN segmentation errors is within 10% for 95% of the 40 test SIRT datasets when using the mono-compartment method. 87.5% (35/40) of automatic liver segmentations from CNN are eligible for clinical use with limited adjustment from the expert in the judgment of 2 experienced liver delineators. This implies a promising future for applying deep learning to the traditional liver segmentation task in the clinical routine of SIRT.

However, the current CNN model may fail in the following cases: poor image quality (low dose or low contrast CT), lesions with large density difference from their surroundings, and extreme liver position and shape. Each of the above cases has many different variations. A small density difference can occur among most organs in the abdomen or between the liver and a neighboring organ. The lesions in the liver may appear homogeneous and round with very low density, large and diffuse, or with high vascularity. The liver can be extremely large or compressed in the sagittal plane and the left lobe may occur in the very left position of the abdomen. These variations and their combinations make them difficult to be defined and quantified. Through visual inspection, it was found that the above three cases and their variations occurred in the training datasets (mainly the challenge datasets) with low frequency. Besides, the DSCs of the challenge datasets for validation and the SIRT datasets for validation are 0.97 and 0.94 respectively when comparing the CNN segmentation with the manual segmentation from the radiographer. This further proves that some discrepancy exists between the SIRT datasets and the challenge datasets.

By using the CNN segmentation as a baseline for adjustment, the inter-observer variability was reduced to a large extent compared with starting the manual liver segmentation from scratch. It can help reduce the random and subjective errors in absorbed dose calculation introduced by inconsistent liver volumes and contours from different observers. The ratio of RVs outside the range from 0.9 to 1.1 is

0% (0/20) between the adjusted segmentations from the two experts and 20% (4/20) between their manual segmentations. This implies that the adjusted segmentations keep the difference of injected activity caused by the inter-observer variability of the liver segmentation within 10% if the mono-compartment model is used.

Currently, the corrections from the experts mainly happen in the regions including the vessels (IVC, portal vein, sushepatic vein) and in the regions where the CNN model has a poor delineation. Since there are no criteria defining the way of including or excluding these vessels for the liver delineation, the experts make the decision based on their own experience and background (e.g. radiographer vs nuclear medicine physician). On CT without intravenous contrast enhancement, the IVC is difficult to discern from normal liver tissue, contrary to contrast-enhanced CT. This further increases the difficulty of liver delineation near the vessel regions. Although the contour difference caused by these vessels does not have an evident influence on dosimetry, it decreases the consistency of liver delineation. This can be solved by proposing a criterion for vessels exclusion agreed upon by the physicians.

It is remarkable that expert 1 needed more time for adjusting a segmentation than for drawing one from scratch, whereas the opposite was the case for expert 2. For this experiment, we allowed the experts to use the segmentation software of their choice. Expert 1 is a radiographer who is doing clinical segmentations since many years, and he did the manual segmentations with the software which he uses also clinically: the Siemens' Syngo MMWP Volume software. However, he found that this software is less suited for correcting existing segmentations and therefore used the MIM software for that, which he had not used before. Expert 2 is a nuclear medicine physician, who is not used to providing manual organ segmentations. He chose the MIM software for both tasks. Consequently, we attribute this discrepancy to the many years of experience of expert 1 with the Siemens software. We cannot claim that correcting a segmentation is always faster than providing one from scratch, as these times depend heavily on the software used for that task and the talents of the operator for using that software efficiently. However, we would argue that when the software is optimized for the task, a skilled operator should be faster at correcting a fairly good segmentation than at creating a new one, since the former task is simpler in principle.

We will introduce our CNN-based correction tool into the SIRT workflow and possibly other clinical workflows involving liver segmentation. Once the experts get used to this tool, shorter time may be spent on liver delineation with better accuracy. As a result, it will become easier for the experts to provide a large amount of liver contours eligible for training the CNN model, further improving the CNN performance. As assistance from the current CNN already improved the inter-observer agreement, we believe this CNN assisted liver segmentation will contribute to improving and standardizing the liver contours used in SIRT planning and help nuclear medicine physicians to obtain more precise dose predictions and better treatment verification.

In summary, we believe that the performance of our current CNN makes it a useful tool for clinical SIRT image analysis. In addition, further improvements are anticipated by including more representative SIRT work-up datasets for training, which will reduce the discrepancy between the characteristics of the training images and those of the typical SIRT images. Besides, the potential of the CNN model to reduce the segmentation time remains to be fully studied in the future. A CNN model for MRI liver segmentation is planned to be developed in the future. The reduction of inter-observer variability for MR is also anticipated.

## 3.6 Conclusion

The CNN-based automatic liver segmentation achieved good results for CT images from SIRT patients, who usually have abnormal liver shapes and high tumor burden. 87.5% of the 40 CNN liver segmentations were considered eligible for clinical use with limited adjustment from the expert. The

inter-observer variability of liver segmentation was reduced considerably when the CNN segmentation was used as a baseline for manual adjustments. As a result, the CNN-based automatic liver segmentation is anticipated to become a valuable tool for clinical routine in the near future.

# 4. Generalized CNN for liver segmentation of CT and MR

## 4.1 Abstract

**Purpose:** In SIRT, both CT and MR are used to acquire accurate total liver segmentations for dose planning and verification. Our goal was to investigate the feasibility of training a generalized CNN for liver segmentation of CT and MR, which can produce smooth liver contours. **Methods:** A U-net model was proposed to obtain contours smoother than those produced by the DeepMedic model. Its performance on CT datasets was compared with the performance of the DeepMedic model by using quantifiable metrics and visual inspection. The U-net and DeepMedic models were trained with 169 CT datasets and validated with 27 CT datasets from public challenges and the hospital. After that, 140 training MR datasets and 35 validation MR datasets were collected from a public challenge and the hospital. The U-net model was trained with only CT datasets (CT-specialized), only MR datasets (MR-specialized), and both CT and MR datasets (generalized). The generalized U-net model was compared with the CT-specialized and MR-specialized models by evaluating their performance on the validation CT and MR datasets. All CNN models were evaluated with Dice Similarity Coefficient (DSC), Mean Surface Distance (MSD), and Hausdorff Distance (HD). **Results:** Both the CT-specialized U-net and DeepMedic models achieved a mean DSC of 0.97 for the validation CT datasets from the challenges. For the validation CT datasets from the hospital, the mean DSC for the CT-specialized U-net model were 0.01 higher than those for the CT-specialized DeepMedic model, and the mean MSD and HD for the CT-specialized U-net model were 0.43 mm and 7.6 mm lower than those for the CT-specialized DeepMedic model, respectively. Both the generalized and CT-specialized U-net models achieved a mean DSC of around 0.97 (p-value: 0.11) for the validation challenge CT datasets and a mean DSC of around 0.96 (p-value: 0.31) for the validation CT datasets from the hospital. Both the generalized and MR-specialized U-net models achieved a mean DSC of around 0.93 (p-value: 0.94) for the validation challenge MR datasets and a mean DSC of 0.93 (p-value: 0.90) for the validation MR datasets from the hospital. The CT-specialized model achieved a mean DSC of 0.20 and 0.47 for the validation challenge and hospital MR datasets, respectively. The MR-specialized model achieved a mean DSC of 0.90 and 0.89 for the validation challenge and hospital MR datasets, respectively. **Conclusion:** The U-net and DeepMedic models had a comparable performance for the challenge CT datasets. However, the DeepMedic model produced poorer contour quality for the CT datasets from the hospital, compared with the U-net model. The images from the hospital are more challenging for segmentation than those from the challenges, due to lower contrast, higher noise level, and abnormal liver shapes. Therefore, the U-net model was selected for liver segmentation in future clinical application. The MR-specialized model achieved reasonable performance for CT liver segmentation, while the CT-specialized model failed for MR liver segmentation. MR images provide more information on liver features than CT images do. The generalized U-net model achieved good performance for liver segmentation compared with the specialized models. It is feasible to produce both CT and MR liver segmentation using a single model without degrading the performance.

**Key Words:** convolutional neural network (CNN); selective internal radionuclide therapy (SIRT); liver segmentation; CT; MR

## 4.2 Introduction

Before performing selective internal radiation therapy (SIRT) on a patient with a liver tumor, several imaging studies are performed for liver volumetry and tumor detection. Contrast-enhanced MR images

are used for HCC detection. For liver metastases, [$^{18}$F]FDG or [$^{68}$Ga]Ga-DOTA-TATE PET/CT or contrast-enhanced CT scans are used. During the pre-treatment workup of SIRT, a $^{99m}$Tc-MAA SPECT/CT simulation is performed to obtain the estimated activity distribution for treatment planning. $^{90}$Y PET/CT or PET/MR is performed to acquire the actual activity distribution during treatment as a verification in the post-treatment study. CT and MR are anatomical imaging modalities and as such are critical for liver and tumor delineation. Besides, liver and tumors delineated on CT and MR can be used as landmarks to guide the registration of multi-modal images [98, 99, 114]. Manual liver and tumor delineation is time-consuming and labor-intensive, which impedes the precise liver and tumor contours from being applied to dose calculation. Therefore, it is crucial to develop an automatic method for liver segmentation of CT and MR. The clinical workflow for obtaining automatic liver segmentation can be simplified if a generalized method is developed for both CT and MR.

Different methods were proposed for generalized liver segmentation on CT and MR. CNNs were the most studied approaches among them. Chartrand *et al*. proposed a semi-automatic method for liver segmentation on CT and MR [115]. A few user-generated liver contours were used to initialize the 3D liver model. Then the model was automatically deformed by a Laplacian mesh optimization approach so that the liver was precisely delineated. The liver delineation was corrected by the user to improve the segmentation. They reported an average volumetric overlap error of 5.1% for CT and 7.6% for MR [115]. Mulay *et al*. proposed a liver segmentation method for CT and MR by combining a holistically-nested edge detection (HED) and a mask-region CNN (R-CNN) [116]. The HED was used to obtain an edge map for contrasted-enhanced CT and MR. After that, the R-CNN was applied to compute liver segmentations from the edge maps. They reported a DSC of 0.94 for CT, 0.91 for T1-weighted MR, and 0.89 for T2-weighted MR [116]. Wang *et al*. implemented a 2D U-net model for liver segmentation on CT and MR [70]. The U-net model was first trained with 300 unenhanced multi-echo spoiled gradient-echo MR images. After that, the U-net model was generalized by using transfer learning with 30 contrast-enhanced MR and CT images. They reported a DSC of 0.94±0.06 for 230 CT images, 0.95±0.03 for 100 T1-weighted MR images, and 0.92±0.05 for 168 T2-weighted MR images [70]. Christ *et al*. trained two cascaded U-net models with CT and MR images for liver and lesion segmentation [117]. The first U-net was trained to segment only livers and the second U-net segmented only lesions within the predicted liver segmentations from the first U-net. A DSC of 94% for CT and a DSC of 91% for MR were reported for liver segmentation [117].

The DeepMedic model introduced in Section 3.3.2 has the advantage of processing large image volumes with limited computation resource and memory by inputting small patches sampled from the image down-sampled by different sampling factors. The feature volumes were extracted from the patches at different resolutions via multiple convolutional layers in three resolution pathways. The feature volumes at different resolutions were up-sampled to the original image resolution and combined through two convolutional layers to output the liver segmentation. The two convolutions cannot ensure sufficient fusion of the up-sampled low-resolution liver features and the liver features at the original resolution, which might lead to a mosaic effect (see Figure 4.3d). As a result, the contours of the thresholded DeepMedic output might exhibit right-angled boundaries (see Figure 4.3b), which never occur in manual liver delineation. The U-net structure [73] proposed a more elegant approach of down- and up-sampling to leverage the limited computation capability and memory for better integration of global and local features. The input image of the original resolution was gradually down-sampled by a small sampling factor in each pathway after several convolutional layers. The low-resolution feature volumes were up-sampled by the same sampling factor and integrated with the feature volumes of higher resolution by several convolutional layers in each pathway. In this way, the feature volumes of different resolutions can be gradually integrated to prevent segmentations of sharp edges. However, the input volume size of U-net is much larger than that of DeepMedic, because it does not extract patches and does not apply down-sampling before the input into the CNN. Therefore, the computational requirements and memory usage for U-net are higher than that for DeepMedic.

In the current SIRT imaging studies, the $^{99m}$Tc-MAA SPECT and $^{90}$Y PET have a relatively low resolution (not below 3 mm) compared with CT and MR. Since liver segmentations are to be mapped on the SPECT and PET for dose calculation, they do not need to be acquired from CT and MR with high resolution. The original image can be resampled to a larger voxel size so that the volume size of CNN input is decreased dramatically. This reduces the demand for computation capability and memory for CNN training. Therefore, the U-net structure becomes usable for liver segmentation in SIRT.

The aim of our study is to investigate the feasibility of developing a generalized U-net model for liver segmentation on CT and MR. For this purpose, the performances of liver segmentation using the DeepMedic model and the U-net model were compared on CT datasets. After that, the U-net model was trained with only CT datasets (CT-specialized), only MR datasets (MR-specialized), both CT and MR datasets (generalized). The generalized model was compared with the CT-specialized and MR-specialized models for CT and MR liver segmentation. CT and MR are two distinct modalities with very different intensity appearance. One aim of these comparisons is to find out if one of the modalities provides more information about liver anatomy than the other, and therefore contributes more to the performance of the generalized CNN.

## 4.3 Materials and methods

### 4.3.1 Data

In our study, the CNN training and validation datasets are from our hospital (65 CT datasets and 135 T1-weighted MR datasets) and public challenges, including the LiTS17 challenge (111 CT datasets), the SLIVER07 challenge (20 CT datasets), and Combined Healthy Abdominal Organ Segmentation (CHAOS) challenge (40 T1-weighted MR datasets). The CT datasets from the challenges have an image in-plane pixel size ranging from 0.56 to 1.0 mm and a slice thickness ranging from 0.7 to 5.0 mm. The CT datasets from the hospital have an image in-plane pixel size between 0.65 to 1.37 mm and a slice thickness from 1.0 to 5.0 mm. For the MR datasets from the challenge, the image in-plane pixel size ranges from 1.44 to 2.03 mm and the slice thickness ranges from 5.5 to 9.0 mm. The MR datasets from our hospital have an image in-plane pixel size between 0.76 and 1.95 mm and a slice thickness between 1.5 and 6.6 mm. Due to the lack of sufficient CT datasets with precise liver delineations for CNN training, the datasets are only divided into the training and validation datasets, without test datasets. The validation datasets can be used to evaluate the performance of CNN models, because they are not directly used in the computation process for the model optimization. The number of the training and validation datasets are presented in Table 4.1. The training and validation datasets were divided through one random selection.

Table 4.1 Number of the training and validation datasets

| | | Training | Validation | Total |
|---|---|---|---|---|
| CT | LiTS17 & SLIVER07 | 119 | 12 | 131 |
| | Hospital | 50 | 15 | 65 |
| T1-weighted MR | CHAOS | 30 | 10 | 40 |
| | Hospital | 110 | 25 | 135 |

The public challenges provided liver delineations from the experts for the CT and MR datasets. The liver delineation of the CT datasets from the hospital was performed by a radiographer with over 10

years of experience in this task. The MR datasets from the hospital were delineated by a student researcher and the liver delineations were verified or corrected by the radiographer. All hospital datasets for this study were evaluated at the KU Leuven after approval by the Ethics Committee Research of the University Hospitals & KU Leuven.

### 4.3.2 U-net model development

The CNN model proposed in this chapter adopts a U-net structure. The U-net is a type of CNN, which was developed and widely used for biomedical image segmentation [73]. Like the DeepMedic structure introduced in Section 3.3.2, the U-net structure makes use of down- and up-sampling and integration of multi-resolution features to output the classification of multiple voxels in an image, using the limited GPU memory. The implementation of these strategies are different for the U-net and DeepMedic structures.

As is shown in Figure 4.1, the proposed U-net model has four resolution levels with a down- and up-sampling rate of 3 between two neighbouring levels. In the encoder (the left half of U-net), the input whole image passes through two convolutional layers to obtain local liver feature volumes. The feature volume is down-sampled by three to obtain a larger receptive field, which is the number of voxels covered by one element of the feature volume through the previous convolutions and down-sampling. The down-sampled feature volume is processed by two convolutions and the produced feature volume with more global information is down-sampled by three again. This process is repeated until the fourth resolution level. By using down-sampling, the global liver context can be rapidly learned without using multiple convolutions, which consume a large amount of computational resources. However, this also sacrifices the image resolution that contains detailed liver features. The problem is solved by adding the feature volume before down-sampling in each resolution level back in the decoder (the right half of U-net). The feature volume in the fourth level is up-sampled by three to reach the resolution of the third level and concatenated with the feature volume from the encoder of the third level. The concatenated volume passes two convolutional layers to integrate the more global feature from the fourth level and the more local feature from the third level. This process is repeated until the first level. The information of the final feature volume is compressed by a classification layer to output a probability map, where each voxel represents its probability of being classified as the liver.

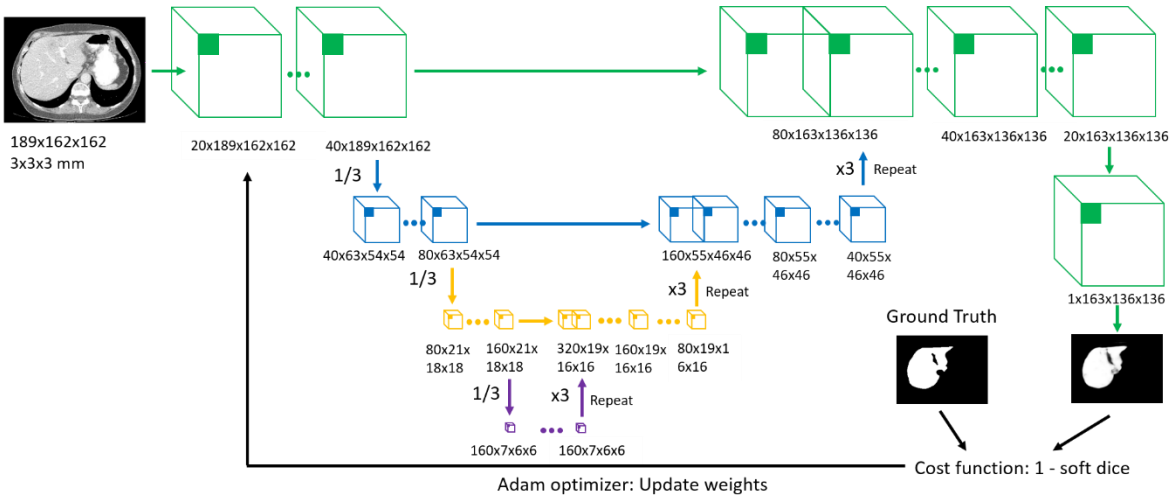



Figure 4.1 Overview of the U-net model. The input image has a fixed size of $189 \times 162 \times 162$ and an isotropic resolution of 3 mm. The output size is $163 \times 136 \times 136$. The down- and up-sampling rate is 3 for the neighboring resolution levels. The size of the convolutional kernel is $3 \times 3 \times 3$.

As is shown in Figure 3.1, the input image is sampled by different sampling factors before being put into the DeepMedic structure. This results in a large loss of details in the image before the details are used for feature extraction by the convolutional layers, especially for the large down-sampling factor.

The feature volumes extracted by multiple convolutions in three resolution pathways are up-sampled to the original resolution and integrated by two convolutional layers before being sent to the classification layer. There is no information communication among three pathways in the process of 10 convolutions. The feature volumes from the three pathways do not have enough communication to be well integrated. Compared with the DeepMedic, the U-net structure performs down- and up-sampling in the encoder and decoder by a small sampling factor at each resolution level. This limits the resolution loss at each level and ensures sufficient information communication for better feature integration.

**Training**. The U-net model used soft DSC as the loss function with the Adam optimizer [118]. The Adam optimizer minimizes the loss function through gradient descent. Compared with the traditional gradient descent methods, it has the advantage of accelerating the optimization process. The initial learning rate was 0.001 and decreased by a factor of 0.2 when the DSC of the full segmentation of the validation datasets did not increase. The model quality was evaluated every 10 epochs[2] on the full segmentation of the validation datasets using the DSC. The model parameters that achieved the highest DSC on the validation datasets were saved during the evaluation. The training process of the generalized U-net model took around 3 days using 2 NVIDIA P100 GPUs with 16 GB DRAM each. The CNN prediction time for CT and MR was around 1 second using the GPUs. When using a CPU of Intel Xeon E5-2699, the time for the CNN prediction of CT and MR was around 17 seconds.

**Data pre-processing**. The 3D CT and T1-weighted MR images were cropped to cover the whole abdomen in each axial slice and the entire liver in the axial direction. The images were resampled to an isotropic voxel size of 3 mm. The CT image was clipped to HU values between -200 and 200 and normalized by a linear mapping to the range between -0.5 and 0.5. The T1-weighted MR was clipped between the minimum intensity and 80% of the maximum intensity of the image, and mapped to the range of [-0.5, 0.5].

**Data post-processing**. The CNN output was treated as a probability map, where each voxel value ranged from 0 to 1. The map was thresholded by 0.5 to obtain the liver segmentation. The thresholded map might contain multiple non-connected regions. The largest region was retained as the liver segmentation and other smaller regions were discarded.

### 4.3.3 Experiments

**Comparison between the U-net and DeepMedic models**. To evaluate the performance of the U-net and DeepMedic models, the two models were trained with 119 CT datasets from the public challenges and 50 CT datasets from our hospital (see Table 4.1), and the manual liver segmentations for these datasets. After training, their performance was evaluated on 12 CT datasets from the challenges and 15 CT datasets from the hospital. The quantified metrics for evaluation, including DSC, MSD, and HD (see Section 3.3.4), were computed between the CNN liver segmentation and the manual liver segmentation. Besides, the liver segmentations from the two models were visually inspected to compare the contouring quality.

**Comparison among the generalized, CT-specialized, and MR-specialized U-net models**. The CT-specialized U-net model refers to the U-net model mentioned in the last paragraph. The MR-specialized model was trained with 30 and 110 T1-weighted MR datasets from the CHAOS challenge and the hospital (see Table 4.1), respectively, and the manual liver segmentations for these datasets. The generalized model was trained with the CT and MR datasets and their liver segmentations used for training the specialized models. The performance of the three models was evaluated on the validation CT and T1-weighted MR datasets from the challenges and the hospital (see Table 4.1) by computing the DSC, MSD, and HD between the CNN liver segmentation and the manual liver segmentation.

[2] One epoch indicates that all input images are processed by the CNN one time.

### 4.3.4 Statistical tests for comparison of CNN models

As is described in Section 4.3.3, multiple models designed with different structures (the U-net and DeepMedic models) and trained with images of different modalities (the CT-specialized, MR-specialized, and generalized U-net models) were compared to evaluate the influence of the CNN structures and the training image modalities on liver segmentation performance. The metrics (DSC, MSD, HD) between the CNN segmentations on the same datasets from different CNN models were compared by using statistical tests to determine if there was significant difference between them. The Shapiro-Wilk test [119] was used to test if the distribution of the metric values for a CNN model is normal. The paired t-test was used if the distribution of the metric values was normal [120]. Otherwise, the Wilcoxon signed-rank test was used [120]. The level of significance was set to 0.05. The above three statistical tests are also used for results comparison in the following chapters.

## 4.4 Results

### 4.4.1 Comparison between the U-net and DeepMedic models

As shown in Figure 4.2, the performances of the U-net and the DeepMedic models on the validation CT datasets did not show statistically significant difference ($p > 0.05$). The U-net and DeepMedic models achieved close DSC values (mean DSC: 0.97) for the 12 validation CT datasets from the challenges. The difference in the mean MSD and HD values of the challenge CT datasets from the two models was within 0.1 mm. For the hospital datasets, the mean DSC of the U-net was 0.01 higher than the mean DSC of the DeepMedic model. The mean MSD and HD of the U-net for the hospital datasets were 0.43 mm and 7.6 mm lower than the mean MSD and HD of the DeepMedic model, respectively. The U-net model improved the metrics for measuring the segmentation and contour quality, especially the HD, for the datasets with poor segmentations using the DeepMedic model. The HD measures the maximum distance between two contours, which means the DeepMedic model produced more liver segmentations with poor contouring in some areas. As mentioned in Section 4.3.2, the DeepMedic structure tends to generate liver segmentations with sharp edges (right-angled edges). One example of the liver segmentations on the same CT from the two models is shown in Figure 4.3. The CNN output was thresholded by 0.5 and resampled to the original image resolution. Right-angled edges occurred within a part of the stomach that was included in the liver segmentation from the DeepMedic model.

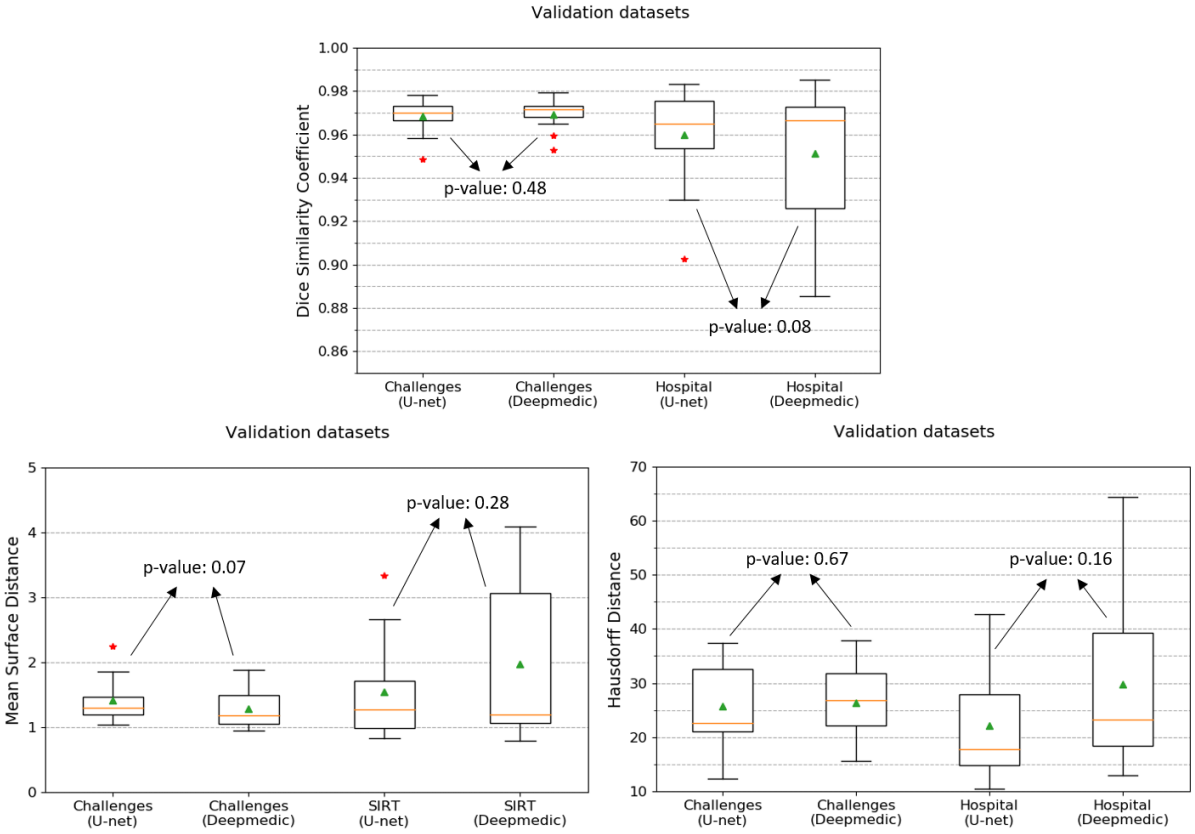



Figure 4.2 Comparison of the DSC (top), MSD (bottom left), and HD (bottom right) for the validation CT datasets from the public challenges and the hospital between the U-net and DeepMedic models. The challenge CT datasets are from the LiTS17 and SLIVER07 challenges. The number of the validation CT datasets are presented in Table 4.1. The median, mean, and outlier are represented with a yellow line, a green triangle, and a red star.

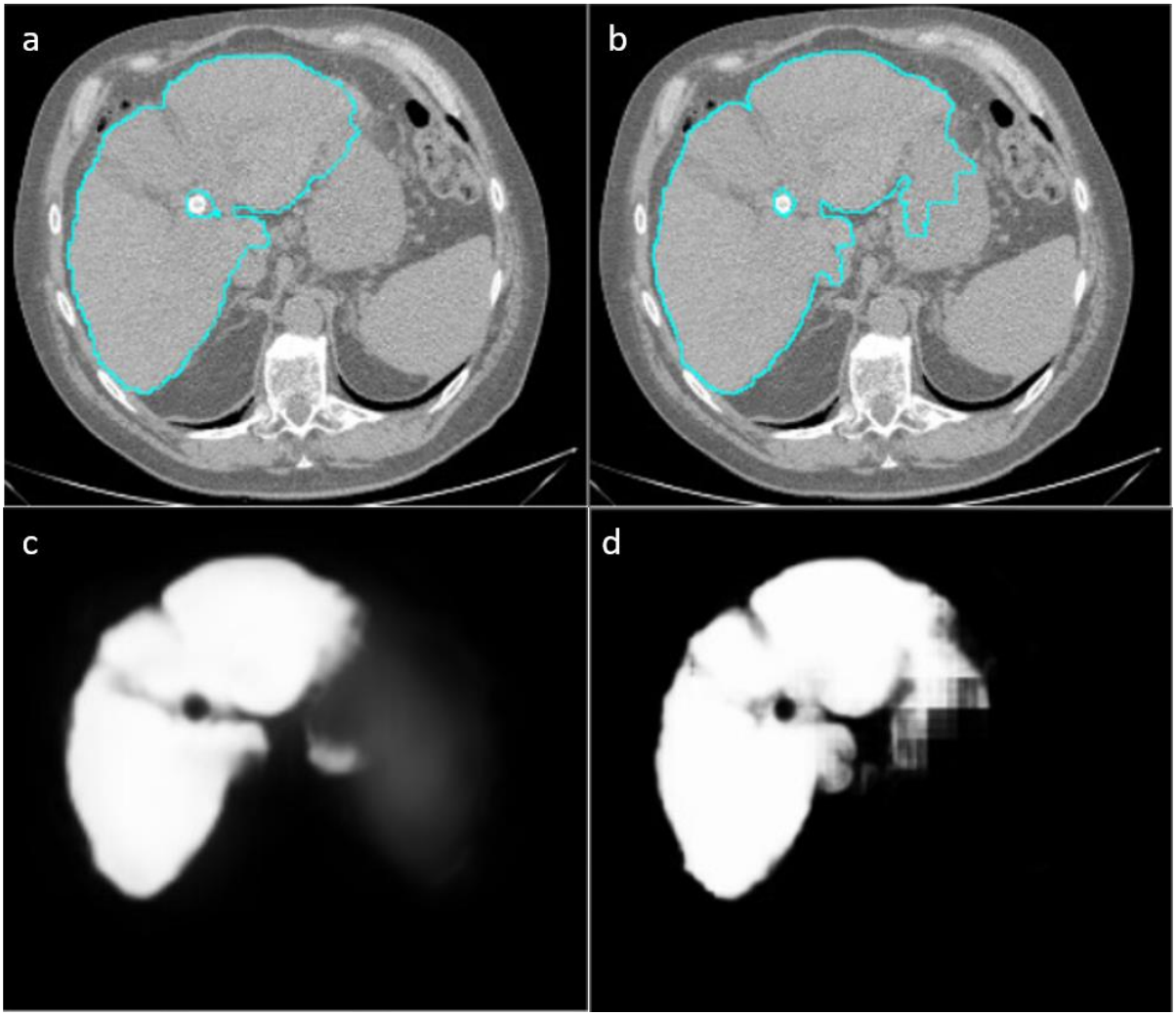



Figure 4.3 Liver segmentations of a validation CT from the CT-specialized U-net and DeepMedic models. The output of the U-net (c) and DeepMedic (d) models was thresholded at 0.5, processed by removing the small disconnected regions, and resampled to the original image resolution. The contours of the post-processed output of the U-net (a) and DeepMedic (b) models were overlayed on the CT image. A mosaic effect was observed in the DeepMedic output (d), which was caused by the insufficient fusion of the high-level and low-level liver features. This resulted in right-angled edges within the stomach that was included in the liver contour (b) from the DeepMedic model.

### 4.4.2 Comparison among the generalized, CT-specialized, and MR-specialized U-net models

As shown in Figure 4.4, the MR-specialized model achieved a smaller mean DSC on the validation CT datasets from the public challenges ($p = 5 * 10^{-5}$, $p = 3 * 10^{-5}$) and the hospital ($p = 2 * 10^{-5}$, $p = 10^{-6}$) than the CT-specialized and generalized models. For the validation MR datasets from the public challenge and the hospital, the mean DSC of the CT-specialized model was much smaller than those of the MR-specialized ($p = 7 * 10^{-5}$, $p = 10^{-5}$) and generalized ($p = 8 * 10^{-5}$, $p = 10^{-5}$) models according to Figure 4.5. The lowest DSC of the CT-specialized model was zero for the validation MR datasets (see Figure 4.5). For the validation challenge and hospital CT datasets, the absolute difference in the mean DSC between the CT-specialized and generalized models were both around 0 ($p = 0.11$, $p = 0.31$) (see Figure 4.6). For the validation challenge and hospital MR datasets, the absolute difference in the mean DSC between the MR-specialized and generalized models were both around 0 ($p = 0.94$, $p = 0.90$) (see Figure 4.6). For the validation challenge and hospital CT datasets, the absolute difference in the mean MSD (HD) between the CT-specialized and generalized

models were 0.05 (2.1) mm ($p = 0.48$ (0.17)) and 0.15 (3.5) mm ($p = 0.78$ (0.25)), respectively (see Figure 4.7 and Figure 4.8). For the validation challenge and hospital MR datasets, the absolute difference in the mean MSD (HD) between the MR-specialized and generalized models were 0.30 (3.2) mm ($p = 0.11$ (0.51)) and 0.02 (0.1) mm ($p = 0.34$ (0.92)), respectively ($p > 0.05$) (see Figure 4.7 and Figure 4.8).

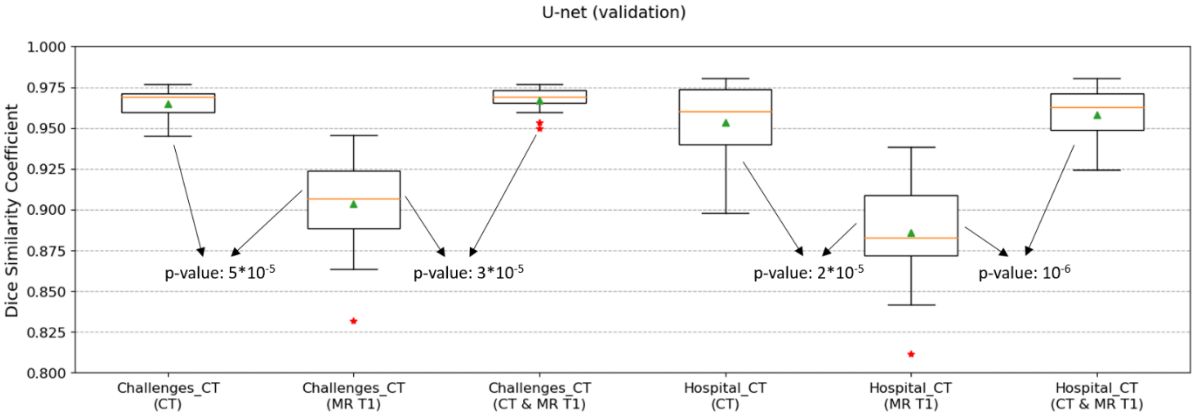



Figure 4.4 Comparison of the DSC for the validation CT datasets from the public challenges and the hospital between the CT-specialized, MR-specialized, and generalized U-net models. (CT): the CT-specialized U-net model; (MR T1): the MR-specialized U-net model; (CT & MR T1): the generalized U-net model. The median, mean, and outlier are represented with a yellow line, a green triangle, and a red star.

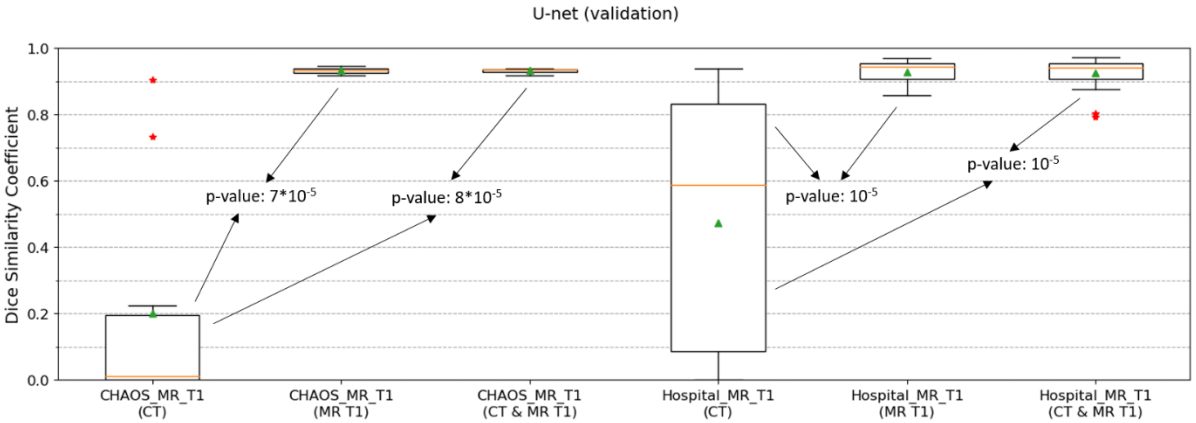



Figure 4.5 Comparison of the DSC for the validation MR datasets from the public challenge (CHAOS) and the hospital between the CT-specialized, MR-specialized, and generalized U-net models. (CT): the CT-specialized U-net model; (MR T1): the MR-specialized U-net model; (CT & MR T1): the generalized U-net model. The median, mean, and outlier are represented with a yellow line, a green triangle, and a red star.

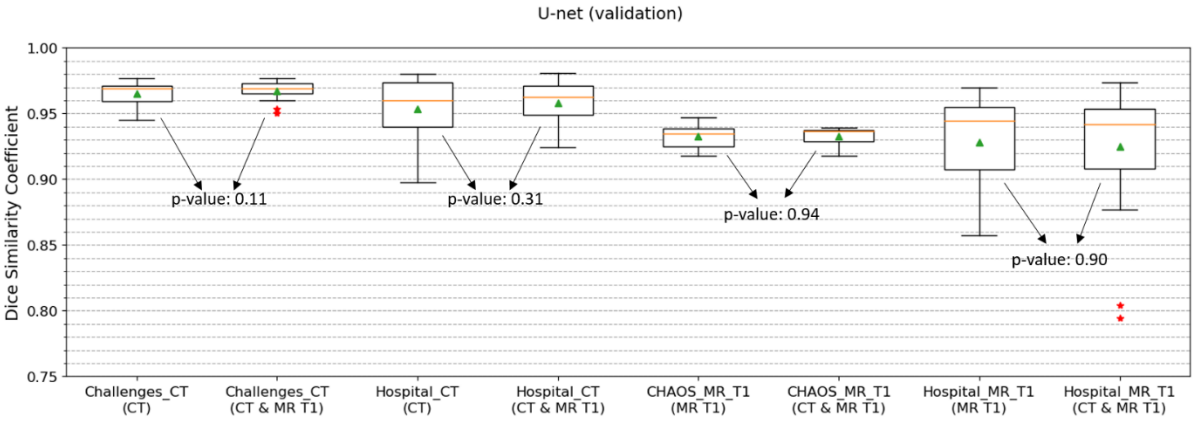



Figure 4.6 Comparison of the DSC for the validation CT and MR datasets from the public challenges and the hospital between the CT-specialized, MR-specialized, and generalized U-net models. (CT): the CT-specialized U-net model; (MR T1): the MR-specialized U-net model; (CT & MR T1): the generalized U-net model. The median, mean, and outlier are represented with a yellow line, a green triangle, and a red star.

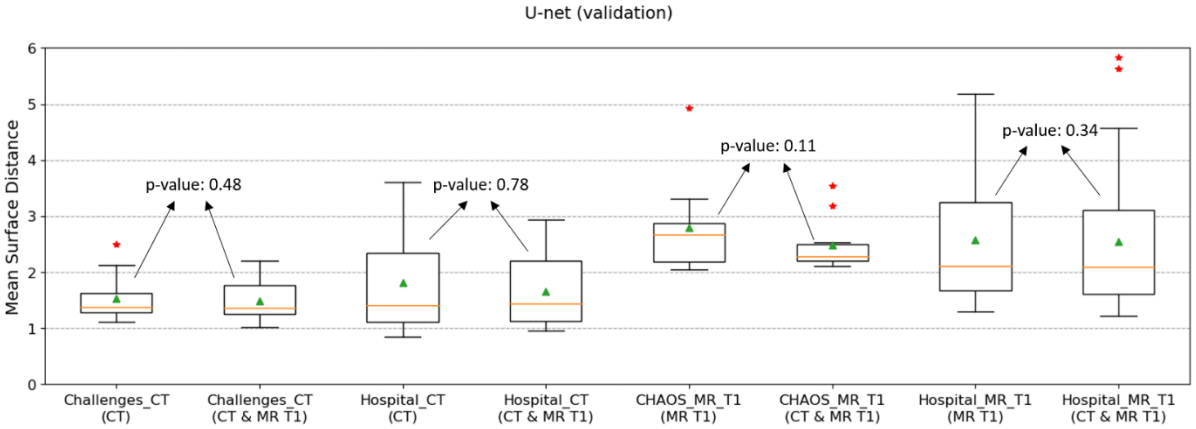



Figure 4.7 Comparison of the MSD for the validation CT and MR datasets from the public challenges and the hospital between the CT-specialized, MR-specialized, and generalized U-net models. (CT): the CT-specialized U-net model; (MR T1): the MR-specialized U-net model; (CT & MR T1): the generalized U-net model. The median, mean, and outlier are represented with a yellow line, a green triangle, and a red star.

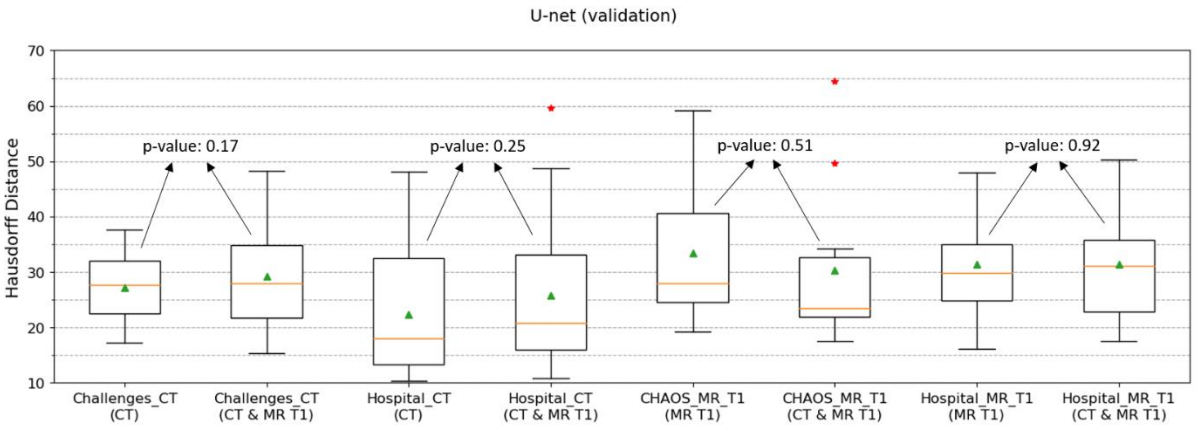



Figure 4.8 Comparison of the HD for the validation CT and MR datasets from the public challenges and the hospital between the CT-specialized, MR-specialized, and generalized U-net models. (CT): the CT-specialized U-net model; (MR T1): the MR-specialized U-net model; (CT & MR T1): the generalized U-net model. The median, mean, and outlier are represented with a yellow line, a green triangle, and a red star.

## 4.5 Discussion

The U-net and DeepMedic models achieved comparable DSCs, MSDs, and HDs for the challenge CT datasets (see Figure 4.2). The CT images from the public challenges had good contrast, low noise level, and relatively normal liver shapes. Both the U-net and DeepMedic models produced liver segmentations with similar quality. The CT datasets from the hospital include CT images with low contrast, high noise level, and abnormal liver shapes, which increase the difficulty in liver segmentation. The U-net model improved the DSCs, MSDs and HDs for the CT datasets from our hospital compared with the DeepMedic model, especially for the mean HD with an improvement of 7.6 mm (see Figure 4.2). The HD measures the maximum surface distance between two liver surfaces. The liver segmentations from the DeepMedic model had a poorer contour quality than those from the U-net model. The DeepMedic model tended to produce liver segmentations with sharp and right-angled edges according to visual inspection (see Figure 4.3), which is not in line with the underlying anatomical reality. As is presented in Figure 4.1, the U-net model has a more elegant way of integrating high-level features with low-level features. The spatial resolution is reduced by a small down-sampling factor at each level in the encoder (the left half of the U-net) to fetch the high-level information. In the decoder (the right half of the U-net), the high-level features from the previous level of lower resolution are up-sampled by the same sampling factor to be integrated with the low-level features from the encoder of the current level. The extraction of high-level features and the fusion of the high-level and low-level features are performed step by step at each level to ensure thorough information communication. For the DeepMedic model (see Figure 3.1), the input image is down-sampled by the sampling factors from one to large values. The resolution information is degraded for the image down-sampled by a large sampling factor before the information is extracted by convolutions. The fusion of the high-level and low-level features is performed at a position close to the end of the network, which does not give enough time for thorough integration. There is no information communication between features of different level before the step of feature fusion. Therefore, the DeepMedic model tended to produce an output with a mosaic effect when it had difficulty in making inference based on the high-level and low-level features for the datasets from the hospital. The output might be dominated by the up-sampled high-level features due to insufficient feature fusion. As a result, the thresholded output had right-angled edges on the boundary. Hence, the U-net model was selected for liver segmentation in the following experiments.

As shown in Figure 4.4 and Figure 4.5, the specialized U-net models had much poorer performance on the images of the modality that they were not specialized for, compared with the generalized model. The MR-specialized model achieved a reasonable performance on the CT datasets with a minimum DSC over 0.80. However, the CT-specialized model failed on the MR datasets with over half of the DSCs smaller than 0.60 and a lowest DSC of zero. Wang *et al*. trained an initial U-net for liver segmentation with 300 unenhanced multiecho 2D-SPGR MR images and further generalized the initial U-net with 30 2D-SPGR MR images, 20 contrast-enhanced T1-weighted MR images, and 10 contrast-enhanced CT images [70]. The initial U-net achieved a mean DSC of 0.82 for 230 contrast-enhanced and unenhanced test CT images [70]. To evaluate the influence of the number of CT images in the training dataset on liver segmentation performance, the number of contrast-enhanced CT images was increased from 1 to 10. The performance gain was evident with the mean DSC from 0.82 to 0.91 as the number of the CT images increased from 0 to 2, and the mean DSC plateaued at 0.94 for 10 CT images [70]. Wang's and our experiment results demonstrated that MR images provide more variations of intensities and textures than CT images, which can facilitate the generalization of the CNN model. The generalized U-net model achieved comparable performance on the CT datasets as the CT-specialized model and comparable performance on the MR datasets as the MR-specialized model. Generalizing the CT-specialized model with the MR datasets did not degrade the segmentation quality for the CT datasets. The same situation held for generalizing the MR-specialized model with the CT datasets. It is feasible to use one generalized U-net model for both CT and MR liver segmentation with performance comparable to the specialized models. Our generalized U-net model achieved similar

DSCs for liver segmentation of CT and T1-weighted MR images, compared to the DSCs reported by Wang *et al*. [70], Mulay *et al*. [116], and Christ *et al*. [117], although the used CT and MR datasets were different.

Due to the lack of sufficient CT and MR datasets with good liver delineations for CNN training, no test datasets were separated from the total datasets. The learning process was performed on the training datasets. The validation datasets were only used to evaluate the performance of different CNN structures and parameters on the unseen datasets. Since the model did not learn liver features directly from the validation datasets, the bias caused by the model tuning with the validation datasets was limited. The collection and annotation of new datasets would consume a lot of time and the main goal of this project was not to fully exploit the influence of different CNN structures and input datasets on liver segmentation. Therefore, we decided not to continue the experiments by collecting more test datasets. However, we believe that the conclusions will be stronger by carrying out unbiased evaluation of the CNN models on the test datasets.

## 4.6 Conclusion

The U-net model achieved better contour quality of liver segmentations for the CT datasets from the hospital than the DeepMedic model. CNN liver segmentation for the clinical CT images is more challenging than for the CT images from the public challenges, which usually have higher contrast, lower noise level, and normal live shapes. The U-net model was selected for the liver segmentation in future studies. The generalized U-net model achieved similar performance of liver segmentation on CT and MR images as the specialized models. It is feasible to perform liver segmentation on both CT and MR images with one generalized U-net model, which does not degrade the performance achieved by the specialized models.

# 5. Segmentation-guided multi-modal registration of liver images for dose estimation in SIRT

This chapter has been published as:

X. Tang, E. Jafargholi Rangraz, R. 's Heeren *et al*, "Segmentation-guided multi-modal registration of liver images for dose estimation in SIRT," *EJNMMI Phys*, vol. 9, pp. 3, 2022, Springer Nature. https://doi.org/10.1186/s40658-022-00432-8.

## 5.1 Abstract

**Purpose:** Selective internal radiation therapy (SIRT) requires a good liver registration of multi-modality images to obtain precise dose prediction and measurement. This study investigated the feasibility of liver registration of CT and MR images, guided by segmentation of the liver and its landmarks. The influence of the resulting lesion registration on dose estimation was evaluated. **Methods:** The liver segmentation was done with a convolutional neural network (CNN) and the landmarks were segmented manually. Our image-based registration software and its liver-segmentation-guided extension (CNN-guided) were tuned and evaluated with 49 CT and 26 MR images from 20 SIRT patients. Each liver registration was evaluated by the root mean square distance (RMSD) of mean surface distance between manually delineated liver contours and mass center distance between manually delineated landmarks (lesions, clips, etc). The root mean square of RMSDs (RRMSD) was used to evaluate all liver registrations. The CNN-guided registration was further extended by incorporating landmark segmentations (CNN&LM-guided) to assess the value of additional landmark guidance. To evaluate the influence of segmentation-guided registration on dose estimation, mean dose and volume percentages receiving at least 70 Gy (V70) estimated on the $^{99m}$Tc-labeled macro-aggregated albumin ($^{99m}$Tc-MAA) SPECT were computed, either based on lesions from the reference $^{99m}$Tc-MAA CT (reference lesions) or from the registered floating CT or MR images (registered lesions) using the CNN- or CNN&LM-guided algorithms. **Results:** The RRMSD decreased for the floating CTs and MRs by 1.0 mm (11%) and 3.4 mm (34%) using CNN guidance for the image-based registration and by 2.1 mm (26%) and 1.4 mm (21%) using landmark guidance for the CNN-guided registration. The quartiles for the relative mean dose difference (the V70 difference) between the reference and registered lesions and their correlations [25th, 75th; r] are as follows: [-5.5% (-1.3%), 5.6% (3.4%); 0.97 (0.95)] and [-12.3% (-2.1%), 14.8% (2.9%); 0.96 (0.97)] for the CNN&LM- and CNN-guided CT to CT registrations, [-7.7% (-6.6%), 7.0% (3.1%); 0.97 (0.90)] and [-15.1% (-11.3%), 2.4% (2.5%); 0.91 (0.78)] for the CNN&LM- and CNN-guided MR to CT registrations. **Conclusion:** Guidance by CNN liver segmentations and landmarks markedly improves the performance of the image-based registration. The small mean dose change between the reference and registered lesions demonstrates the feasibility of applying the CNN&LM- or CNN-guided registration to volume-level dose prediction. The CNN&LM- and CNN-guided registrations for CTs can be applied to voxel-level dose prediction according to their small V70 change for most lesions. The CNN-guided MR to CT registration still needs to incorporate landmark guidance for smaller change of voxel-level dose estimation.

**Key Words:** selective internal radiation therapy (SIRT); liver registration; convolutional neural network (CNN); internal dosimetry; multi-modality images

## 5.2 Introduction

Selective internal radiation therapy (SIRT) or radioembolization is increasingly applied for the treatment of surgically unresectable primary liver malignancies and secondary metastases. During this

treatment, microspheres loaded with β-emitting radionuclides, including yttrium-90 ($^{90}$Y) or holmium-166 ($^{166}$Ho), are infused into selected branches of the hepatic artery according to the vascular anatomy mapped by angiography [121]. Since the selected branches dominate the blood supply to tumors [100], these radioactive microspheres are trapped within the tumors. The high energy, small tissue penetrating range, and concentration in tumors allow these microspheres to deposit higher energy per mass in tumors, while preventing healthy liver parenchyma dysfunction by limiting their irradiation.

In the SIRT planning, the absorbed dose is used to measure the amount of energy per mass (in Gy or J/kg) from ionizing radiation deposited in a volume of interest (VOI), including tumors and healthy liver parenchyma. It serves as a toxicity indicator for tumors and normal tissues and a criterion for determining the amount of injected activity. There are several different methods to determine the injected activity, including mono-compartment [45] and multi-compartment [108] methods and voxel-based approaches [45]. The dose calculation better reflects the underlying biology as the VOI changes from the volume level (the whole liver, tumors and non-tumoral parts) to the voxel level.

During the pre- and post-treatment studies, multi-modality images are acquired for VOI delineation and dose calculation. In the pre-treatment study, a SPECT/CT scan is performed shortly after administration of $^{99m}$Tc-labeled macro-aggregated albumin ($^{99m}$Tc-MAA) particles into selected branches of hepatic artery to mimic the activity distribution of $^{90}$Y microspheres in the liver [104]. Cone beam CT (CBCT) images are used for delineation of liver perfusion territories (LPTs) [109]. $^{18}$F-fluorodeoxyglucose ([$^{18}$F]FDG) or [$^{68}$Ga]DOTATATE PET/CT scans are performed for [$^{18}$F]FDG- or [$^{68}$Ga]DOTATATE-avid tumors. Contrast-enhanced and diffusion-weighted MR images are acquired for tumors that are not [$^{18}$F]FDG- or [$^{68}$Ga]DOTATATE-avid. Following the administration of $^{90}$Y-microspheres, a post-treatment study is performed to obtain the PET/CT or PET/MR images of the actual activity distribution inside the liver [122]. This PET image can be used to compute the absorbed dose in different VOIs for evaluation of the treatment irradiation distribution.

Registration of multi-modality images plays an essential role in information integration for SIRT dosimetry. Different methods of liver registration for intra-modality (CT or MR) and inter-modality images (CT and MR) have been studied. These methods include surface-based [90, 91, 92, 94], vessel-based [91, 92, 93], intensity-based [94, 95, 123], and segmentation-based registrations [98, 99]. Some studies combine both liver surfaces and vascular structures for better registration of tumors inside the liver [91, 92], since most tumors are found near vessels. Most intensity-based methods adopt mutual information for image similarity measurement. The segmentation-based methods use liver segmentations obtained either with histogram-based thresholding [98] or with a convolutional neural network (CNN) [99] for guidance of rigid or affine registration. For the application of liver registration in SIRT, Alsultan *et al.* used a rigid registration method in Simplicit90Y (Mirada Medical Ltd, Oxford, UK) to register contrast-enhanced CT images to low-dose CT images for efficacy evaluation of coil embolization to acquire intrahepatic redistribution [124]. Spahr *et al.* implemented a registration framework based on normalized gradient fields for liver registration of multi-modality images, and evaluated their algorithm through landmarks and deformation field analyses [97]. Nodari *et al.* used a multi-modality deformable registration algorithm in MIM SurePlan (v7.0.1; MIM software Cleveland, USA), performed by a trained medical physicist, to register the liver tumor contours from MR images to $^{99m}$Tc-MAA-SPECT/CT and $^{90}$Y-PET/CT [125]. Their study indicates that different tumor contours from anatomical and scintigraphic images have no significant impact on mean dose, and registering anatomical tumor contours to scintigraphic images is feasible for improving therapeutic strategy [125]. Besides, our in-house non-rigid liver registration regularized by a spring model was applied in our previous studies on the development of pre-treatment dosimetry [109] and evaluation of the predictive value of $^{99m}$Tc-MAA-based dose planning [126].

This study describes a (semi-)automatic segmentation-guided registration algorithm, and evaluates its performance for registering liver contours and anatomic landmarks and its influence on dose estimation for SIRT. Our in-house image-based registration algorithm was modified for guidance by

only CNN liver segmentations and by both CNN liver segmentations and manually delineated landmarks. (Semi-)automatic registration algorithms have the advantage of providing the physicians with integrated image information for more precise dosimetry, while not creating much cumbersome and time-consuming work in the clinical workflow. In our experience, the current lack of automation impedes the full use of multi-modality information. The registration can be challenging, because in the $^{99m}$Tc-MAA SPECT/CT protocol, the CT is acquired without contrast enhancement, and some MR images from the $^{90}$Y-PET/MR studies have severe artefacts. In this study, we wanted to assess (semi-)automatic liver registration and demonstrate its feasibility and value in a clinical context, so that it can contribute to a personalized and precise SIRT treatment with fewer manual interactions.

## 5.3 Materials and methods

### 5.3.1 Data

The training datasets for the CNN model contain 119 CT images from Liver Tumor Segmentation (LiTS17) and SLIVER07 challenges, 30 MR T1 images from the CHAOS challenge, and 50 SIRT CT and 110 MR T1 images from our hospital.

For the registration experiment, 49 CT and 26 MR images from 20 SIRT patients were selected according to several criteria. The selection procedure is presented in Figure 5.1. Since the algorithm is designed for registering both CT and MR images in the pre- and post-treatment studies to the $^{99m}$Tc-MAA CT[3], there should be at least one CT and one MR from these studies in addition to the CT image from the $^{99m}$Tc-MAA-SPECT/CT for each patient. Since thresholding a $^{99m}$Tc-MAA SPECT for lesions may result in overestimation of the tumor targeting performance, landmark delineation (including lesions) for the $^{99m}$Tc-MAA study was performed on its CT image. $^{99m}$Tc-MAA CTs in our clinical routine are not contrast enhanced. Therefore, it was important to select patients with visible landmarks in $^{99m}$Tc-MAA CTs. After patient selection, landmarks visible in all CT and MR images for each patient were chosen and manually delineated by a nuclear medicine physician. Most landmarks were lesions and some of them were vessel knots and metal clips. For evaluation of liver registration, liver contours were manually delineated by a trained researcher and then corrected by a radiographer with over 10 years of experience in liver delineation. For the hyper-parameter tuning of the registration algorithm, 25 CT and 14 MR images from 10 patients were randomly selected for training, 24 CT and 12 MR images from the other 10 patients were used for testing. The characteristics of the 20 SIRT patients are presented in Table 5.1.

[3] The term "$^{99m}$Tc-MAA CT" is a shortcut for "the CT image acquired during the $^{99m}$Tc-MAA SPECT/CT imaging session", and similar terms for the other multi-modal systems. This slight abuse of notation will be used throughout the dissertation to improve the readability.

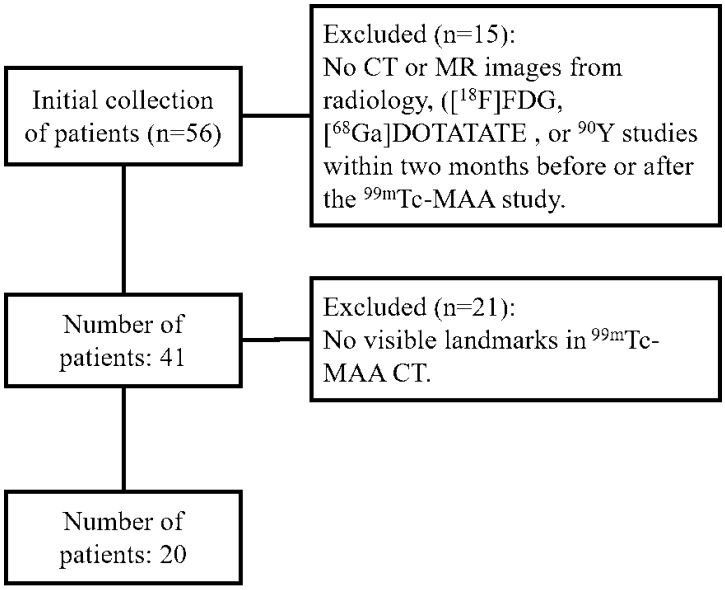



Figure 5.1 SIRT patient selection for the registration experiment.

Table 5.1 Characteristics of the SIRT patients for registration experiment

| Characteristics | Training | Test |
|---|---|---|
| No. of SIRT patients | 10 | 10 |
| Age (y), median [range] | 66.5 [46, 75] | 64.5 [25, 78] |
| Sex (female/male) | 5/5 | 4/6 |
| Weight (kg), median [range] | 71 [46,105] | 75 [46,95] |
| Height (m), median [range] | 1.70 [1.55, 1.74] | 1.70 [1.54, 1.78] |

All SIRT datasets for this research were evaluated at KU Leuven after approval by the Ethics Committee Research of UZ / KU Leuven.

### 5.3.2 CNN Structure for Liver Segmentation

The CNN model used in the paper adopts a U-net structure [73], which was developed for biomedical image segmentation. The U-net structure (see Figure 5.2), adapted for 3D images, consists of four resolution hierarchies with three skip connections to combine the high-level liver features from low resolution hierarchies with the detailed liver features from high resolution hierarchies. Our previous research demonstrated that the CNN segmentations resulted in good segmentation quality without consuming much time and work [127].

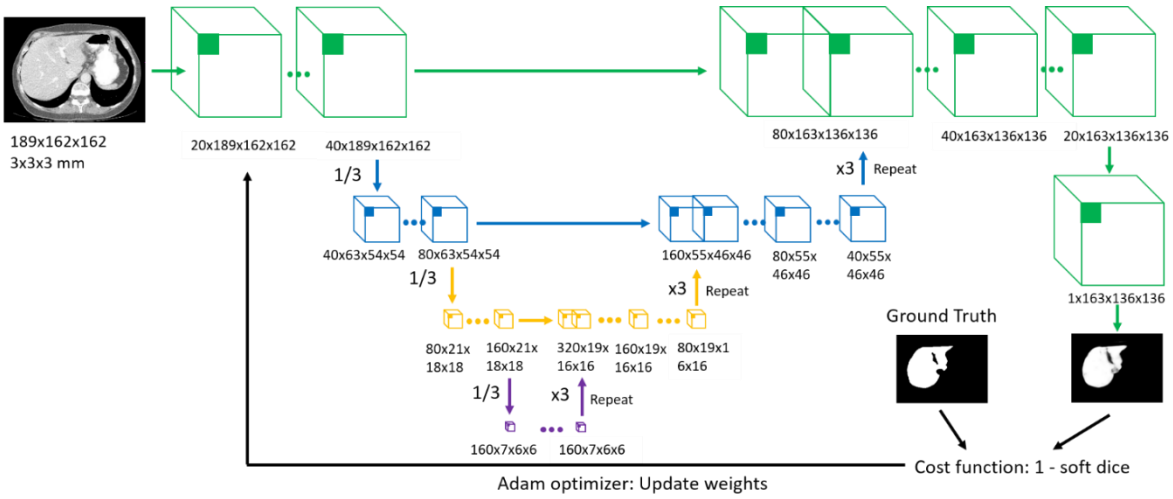



Figure 5.2 Overview of the CNN structure. The model consists of four resolution hierarchies with three skip connections. The input image has the voxel size of $3 \times 3 \times 3$ mm$^3$. The input image size is $189 \times 162 \times 162$ and the output image size is $163 \times 136 \times 136$. The up- and down-sampling rate is three and the up-sampling is implemented by repetition. The convolutional kernel size for all layers is $3 \times 3 \times 3$.

### 5.3.3 Segmentation-guided registration

The in-house image-based registration algorithm (see Figure 5.3) consists of affine registration followed by non-rigid registration. The image similarity metric for affine registration is computed by mutual information. The 12 affine parameters, generated after affine registration, are converted into an initial displacement field for non-rigid registration, where each element represents the displacement of the corresponding voxel in the floating image. The image similarity metric for non-rigid registration is also mutual information. To avoid topology-violating deformations during non-rigid registration, the voxel displacement is regularized through a spring model [96]. Each pair of neighboring voxels is assumed to be connected through a spring, which opposes distance changes. The spring rigidity is used to adjust the regularization power. Therefore, the image-based non-rigid registration algorithm minimizes the weighted sum of the image similarity loss ($L_I$) and the regularization loss ($L_R$). The computation and optimization of $L_I$ and $L_R$ are explained in detail in [96].

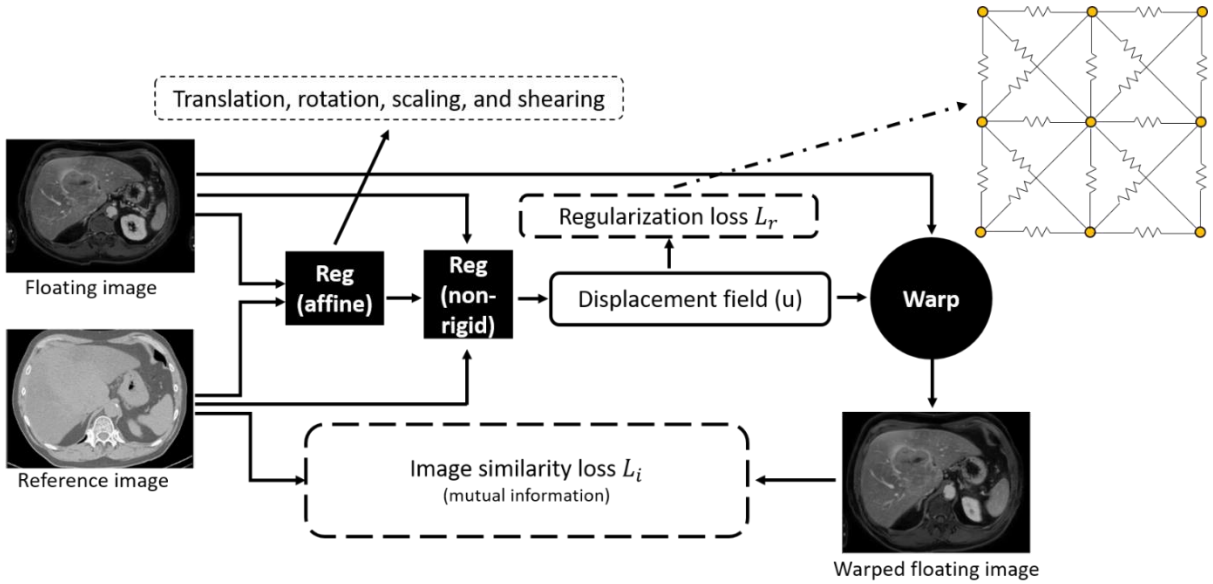



Figure 5.3 Overview of the in-house image-based registration algorithm.

This image-based registration algorithm was extended through guidance by CNN liver segmentations (see Figure 5.4). CNN liver segmentations for the reference and floating images are registered via affine transformation to provide initialization of the non-rigid registration. In the next stage, images and their CNN liver segmentations are simultaneously non-rigidly registered. The segmentation similarity loss ($L_S$) is computed as the sum of squared differences between the reference and warped segmentations, representing the segmentations as binary images. The computation and optimization of $L_S$ are the same as that of $L_I$ when using sum of squared distance as the similarity measurement, as described in [96]. The final loss of the non-rigid registration guided by CNN liver segmentations (CNN-guided) is the weighted sum: $w_I L_I + w_S L_S + L_R$.

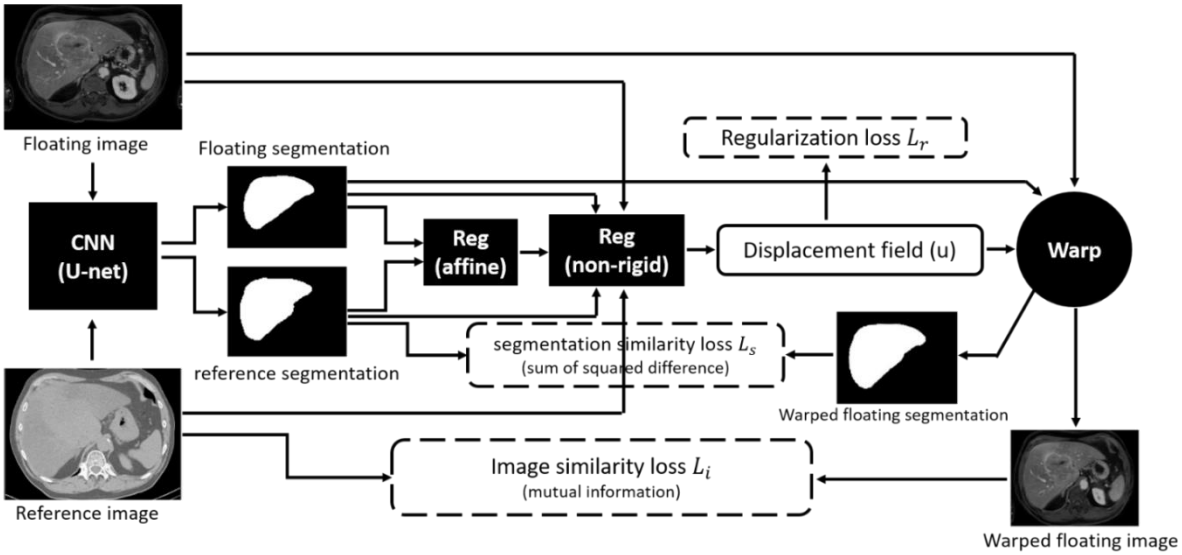



Figure 5.4 Overview of the registration algorithm guided by CNN liver segmentations.

To explore the possibility of landmark guidance for better lesion registration, a labeled map containing both CNN liver segmentations and landmarks as two classes replaces the CNN liver segmentations in the CNN-guided algorithm with other settings unchanged. This algorithm guided by both CNN liver segmentations and landmarks (CNN&LM-guided) is compared with the image-based and CNN-guided algorithms in the following experiments.

### 5.3.4 Experiments

**CNN liver segmentation.** The CNN model trained on CT and MR images from the public challenges and our hospital was implemented to generate automatic liver segmentations for the 49 CT and 26 MR images from 20 SIRT patients selected for the registration experiment. The results were evaluated through the dice similarity coefficient (DSC) between the CNN liver segmentation and the ground truth liver segmentation. The DSC quantifies the overlap between the segmentations, its definition and computation are explained in [112].

**Registration of multi-modality images.** For affine registration, the results from using either images or CNN liver segmentations were compared. For non-rigid registration, 1D ($w_S = 0$) and 2D grid searches were implemented to find the optimal weights without and with guidance of CNN liver segmentations. The non-rigid registration was initialized by either image-based or CNN-based affine registration. To evaluate the value of landmark guidance for better lesion registration, 2D grid searches were implemented to find the optimal weights for the CNN&LM-guided non-rigid registration initialized by the CNN&LM-based affine registration. Each liver registration was evaluated by the root mean squared distance (RMSD) of mean surface distance (MSD) between liver contours and mass center distance (MCD) between landmarks: $RMSD =$

$\sqrt{(MSD^2 + \sum_{i=1}^{N_{landmarks}} MCD_i^2)/(1 + N_{landmarks})}$. Mean surface distance is designed to measure the contour difference between two segmentations [113]. Since a landmark can appear with different shapes and volumes in images of different modality, mass center distance, which is independent of shape and volume changes, is used to evaluate landmark registration. The root mean square of RMSDs (RRMSD) from all liver registrations was used to evaluate each pair of weights. The optimal weights, with the lowest RRMSD, were found through grid searches on the training datasets for CT to CT and MR to CT registrations, respectively. After that, different registration settings were compared on the test datasets through RMSD and RRMSD.

**Dose estimation.** The CNN- and CNN&LM-guided registrations using the optimal weights were implemented to generate registered landmarks (including lesions) for the floating CT and MR images (registered floating landmarks) from the test datasets for comparison with the landmarks delineated on $^{99m}$Tc-MAA CTs (reference landmarks). A five-scale Likert score, with its criteria presented in Table 5.2, was used for grading the floating landmarks registered to the reference landmarks. Dose estimation using the registered floating lesions from the CNN- and CNN&LM-guided algorithms was compared with dose estimation using the reference lesions through the mean dose and the volume percentage receiving at least 70 (V70) and 100 Gy (V100) in the lesion. The absorbed doses of 70 and 100 Gy illustrate an intermediate and high tumor response probability. The injected activities prescribed for the left and right LPTs for each patient were used for dose estimation. LPTs are delineated on CBCTs in our clinical workflow. Dose estimation computed on the left and right LPTs requires registration of CBCT to $^{99m}$Tc-MAA-SPECT/CT, which introduces extra potential sources of registration errors. To focus on evaluating the influence of CT and MR registration on dose estimation in this study, the injected activities for the left and right LPTs are summed and distributed in a fractional uptake map generated by normalizing all counts of the $^{99m}$Tc-MAA SPECT within the manually delineated liver contour. The computation of the absorbed dose was based on the local deposition model.

Table 5.2 Likert score criteria for scoring landmark registration

| Score | Criteria |
|---|---|
| 1 | Major misalignment exists for the registration. Major impact on dosimetry is expected. Dosimetry results are deemed unreliable. |
| 2 | Pronounced misalignment exists for the registration. Substantial impact on dosimetry is expected. |
| 3 | Moderate misalignment exits for the registration. Moderate impact on dosimetry is expected. |
| 4 | Little misalignment exists for the registration. No significant impact on dosimetry is expected. |
| 5 | Near perfect alignment for the registration. No intervention is warranted and dosimetry is deemed reliable. |

## 5.4 Results

### 5.4.1 CNN liver segmentation

The DSCs between CNN liver segmentations of CT and MR images from 20 SIRT patients for registration and manual liver segmentations are presented in Figure 5.5. The median DSCs for CT and MR images are around 0.95 and 0.93, respectively. The DSCs for MR images have an outlier with a

very low value of 0.72. Through visual inspection, the CNN segmentation for this MR image misses a large part of the liver volume, and is incapable of registration guidance. Therefore, the case involving this MR image was excluded for MR to CT registration.

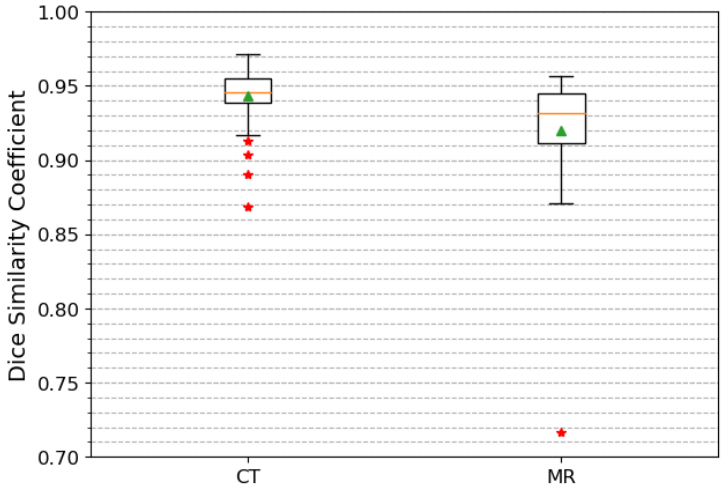



Figure 5.5 DSCs between manual liver segmentations and CNN liver segmentations of 49 CT and 26 MR images from the 20 SIRT patients selected for the registration experiment. The orange line, green triangle, and red stars represent median, mean, and outliers, respectively. The box corresponds to the first and third quartiles and the whiskers give the range (except for the outliers).

### 5.4.2 Registration of multi-modality images

The results of CT to CT registrations are presented in Figure 5.6, using the optimal weights for each deformable registration. The RRMSD for CNN-based affine registration is 3.1 mm (27%) smaller than that for image-based affine registration. The image-based non-rigid registration increases the RRMSD for CNN-based affine registration by 0.4 mm (5%). The RRMSD for the CNN-guided affine and non-rigid registration[4] is 1.0 mm (11%) smaller than that for image-based affine and non-rigid registration. The CNN&LM-guided affine and non-rigid registration[5] has a RRMSD 2.1 mm (26%) smaller than the CNN-guided one.

[4] "The CNN-guided affine and non-rigid registration" always refers to the CNN-based affine and CNN-guided non-rigid registration.

[5] "The CNN&LM-guided affine and non-rigid registration" always refers to the CNN&LM-based affine and CNN&LM-guided non-rigid registration.

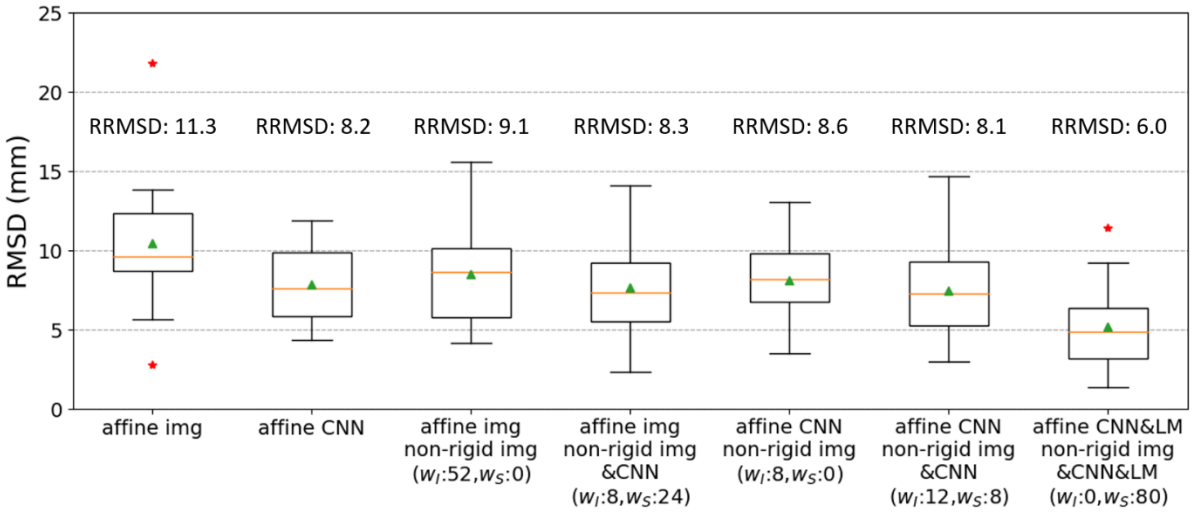



Figure 5.6 RMSDs and RRMSDs of the test datasets (14 floating CT images) for comparison of image-based, CNN-guided, and CNN&LM-guided CT to CT registrations.

The results of MR to CT registrations are presented in Figure 5.7, using the optimal weights for each deformable registration. The CNN-based affine registration has a RRMSD 3.4 mm (34%) smaller than the image-based one does. The optimal $w_I$ and $w_S$ for CNN-guided non-rigid registration initialized by CNN-based affine registration were both zero. Therefore, the optimal performance of CNN-guided affine and non-rigid registration is the same as that of the CNN-based affine registration. This indicates that both image-based and CNN-guided non-rigid registration degrade the results of CNN-based affine registration. The RRMSD for CNN&LM-guided affine and non-rigid registration is 1.4 mm (21%) smaller than that for CNN-based affine registration.

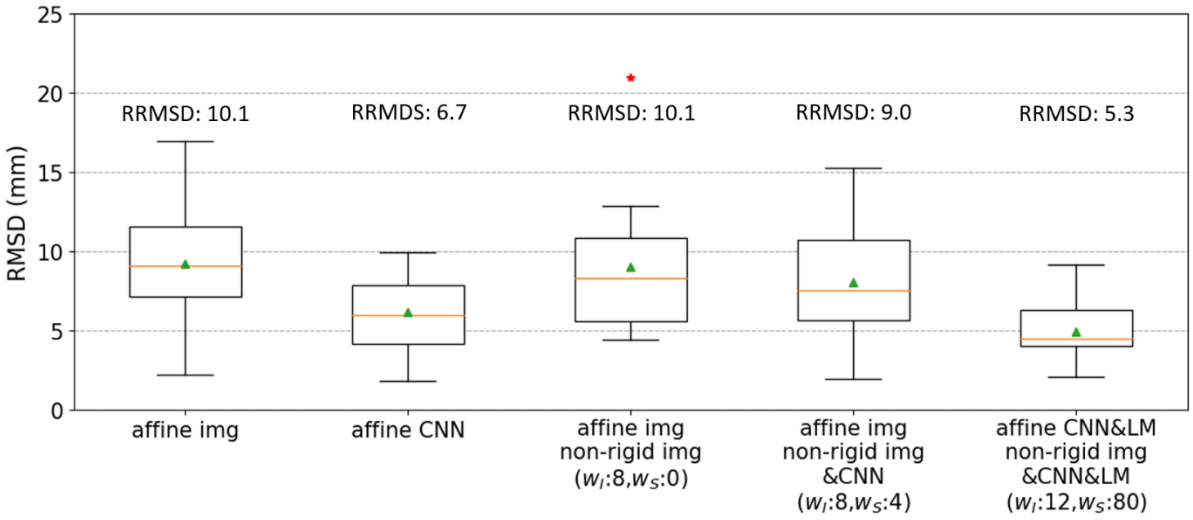



Figure 5.7 RMSDs and RRMSDs of the test datasets (12 floating MR images) for comparison of image-based, CNN-guided, and CNN&LM-guided MR to CT registrations.

### 5.4.3 Dose estimation

According to the results in the registration experiment, the optimal performance for CNN&LM- and CNN-guided CT to CT registration and CNN&LM-guided MR to CT registration were achieved when using both affine and non-rigid registrations, while the optimal performance for CNN-guided MR to CT registration was achieved when using only affine transformation. These settings using the optimal weights were used to generate the registered landmarks for dose estimation.

**Registration of CT to $^{99m}$Tc-MAA CT.** The Likert scores for registration of all landmarks (lesions, vessels, clips, etc) are presented in Figure 5.8a. Over two thirds of registered landmarks had a score of at least 4, which means little or almost no misalignment and insignificant impact on dosimetry. There were in total 29 lesions from 10 SIRT patients used for dose estimation. The relative difference of mean dose and the difference of V70 and V100 between the reference and registered floating lesions using the CNN&LM- and CNN-guided registrations are shown in Figure 5.9. There were 59% and 45% of lesions having an absolute relative difference of mean dose smaller than 10% for the CNN&LM- and CNN-guided registrations, respectively. Around 79% (76%) and 83% (69%) of lesions have an absolute V70 (V100) difference smaller than 10% for the CNN&LM- and CNN-guided registrations, respectively. As shown in Figure 5.10 , the mean dose, V70, and V100 for the floating lesions registered by both the CNN&LM- and CNN-guided algorithms had a strong correlation ($r \geq 0.95$) with the mean dose, V70, and V100 for the reference lesions.

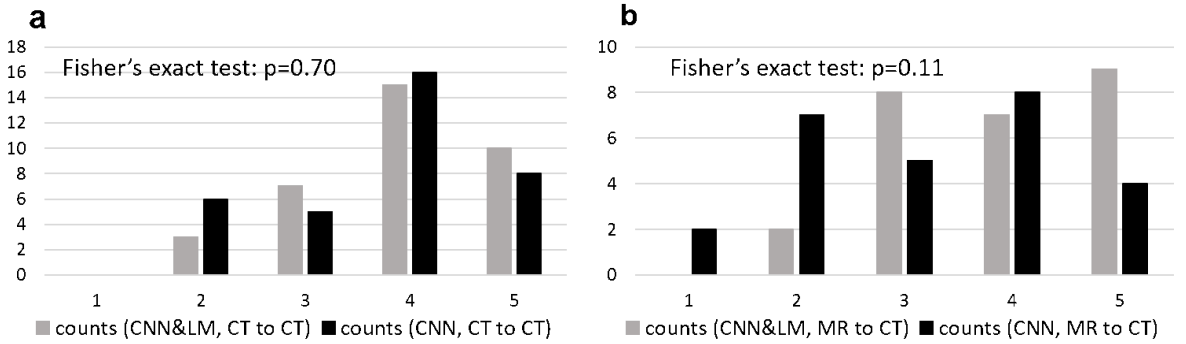



Figure 5.8 Likert scores for the registered landmarks of floating CT (a) and MR (b) images using the CNN&LM-guided and CNN-guided registrations.

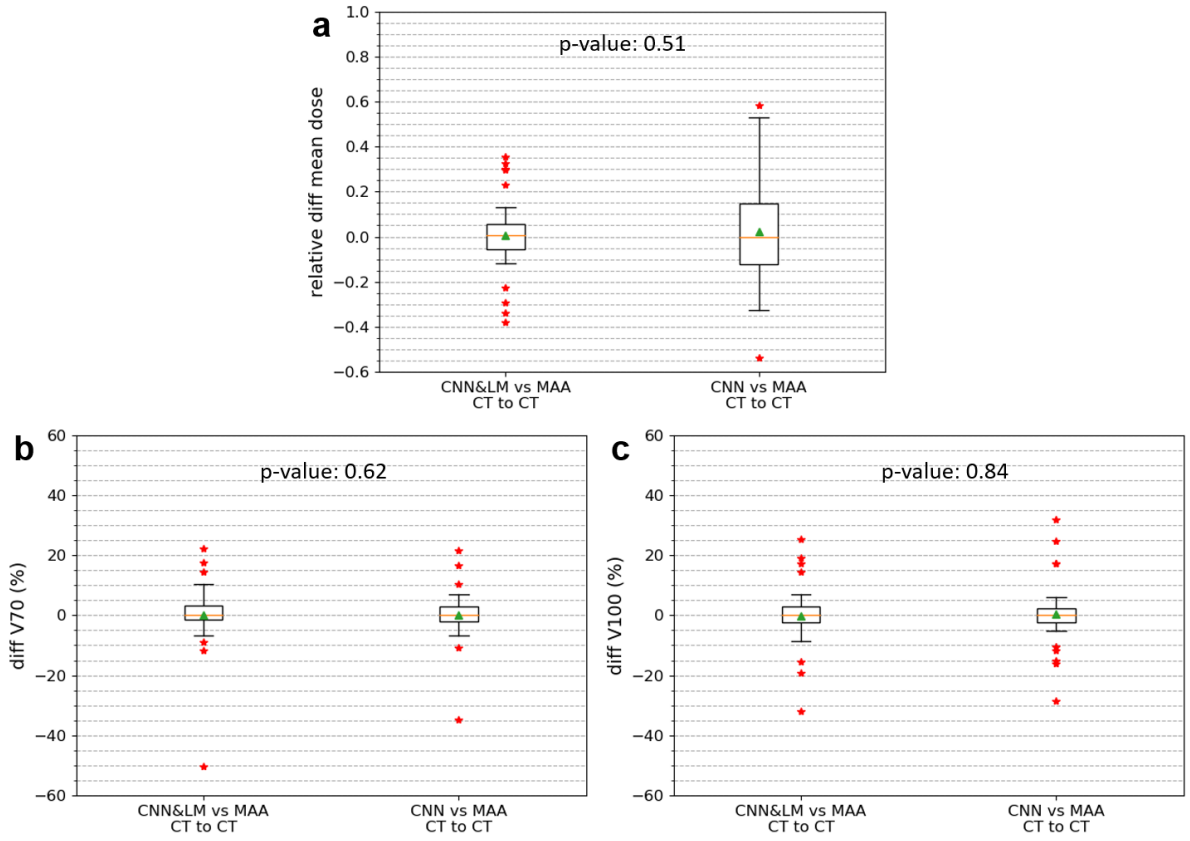



Figure 5.9 Relative difference of mean dose (a) and difference of V70 (b) and V100 (c) between the reference and registered floating lesions using the CNN&LM-guided or CNN-guided CT to CT registration. MAA: the reference

lesions. CNN&LM: the floating lesions registered by the CNN&LM-guided registration. CNN: the floating lesions registered by the CNN-guided registration. The relative difference of mean dose is computed by (mean dose (floating) – mean dose (reference) / mean dose (reference)). The difference of V70 and V100 is computed by (V70 or V100 (floating) – V70 or V100 (reference)).

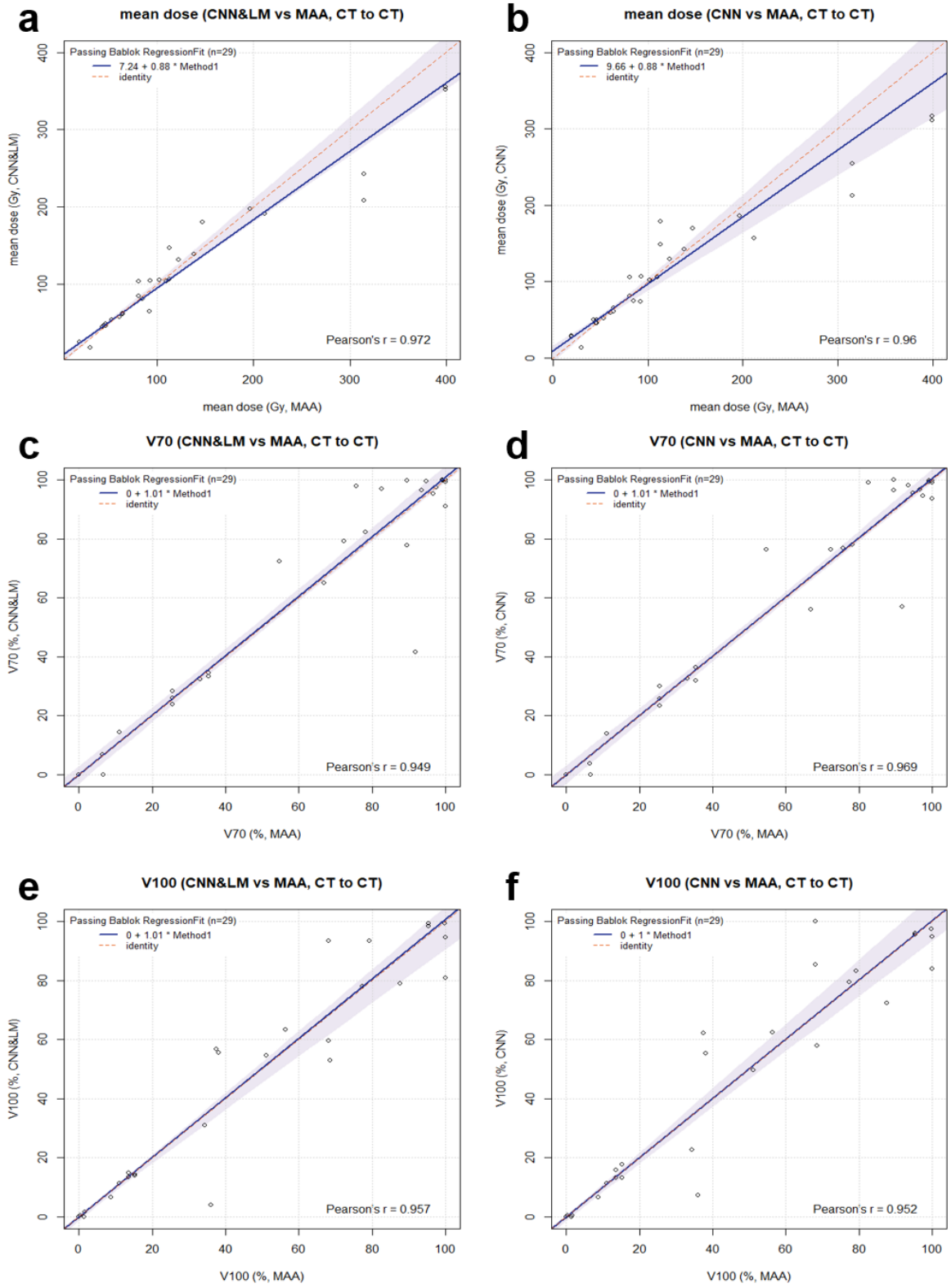



Figure 5.10 Passing-Bablok plots for the mean dose, V70, and V100 estimated on the floating CT lesions registered by the CNN&LM- or CNN-guided algorithm versus the mean dose, V70, and V100 estimated on the

reference lesions. MAA: the reference lesions. CNN&LM: the floating lesions registered by the CNN&LM-guided registration. CNN: the floating lesions registered by the CNN-guided registration.

**Registration of MR to $^{99m}$Tc-MAA CT.** According to the Likert scores for all registered landmarks shown in Figure 5.8b, around 92% and 63% of lesions registered by the CNN&LM- and CNN-guided algorithms had a score equal to or better than "moderate misalignment and impact on dosimetry", respectively. There were in total 23 lesions from 10 SIRT patients for dose estimation. The relative difference of mean dose and the difference of V70 and V100 between the reference and registered floating lesions are presented in Figure 5.11. Around 70% and 43% of lesions had an absolute difference of mean dose smaller than 10 % for the CNN&LM- and CNN-guided registrations, respectively. The CNN&LM- and CNN-guided registrations have around 70% (70%) and 61% (61%) of lesions with an absolute V70 (V100) difference smaller than 10%, respectively. The correlation of mean dose, V70, and V100 between the reference and registered floating lesions is presented in Figure 5.12. A weaker correlation ($r < 0.90$) of V70 and V100 for the CNN-guided registration is observed than the correlation of mean dose for the CNN&LM- and CNN-guided registrations and of V70 and V100 for the CNN&LM-guided registration.

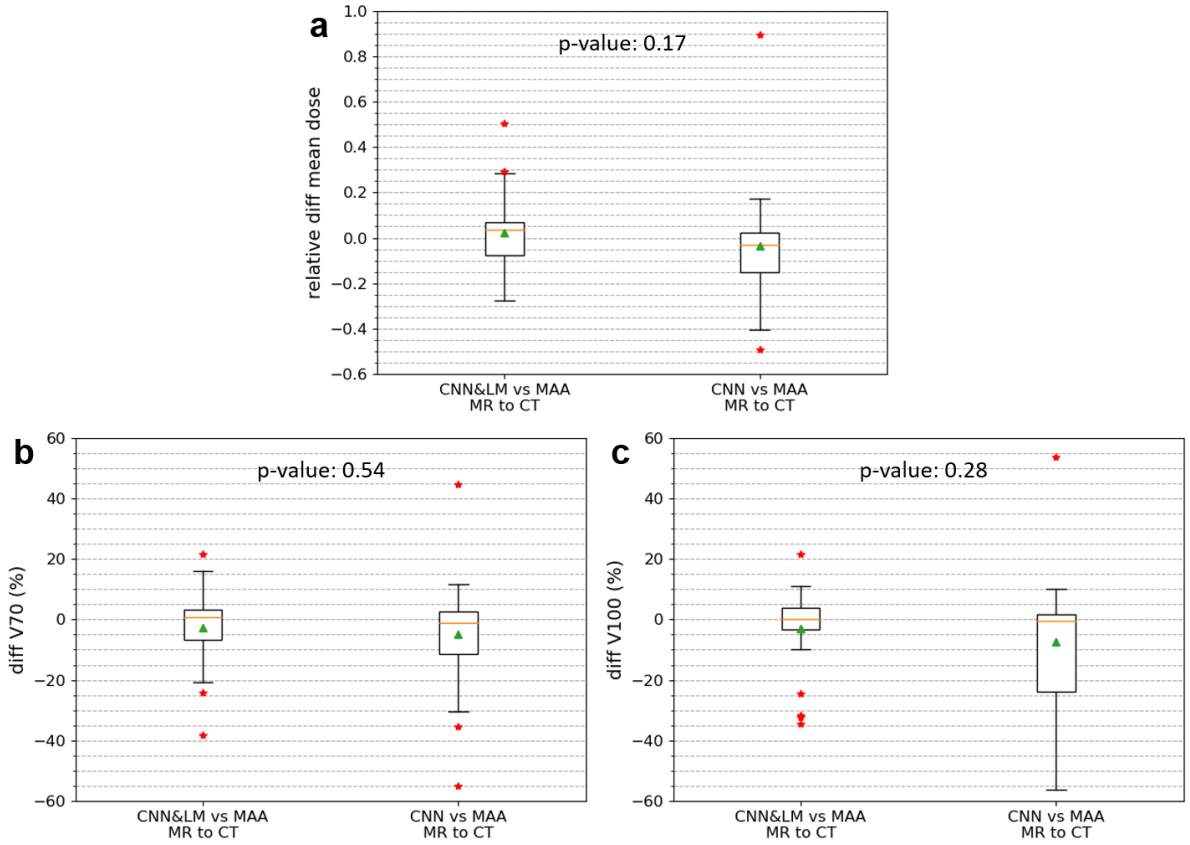



Figure 5.11 Relative difference of mean dose (a) and difference of V70 (b) and V100 (c) between the reference and registered floating lesions using the CNN&LM-guided or CNN-guided MR to CT registration. MAA: the reference lesions. CNN&LM: the floating lesions registered by the CNN&LM-guided registration. CNN: the floating lesions registered by the CNN-guided registration. The relative difference of mean dose is computed by (mean dose (floating) – mean dose (reference) / mean dose (reference)). The difference of V70 and V100 is computed by (V70 or V100 (floating) – V70 or V100 (reference)).

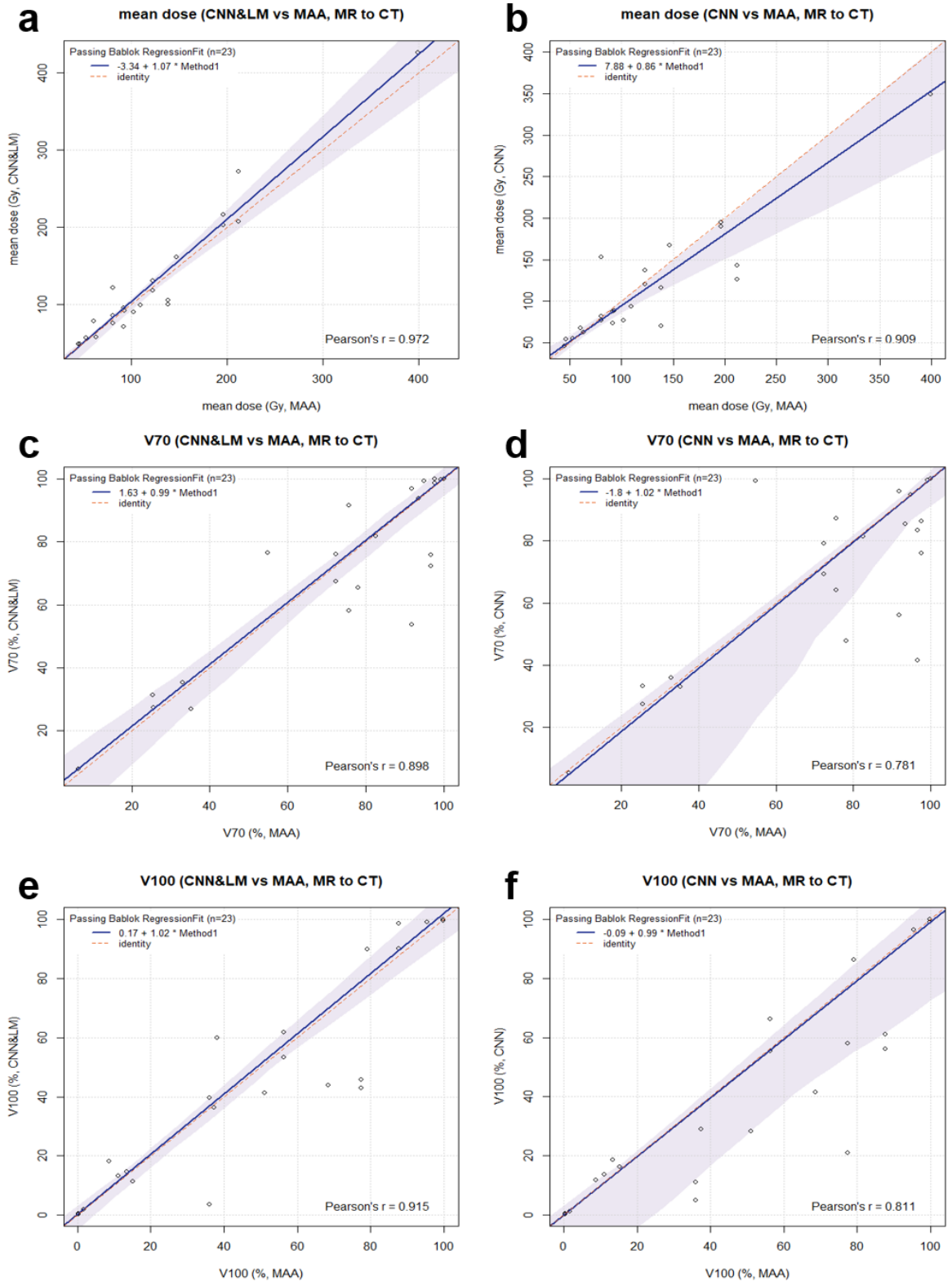



Figure 5.12 Passing-Bablok plots for the mean dose, V70, and V100 estimated on the floating MR lesions registered by the CNN&LM- or CNN-guided algorithm versus the mean dose, V70, and V100 estimated on the reference lesions. MAA: the reference lesions. CNN&LM: the floating lesions registered by the CNN&LM-guided registration. CNN: the floating lesions registered by the CNN-guided registration.

Some examples of the reference and floating lesions registered by the CNN&LM- and CNN-guided methods and their dose volume histograms (DVHs) are presented in Figure 5.13.

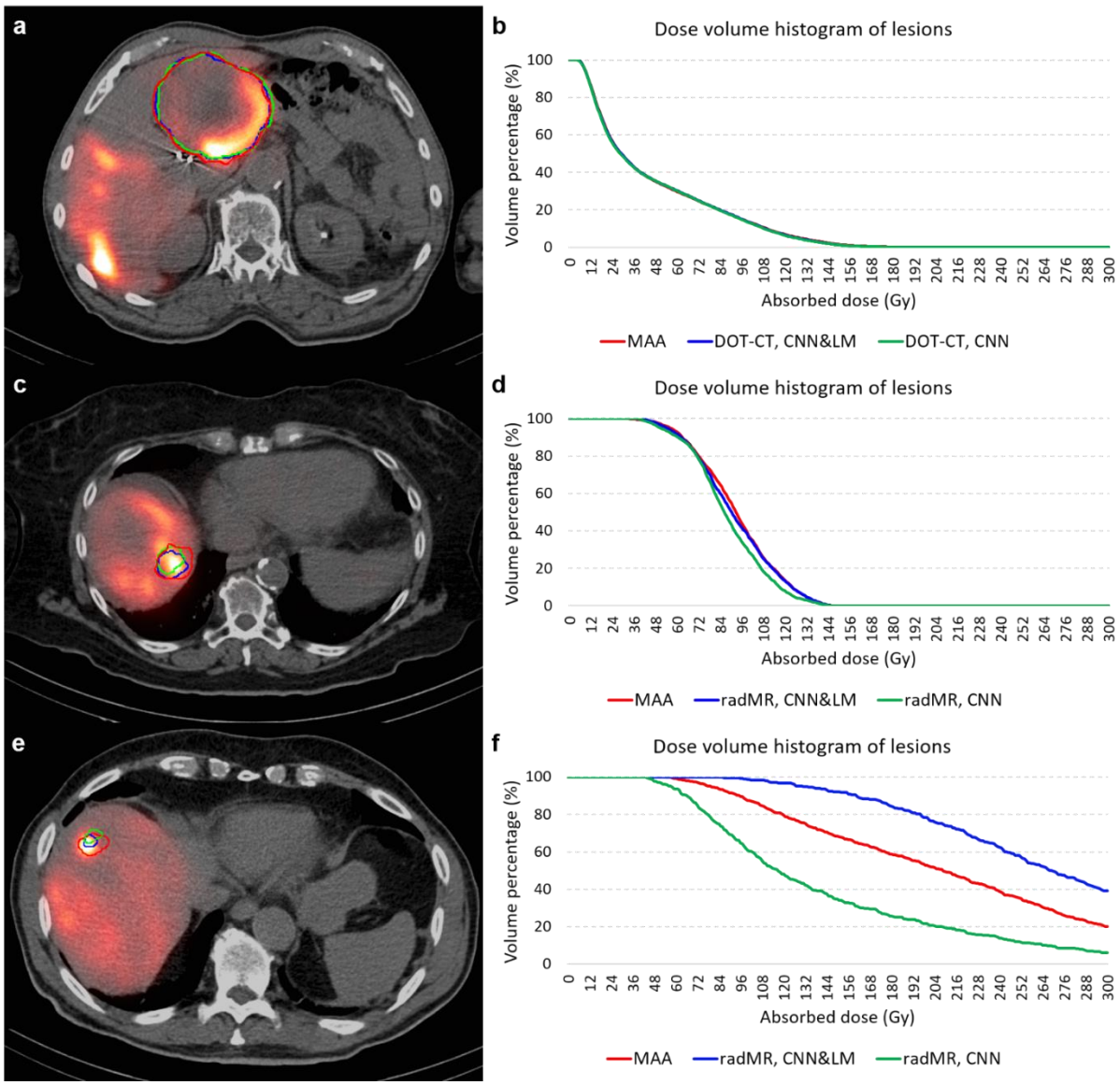



Figure 5.13 Examples of the reference and floating lesions registered by the CNN&LM- and CNN-guided methods and their dose volume histograms from three patients. The red, blue, and green contours represent the reference lesion (MAA), the floating lesion registered by the CNN&LM-guided method (CNN&LM), and the floating lesion registered by the CNN-guided method (CNN). **a, b** the registered floating lesions from the CT of the DOTATATE (DOT) study; CNN&LM: score is 5, relative difference of mean dose is -0.1%, the differences of V70 and V100 are 0.3% and -0.1%, CNN: score is 5, relative difference of mean dose is -0.9%, the differences of V70 and V100 are 0.2% and -0.4%. **c, d** the registered floating lesions from the radiology MR (radMR); CNN&LM: score is 3, relative difference of mean dose is -1.3%, the differences of V70 and V100 are -0.7% and -1.1%, CNN: score is 2, relative difference of mean dose is -5.2%, the differences of V70 and V100 are -1.0% and -8.4%. **e, f** the registered floating lesions from radMR; CNN&LM: score is 3, relative difference of mean dose is 28.4%, the differences of V70 and V100 are 2.5% and 11.1%, CNN: score is 2, relative difference of mean dose is -32.5%, the differences of V70 and V100 are -11.1% and -26.5%.

## 5.5 Discussion

### 5.5.1 Registration of multimodality images

The CNN-based affine registration of CT to CT and MR to CT improves the RRMSD by 3.1 mm (27%) and 3.4 mm (34%), respectively, compared with the image-based affine registration. This substantial decrease shows the advantage of using CNN liver segmentations for affine registration

without introducing non-affine deformation. The RRSMD increases by 0.4 mm (5%) when the CNN-guided affine registration is followed by the image-based non-rigid registration for CT images. The CNN-guided non-rigid registration of CT images decreases the RRMSD for the CNN-based affine registration by 0.1 mm (1%). This indicates the slightly negative influence of using only images (including non-contrast-enhanced $^{99m}$Tc-MAA CT) and the limited improvement of using CNN liver segmentations for non-rigid registration of CT images, given the good initialization provided by the CNN-based affine registration. The optimal weights found for the CNN-guided affine and non-rigid MR to CT registration were both zero indicating that using both images and CNN liver segmentations for non-rigid MR to CT registration could not improve the results from the CNN-based affine registration. This might be caused by relatively poorer CNN liver segmentations for MR images than for CT images. Some CNN liver segmentations for MR images were found to miss some low-intensity lesion regions. This can cause big misalignment in these regions due to liver surface matching guided by CNN liver segmentations during the non-rigid registration. Nevertheless, the CNN-guided registration improves the RRMSD by 1.0 mm (11%) and 3.4 mm (34%) for the floating CT and MR images compared with the image-based registration. Even if there might be errors in CNN liver segmentations compared to the "ground truth", these unedited CNN segmentations are still helpful for improving the image-based registration. This enables the automation of the liver-segmentation-guided registration without the need of extra manual correction for CNN liver segmentations.

Through landmark guidance, the RRMSD is decreased by 2.1 mm (26%) and 1.4 mm (21%) compared with the CNN-guided registration for the floating CT and MR images, respectively. Since we currently don't have an automatic lesion segmentation tool for CT and MR images, manually delineated landmarks were used for both registration guidance and evaluation. This might cause a self-fulfilling effect for registration evaluation. However, it does not change the feasibility of using landmarks for better registration, since manually delineated landmarks can be taken as perfect automatic lesion segmentations. This indicates that developing automatic lesion segmentation would be beneficial for registration guidance. Besides, manually delineated lesions approved by the physician are usable in the clinical context.

### 5.5.2 Dose estimation

A strong correlation ($r > 0.9$) of mean dose estimation existed between the reference and floating lesions registered by the CNN&LM- and CNN-guided registrations. Landmark guidance for the CNN-guided registration resulted in a smaller difference of mean dose for CT to CT registration than for MR to CT registration. Since mean dose is computed on the volume level, it appears less sensitive to contour changes than the voxel-level dosimetry.

For the voxel-level dosimetry, the CNN&LM- and CNN-guided registrations had around 79% (76%) and 83% (69%) of lesions with an absolute V70 (V100) difference between the reference and floating CT lesions smaller than 10%, respectively. A very strong correlation ($r \geq 0.95$) of V70 and V100 existed between the reference and floating CT lesions for the two methods. Landmark guidance for CT to CT registration made small improvement for the voxel-level dosimetry, which was also reflected in the Likert scores given by the physician. Around 70% (70%) and 61% (61%) of MR lesions had an absolute V70 (V100) difference smaller than 10% for the CNN&LM- and CNN-guided registrations, respectively. A weaker correlation ($r < 0.82$) of V70 and V100 was observed between the reference lesions and the floating MR lesions for the CNN-guided registration than for the CNN&LM-guided registration. Landmark guidance for MR to CT registration helped decrease the discrepancy of the voxel-level dose estimation caused by lesion registration.

It was found that the relative difference of mean dose and the difference of V70 and V100 were smaller than 10% for most lesions with a volume over 50 cc. It is reasonable that small lesions with a small shift can create a relatively large voxel change. Besides, lesions delineated on images of different modality can appear with diverse shape and volume, due to different lesion information expressed in

different images or tumor development. Small volumes and large shape and volume differences accounted for a V70 (V100) difference over 10% for 4 (3) out of 6 (7) and 3 (4) out of 5 (9) floating CT lesions registered by the CNN&LM- and CNN-guided algorithms and for 4 (2) out of 7 (7) and 5 (3) out of 9 (9) floating MR lesions registered by the CNN&LM- and CNN-guided algorithms. Good lesion registration does not ensure small difference of dose estimation, since the small size and large shape and volume difference are the other two critical factors with significant impact on dose estimation. It is difficult to eliminate the shape and volume difference between lesions delineated on different images, since each modality reflects a different aspect of lesion appearance. Therefore, it is beneficial to co-register all multi-modality images for joint lesion delineation by the physician, to approach the ground truth delineation by making full use of all information.

Poor lesion registration does not necessarily lead to significant changes of dose estimation. As presented in Figure 5.13c and Figure 5.13d, the lesion (green) of radiology MR registered by the CNN-guided method, scored with 2, does not have a good overlap with the reference lesion (red), while both lesion contours include most of the high-uptake region. The relative mean dose difference and the V70 and V100 difference between the reference and registered lesions are -5.2%, -1.0%, and -8.4%, respectively. As long as the reference and registered lesions include a similar area of high- and low-uptake regions, the difference of dose estimation can be insignificant despite of poor registration. This indicates that the volume percentage does not necessarily reflect the true energy deposition for each voxel.

In our standard workflow, the liver and lesions are manually delineated on the anatomic image by using the delineation tools from a clinical software package used by the physician for SIRT planning. After that, the delineated volumes of interest (VOIs) are mapped to the $^{99m}$Tc-MAA SPECT/CT by using the manual or semi-automatic registration tools of the software to register the anatomic image to the $^{99m}$Tc-MAA CT. To shorten the processing time, the physician delineates the VOIs directly on the $^{99m}$Tc-MAA SPECT in selected cases, obviating the need for registration. However, proper registration of anatomical images to the $^{99m}$Tc-MAA SPECT/CT allows lesion delineation on the anatomical images and correlation to the SPECT findings, as recommended in recent international guidelines for SIRT [128]. The standard workflow requires the physician's interaction during the entire process. In general, the time to complete the standard workflow is around 30 to 45 min. The segmentation-guided workflow consists of liver segmentation, registration, and lesion delineation. To facilitate clinical application and evaluation of these new tools, we have incorporated the entire workflow into the clinical software platform. Liver segmentation is fully automated by the CNN. It takes no more than 5 min for the trained CNN model to generate one liver segmentation by a CPU-based computation server. After that, the CNN liver segmentations are checked and corrected, if necessary, by the physician to ensure its usability for registration guidance, which takes 1 to 5 min. The segmentation-guided registration algorithm is performed by a CPU-based server without parallel computation, which takes around 15 min for each registration in general. Since the registration workflow is fully automated without manual interaction needed, the processing time is acceptable for routine clinical use. It could be speeded up by implementing parallel computation. Lesion delineation is manually performed by using the delineation tools of the clinical software, which takes around 10 min and needs to be automated in the future. In total, the segmentation-guided workflow can take around 30 min. The automated processing takes 20 min, which does not need the physician's interaction. This makes the segmentation-guided workflow a useful tool for the physician.

In summary, the performance of the CNN&LM- and CNN-guided registrations makes them useful tools for SIRT treatment planning and verification. The deployment of these semi- and automatic registration tools would allow for dose prediction and measurement based on multi-modality images without introducing much manual interaction and workload, which currently impede the application of image analysis tools in the clinical workflow. The pre- and post-treatment studies contain many images with relatively poor quality, including non-contrast-enhanced CTs from the $^{99m}$Tc-MAA study and MR

with severe shading or bias artifacts from the $^{90}$Y-PET/MR. Nevertheless, these registration algorithms can produce reasonably good results for these low-quality images giving them practical value for clinical application. Based on these results, we will study the development of an automatic liver lesion segmentation method for fully automatizing the CNN&LM-guided registration. The clinical influence of these registration methods remains to be fully evaluated in a daily SIRT workflow from a volume-level and voxel-level perspective.

## 5.6 Conclusion

Registration guidance using CNN liver segmentations and landmarks greatly improved the performance of the in-house image-based registration. The CNN&LM- and CNN-guided registrations for CT and MR images can be used for the volume-level dosimetry, since the mean doses obtained from the reference and floating lesion contours were very similar. A small V70 and V100 change for most lesions of the floating CT images using the CNN&LM- and CNN-guided registrations demonstrates the feasibility of their application to the voxel-level dosimetry with the physician's checkup. Landmark guidance is needed for the CNN-guided MR to CT registration to be applied to the voxel-level dosimetry. As a result, the CNN&LM- and CNN-guided registration algorithms could become valuable semi- and automatic tools applicable for SIRT dosimetry based on integration of multi-modality information.

# 6. Comparison of registration errors and dose estimation difference between the segmentation-guided registration algorithm and the commercial registration tool

## 6.1 Abstract

**Purpose:** A commercial medical imaging software (MIM software Cleveland, USA) was introduced into the hospital to provide solutions for radiation oncology, radiology, nuclear medicine, and so on during the Ph.D. project. The current study compared the performance of the segmentation-guided registration introduced in Chapter 5 with the performance of the multi-modality deformable image registration (DIR) tool of MIM introduced in the same period for liver registration of CT and MR. **Methods:** The multi-modality DIR of MIM was tuned and evaluated with the same datasets (49 CT and 26 MR images from 20 SIRT patients) used in Chapter 5. The contours of the liver and landmarks (including lesions) manually delineated for the study of Chapter 5 were used to measure the registration error and dose estimation difference caused by registration. The registration error of the multi-modality DIR was compared with the error of the registrations guided by CNN liver segmentation (CNN-guided) and by CNN liver segmentations and manual landmark segmentation (CNN&LM-guided), using the metric RRMSD introduced in Section 5.3.4. Mean dose and volume percentage receiving at least 70 Gy (V70) in the lesions were calculated based on the $^{99m}$Tc-MAA SPECT for dose estimation. The lesion contours delineated on the floating CT and MR images were registered to $^{99m}$Tc-MAA CT (reference) using the CNN-guided registration, CNN&LM-guided registration, and the multi-modality DIR of MIM. The dose metrics estimated with the registered lesion contours were compared with the metrics estimated with the lesion contours delineated on the $^{99m}$Tc-MAA CT to evaluate the dose estimation difference caused by different registration methods. **Results:** The RRMSDs for the CNN-guided, CNN&LM-guided, and MIM's multi-modality DIR methods were 8.1 mm, 6.0 mm, and 6.3 mm for the CT to CT registration, and 6.7 mm, 5.3 mm, and 5.0 mm for the MR to CT registration. The quartiles for the relative mean dose difference (the V70 difference) between the reference and registered lesions [25th, 75th] are as follows: [-12.3% (-2.1%), 14.8% (2.9%)], [-5.5% (-1.3%), 5.6% (3.4%)], and [-8.0% (-1.6%), 10.2% (2.5%)] for the CNN-guided, CNN&LM-guided, and multi-modality DIR methods for CT to CT registration, [-15.1% (-11.3%), 2.4% (2.5%)], [-7.7% (-6.6%), 7.0% (3.1%)], and [-6.5% (-3.8%), 10.2% (3.6%)] for the CNN-guided, CNN&LM-guided, and multi-modality DIR methods for MR to CT registration. **Conclusion:** The multi-modality DIR tool of MIM achieved comparable performance with the CNN&LM-guided registration method in terms of the registration error and dose estimation difference. The registration tool proved its value for clinical application in SIRT.

**Key Words:** selective internal radiation therapy (SIRT); liver registration; multi-modality images; commercial registration tool

## 6.2 Introduction

Radiology, nuclear medicine, and radiation oncology are medical disciplines highly relying on medical imaging. In recent years, medical imaging software companies grew rapidly and provide solutions for image visualization, image analysis, treatment planning and verification, and so on. The companies collaborate with the hospitals and institutions to obtain medical data, medical expertise, and user feedback to upgrade their software.

MIM software (MIM Software Inc., Cleveland, OH) became the central processing system of the nuclear medicine department of our hospital around four years ago. It provided a series of tools for image visualization, image fusion, semi-automatic image annotation, image registration, and dosimetry analysis. There were some studies using the deformable registration tool of MIM to perform CT registration. Calusi *et al*. used CT scans of two phantoms mimicking different degrees of tumor shrinking and anatomical bending to evaluate the performance of the deformable image registration (DIR) tool [129]. The HU values of the internal structures in one CT scan were artificially modified to create different contrast levels. It was found that the registration error increased with increasing bending and volume difference and decreasing contrast. Fukumitsu *et al*. compared the registration error obtained when aligning the liver CT images acquired before and after treatment using the DIR tools from MIM and Velocity AI (Velocity Medical Solutions, GA) [130]. The two DIR tools achieved comparable performance on liver registration. The registration error of both tools decreased for contrast-enhanced CT images. The DIR tool of MIM has been upgraded to perform multi-modal image registration in the past years and its potential for application to registration of CT and MR images was not yet evaluated in a clinical context.

This study compares the performance of the segmentation-guided registration methods introduced in Chapter 5 and the multi-modality deformable image registration (DIR) tool of MIM with regard to the errors of liver registration for CT and MR images and their influence on dose estimation for SIRT. Relatively large registration errors do not necessarily lead to a large difference in dose estimation, and relatively small registration errors could lead to large dose estimation differences. As long as the dose estimation difference caused by registration errors is acceptable, the registration method can be applied in SIRT. Therefore, the influence of registration errors on dose estimation is critical for evaluating the clinical value of the registration methods. The liver and landmark contours manually delineated for the study in Chapter 5 were used to evaluate the registration. Through this study, we wanted to assess the value of the multi-modality DIR tool of MIM for clinical application with regard to registration of CT and MR images.

## 6.3 Materials and methods

### 6.3.1 Data

The MIM DIR tool was tuned and evaluated with the same datasets described in Section 5.3.1, which included 49 CT and 26 MR images from 20 SIRT patients. The datasets were randomly divided into 25 CT and 14 MR images from 10 patients for training, and 24 CT and 12 MR images from the other 10 patients for testing, similar to the division in Section 5.3.1. Each patient contained one $^{99m}$Tc-MAA CT as the reference image (the image that other images are registered to) and at least one CT and MR as the floating images (the images registered to the reference image).

### 6.3.2 Multi-modality deformable image registration tool of MIM

MIM developed a multi-modality DIR tool mainly for MR to CT registration. However, it is a general deformable registration method that can be used for CT to CT, MR to MR, and CBCT to CT registrations. It uses a similarity scoring metric as a feature to match voxels in two images, and diffusion regularization to constrain the deformation. This tool computes high-dimensional feature descriptors by evaluating each voxel in the context of its neighbouring voxels and maximizes the correspondence of the feature descriptors. The MIM DIR tool has a single parameter: the regularization power is controlled by a smoothness factor, which ranges from 0 to 2. When the smoothness factor increases, the voxels have less displacement freedom to account for local changes in anatomy.

### 6.3.3 Experiments

**Registration of multi-modality images.** The multi-modality DIR tool was used to register the floating CT and MR images to the reference image ($^{99m}$Tc-MAA CT) in the training datasets by using a set of smoothness factors ranging from 0 to 2 with an interval of 0.2. The registration of the manually delineated liver and landmarks for each floating image was evaluated by the RMSD that was described in Section 5.3.4. The registration performance for all floating images of the same modality (CT or MR) using the same smoothness factor was evaluated with the RRMSD that was described in Section 5.3.4. The smoothness factors that achieved the highest RRMSD were used for registering the floating CT and MR images to the reference CT image in the test datasets, respectively. The RMSD and RRMSD for the test datasets were used to compare the performance of the registration guided by CNN liver segmentations (CNN-guided), the registration guided by CNN liver segmentations and manually delineated landmarks (CNN&LM-guided), and the multi-modality DIR tool of MIM.

**Dose estimation.** The mean dose and the volume percentage that receives at least 70 Gy (V70) and 100 Gy (V100) were calculated within the lesion contours based on the $^{99m}$Tc-MAA SPECT for the absorbed dose estimation. The same injected activities described in Section 5.3.4 were used for the dose calculation. The CNN-guided registration, the CNN&LM-guided registration, and the multi-modality DIR tool of MIM with the optimal performance on the training datasets were used to register lesions manually delineated on the floating CT and MR images to the $^{99m}$Tc-MAA CT for the test datasets. The $^{99m}$Tc-MAA CT and $^{99m}$Tc-MAA SPECT were automatically aligned during the imaging session. Therefore, the registered lesions can be directly mapped to the $^{99m}$Tc-MAA SPECT. The dose estimation metrics computed by using the registered lesions through the three registration methods were compared with the metrics computed by using the lesions manually delineated on the $^{99m}$Tc-MAA CT to evaluate the dose estimation difference caused by different registration methods.

The metrics for measuring registration errors and dose estimation difference by using different registration methods were compared through the statistical tests mentioned in Section 4.3.4.

## 6.4 Results

### 6.4.1 Registration of multi-modality images

No evident change of RRMSD for the training datasets was observed for both CT to CT and MR to CT registration using the multi-modality DIR of MIM when the smoothness factor increased from 0 to 2. Therefore, the default smoothness factor of 0.5 was used for the test datasets. The results of CT to CT registration are presented in Figure 6.1 for the CNN-guided registration, the CNN&LM-guided registration, and the multi-modality DIR with the hyperparameters achieving the optimal performance on the training datasets. For CT to CT registration, the RRMSD for the multi-modality DIR is 1.8 mm smaller than that for the CNN-guided registration ($p = 0.005$) and 0.3 mm larger than that for the CNN&LM-guided registration ($p = 0.89$). For MR to CT registration, the RRMSD for the multi-modality DIR is 1.7 mm smaller than that for the CNN-guided registration ($p = 0.003$) and 0.3 mm smaller than that for the CNN&LM-guided registration ($p = 0.41$). There is no statistically significant difference in the registration error between the CNN&LM-guided registration and the multi-modality DIR of MIM for both CT to CT and MR to CT registrations.

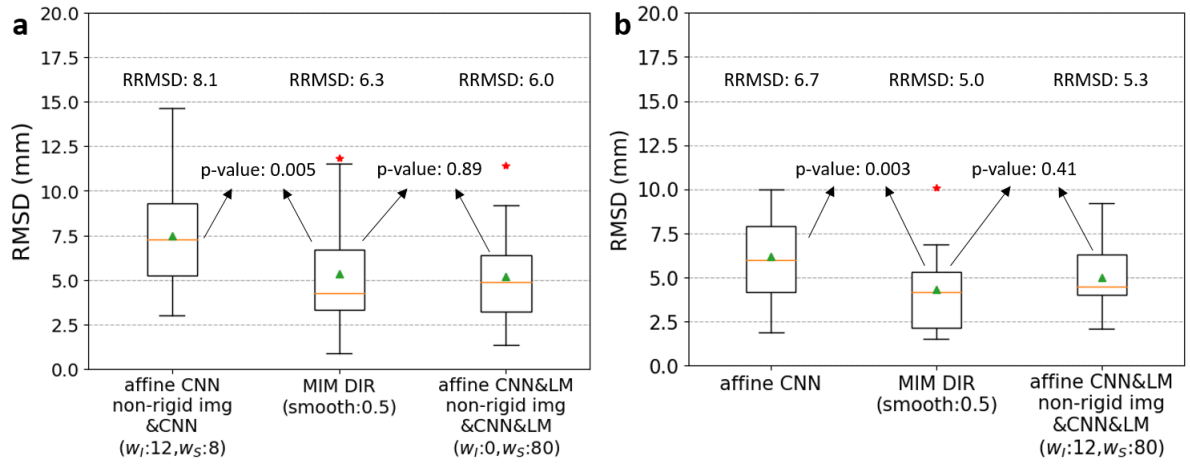



Figure 6.1 (a) RMSDs and RRMSDs of the test CT datasets (14 floating CT images) for comparison of the CNN-guided registration, and the CNN&LM-guided registration, and the multi-modality DIR of MIM. (b) RMSDs and RRMSDs of the test MR datasets (12 floating MR images) for comparison of the CNN-guided registration, the CNN&LM-guided registration, and the multi-modality DIR of MIM. The green triangle, the yellow line, and the red stars represent the mean, median, and outliers.

## 6.4.2 Dose estimation

**Registration of CT to $^{99m}$Tc-MAA CT.** There were in total 29 lesions delineated on both the $^{99m}$Tc-MAA CT and the floating CT images from the 10 SIRT patients for testing. The relative difference of mean dose and the V70 and V100 difference between the reference lesions and the floating lesions registered through the CNN&LM-guided registration, the CNN-guided registration, and the multi-modality DIR of MIM are presented in Figure 6.2. Around 59%, 45%, and 45% of lesions had an absolute relative mean dose difference smaller than 10% for the CNN&LM-guided registration, the CNN-guided registration, and the multi-modality DIR of MIM, respectively. The CNN&LM-guided registration, the CNN-guided registration, and the multi-modality DIR of MIM had 79% (76%), 83% (69%), and 83% (69%) of lesions with an absolute V70 (V100) difference smaller than 10%, respectively. As is presented in Figure 6.3, a strong correlation ($r \geq 0.90$) for the mean dose, V70, and V100 existed between the floating CT lesions registered by the three registration methods and the reference lesions.

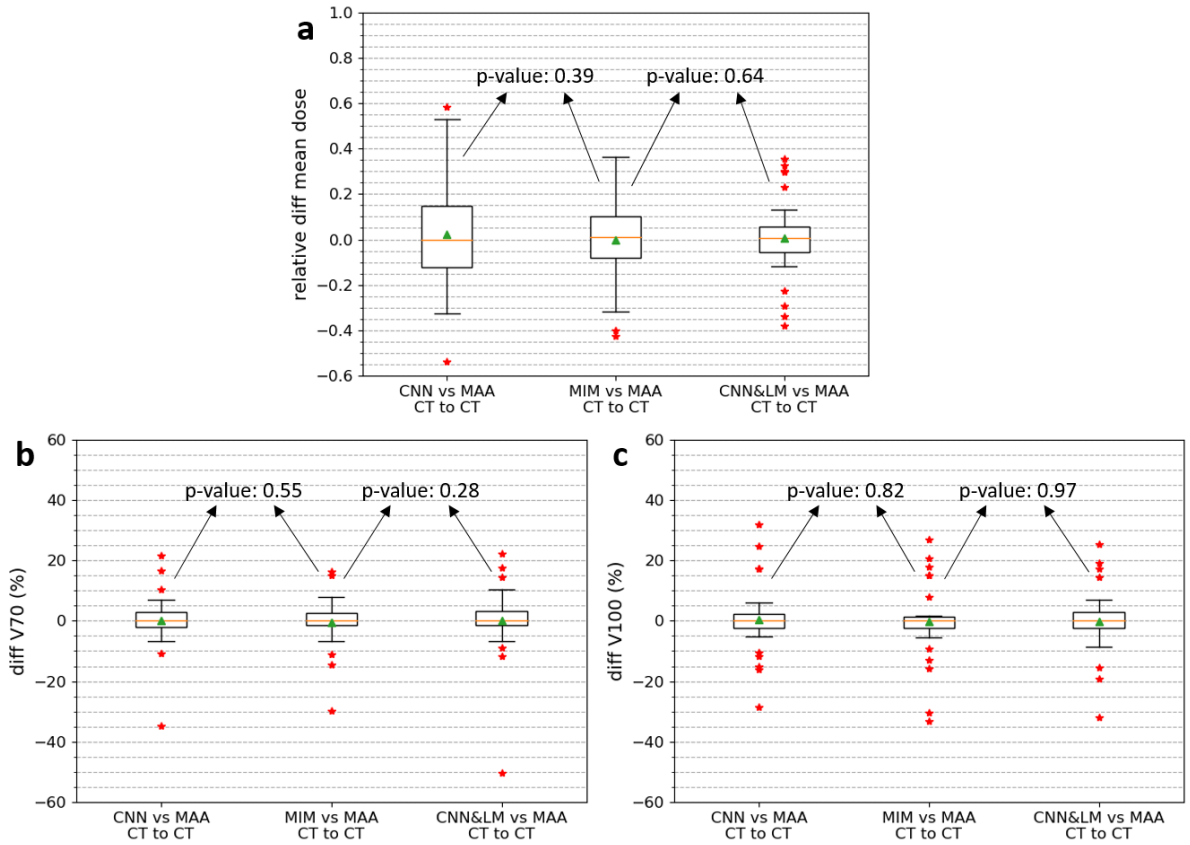



Figure 6.2 Relative difference of mean dose (a) and difference of V70 (b) and V100 (c) between the reference and registered floating lesions using the CNN&LM-guided registration, the CNN-guided registration, and the multi-modality DIR of MIM for the floating CT images. MAA: the reference lesions. CNN&LM: the floating lesions registered by the CNN&LM-guided registration. CNN: the floating lesions registered by the CNN-guided registration. MIM: the floating lesions registered by the multi-modality DIR of MIM. The relative difference of mean dose is computed by (mean dose (floating) – mean dose (reference)) / mean dose (reference). The difference of V70 and V100 is computed by (V70 or V100 (floating)) – (V70 or V100 (reference)). The green triangle, the yellow line, and the red stars represent the mean, median, and outliers.

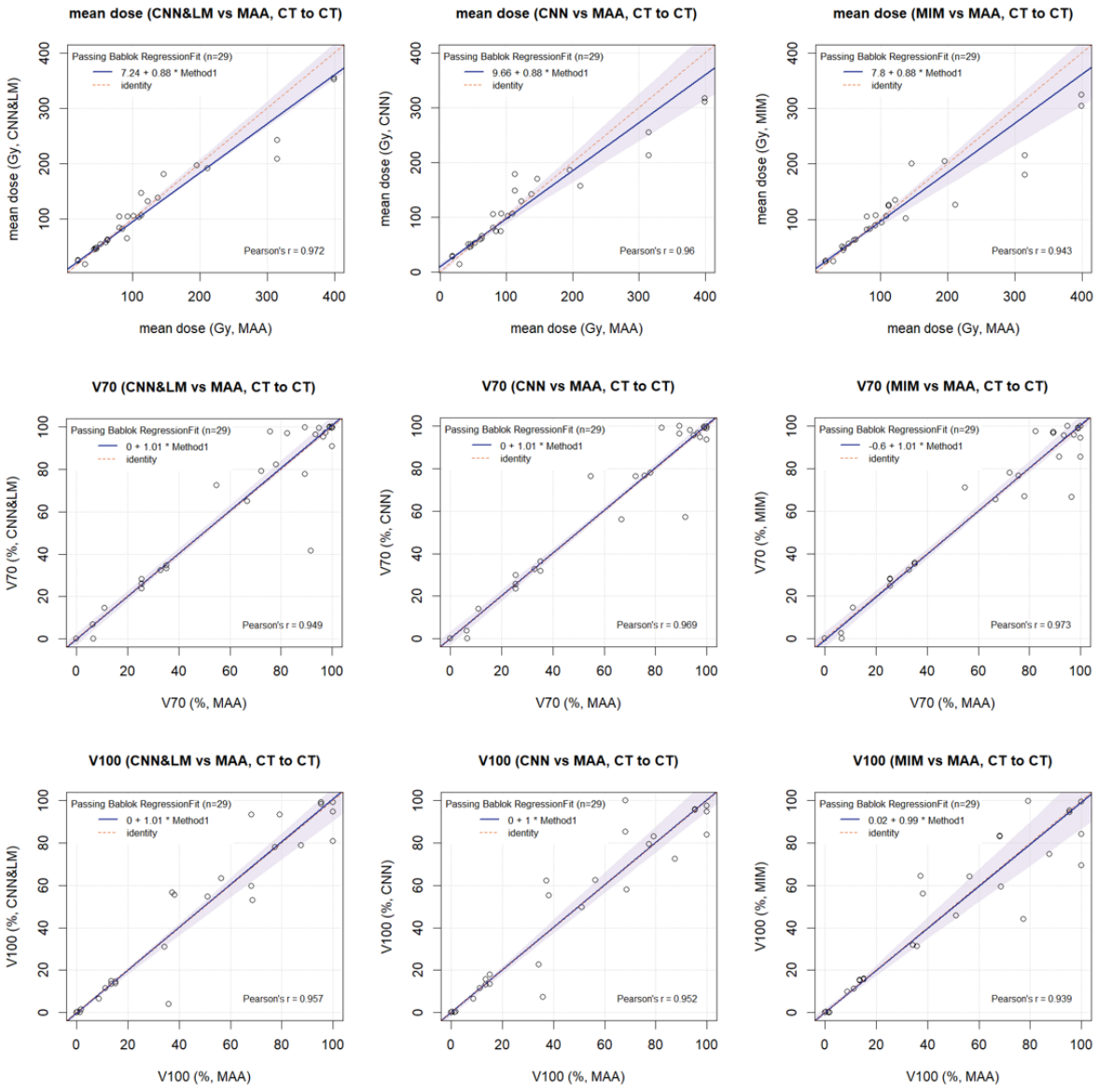



Figure 6.3 Passing-Bablok plots for the mean dose, V70, and V100 estimated on the floating CT lesions registered by the CNN&LM-guided registration, the CNN-guided registration, and the multi-modality DIR of MIM and the mean dose, V70, and V100 estimated on the reference lesions. MAA: the lesions delineated on the $^{99m}$Tc-MAA CT. CNN&LM: the lesions delineated on the floating CT and registered by the CNN&LM-guided registration. CNN: the lesions delineated on the floating CT and registered by the CNN-guided registration. MIM: the lesions delineated on the floating CT and registered by the multi-modality DIR of MIM.

**Registration of MR to $^{99m}$Tc-MAA CT.** There were in total 23 lesions delineated on both the $^{99m}$Tc-MAA CT and the floating MR images from the 10 SIRT patients for testing. The relative difference of mean dose and the V70 and V100 difference between the reference lesions and the floating MR lesions registered through the three registration methods are presented in Figure 6.4. There were 61%, 43%, and 61% of lesions having an absolute relative mean dose difference smaller than 10% for the CNN&LM-guided registration, the CNN-guided registration, and the multi-modality DIR of MIM, respectively. Around 70% (70%), 61% (61%), and 70% (74%) of lesions had an absolute V70 (V100) difference smaller than 10% for the CNN&LM-guided registration, the CNN-guided registration, and the multi-modality DIR of MIM, respectively. As is presented in Figure 6.5, the three registration methods had a strong correlation ($r > 0.90$) for the mean dose between the registered floating MR lesions and the reference lesions. The correlation ($r < 0.90$) of V70 between the registered floating MR

lesions and the reference lesions for the three registration methods is weaker than the correlation ($r >$ 0.90) of V100 for the CNN&LM-guided registration and the multi-modality DIR of MIM.

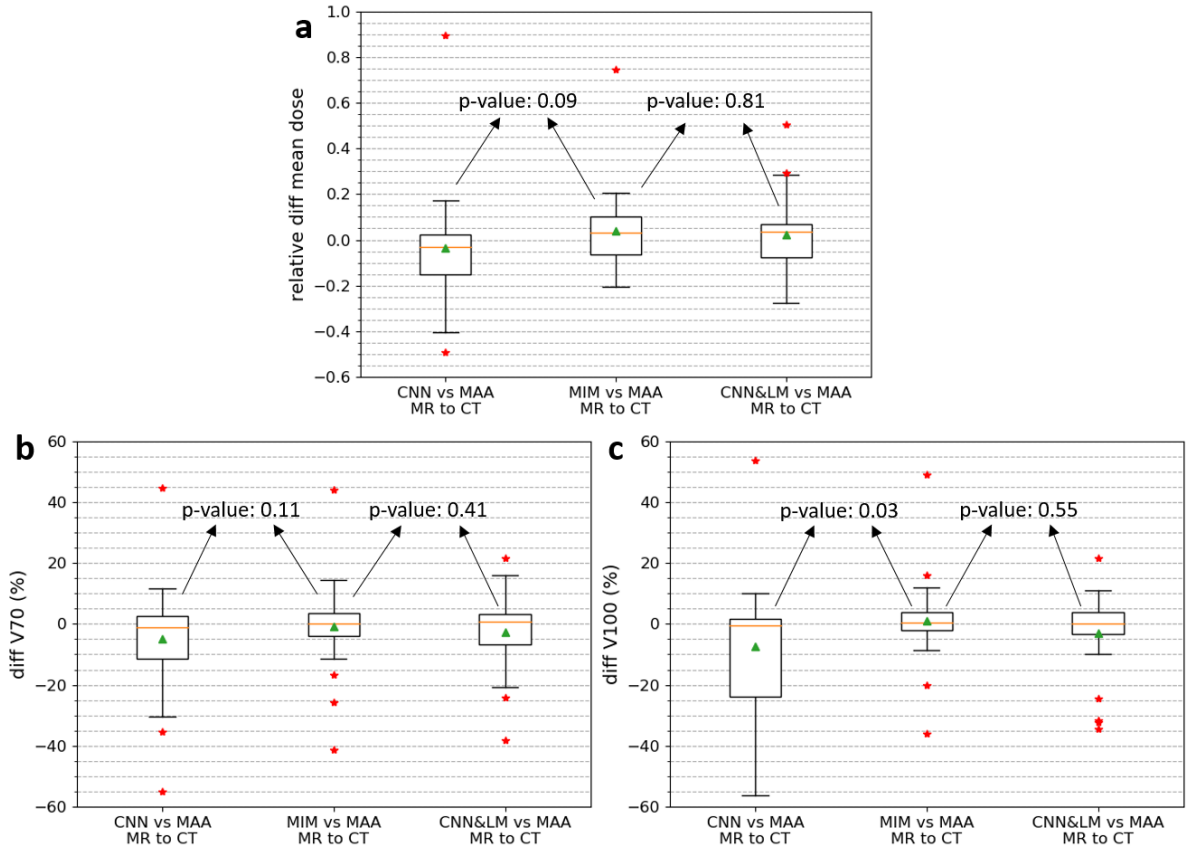



Figure 6.4 Relative difference of mean dose (a) and difference of V70 (b) and V100 (c) between the reference lesions and the registered floating MR lesions using the CNN&LM-guided registration, the CNN-guided registration, and the multi-modality DIR of MIM. MAA: the reference lesions. CNN&LM: the floating lesions registered by the CNN&LM-guided registration. CNN: the floating lesions registered by the CNN-guided registration. MIM: the floating lesions registered by the multi-modality DIR of MIM. The relative difference of mean dose is computed by (mean dose (floating) – mean dose (reference)) / mean dose (reference). The difference of V70 and V100 is computed by (V70 or V100 (floating)) – (V70 or V100 (reference)). The green triangle, the yellow line, and the red stars represent the mean, median, and outliers.

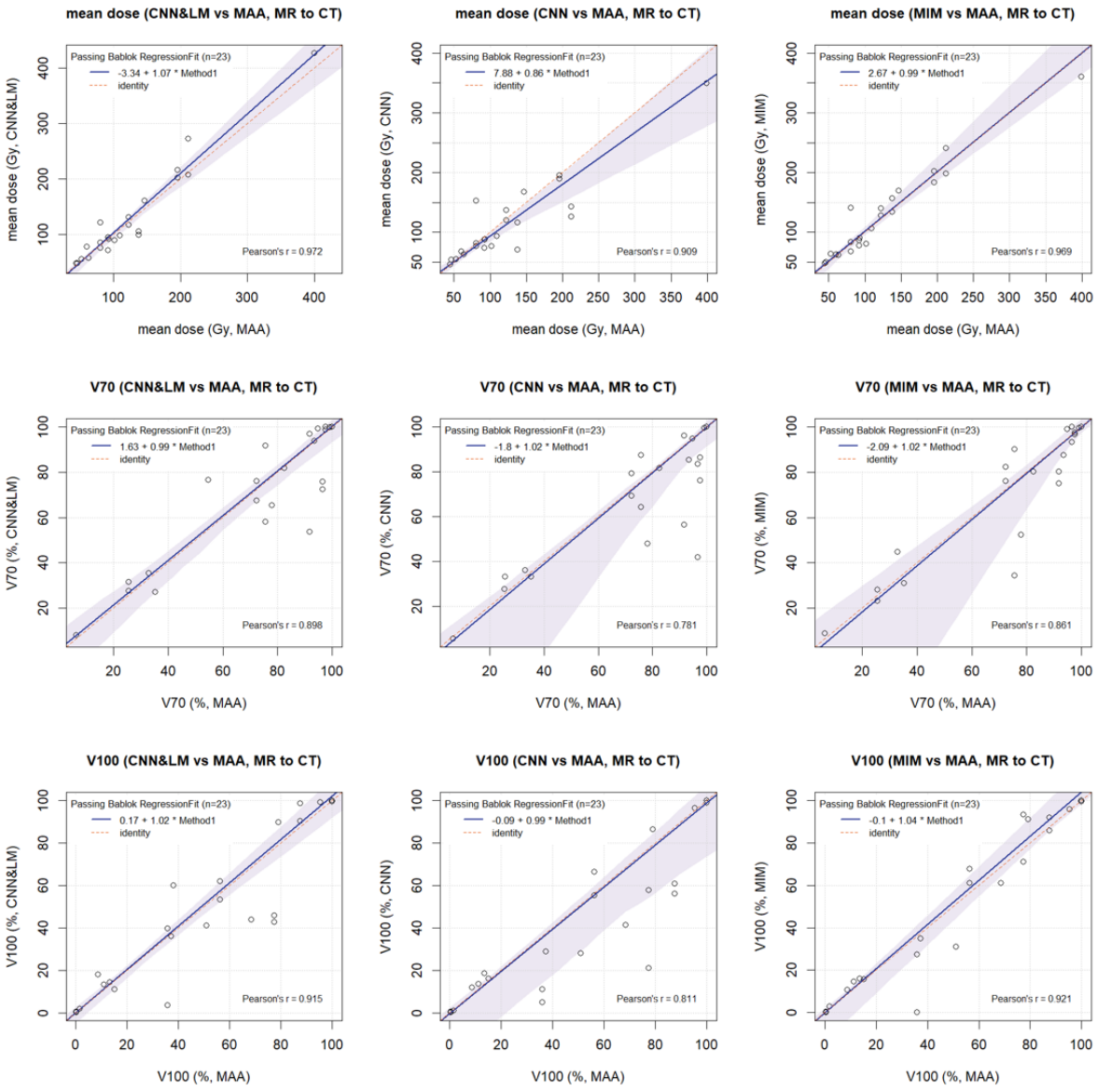



Figure 6.5 Passing-Bablok plots for the mean dose, V70, and V100 estimated on the floating MR lesions registered by the CNN&LM-guided registration, the CNN-guided registration, and the multi-modality DIR of MIM and the mean dose, V70, and V100 estimated on the reference lesions. MAA: the lesions delineated on the $^{99m}$Tc-MAA CT. CNN&LM: the lesions delineated on the floating MR and registered by the CNN&LM-guided registration. CNN: the lesions delineated on the floating MR and registered by the CNN-guided registration. MIM: the lesions delineated on the floating MR and registered by the multi-modality DIR of MIM.

## 6.5 Discussion

The multi-modality DIR tool of MIM achieved comparable performance with the CNN&LM-guided method for CT to CT and MR to CT registration in terms of the registration error and dose estimation difference. The RRMSD difference of the CNN&LM-guided registration and the multi-modality DIR is 0.3 mm for both CT to CT and MR to CT registration, which was not significant. The multi-modality DIR does not require the liver and lesion segmentations on the reference and floating images to guide the registration. This avoids the VOI segmentation on multiple images. The CNN-guided registration, the CNN&LM-guided registration, and the multi-modality DIR achieved a strong correlation ($r > 0.9$) of mean dose between the reference lesions and the registered floating lesions. The multi-modality DIR

achieved a similar number of lesions that have the absolute relative mean dose difference and the absolute V70 and V100 difference smaller than 10%, compared with the CNN&LM-guided registration. The multi-modality DIR workflow starts with a rigid fusion of two images. After that, it is optional to manually adjust the image alignment. The adjusted alignment is used to initialize the deformable registration. In general, the whole DIR process takes around 5 min without any manual alignment.

There is still large room for increasing the performance of the current registration methods. Since guidance of the liver and lesion segmentations helped improve the in-house image-based registration method, incorporating the liver and lesion segmentations into the multi-modality DIR of MIM might be able to further boost the registration performance. Similar studies were conducted to demonstrate the improvement by using guidance of contours or landmarks for deformable image registration. Gu *et al*. developed a contour-guided deformable method for registration of the planning CT and treatment CT for adaptive radiotherapy [131]. The contour-guided deformable registration improved the registration accuracy for one pelvic cancer and five head-and-neck patient data, compared with their original deformable registration [131]. Kearney *et al*. developed a deformable registration method guided by landmarks for the planning CT and CBCT [132]. It was found that the landmark-guided registration achieved better performance than their original deformable registration for six head-and-neck cancer patients [132]. Besides, the DIR assigns a uniform regularization strength to the whole image. It cannot focus on liver registration by assigning more regularization strength to the liver and less strength to the non-liver part to give the liver more flexibility to move as a whole object. Therefore, it can be helpful to define different regularization power inside and outside the liver.

## 6.6 Conclusion

The multi-modality DIR tool of MIM achieved similar registration error as the CNN&LM-guided method for both CT to CT and MR to CT registration. The mean dose estimated on the floating lesion registered by the DIR tool was similar to the mean dose estimated on the reference lesions. A small V70 and V100 difference for most floating CT and MR lesions was observed for both the CNN&LM-guided registration and the multi-modality DIR. This demonstrates the feasibility of using the multi-modality DIR for both the volume-level and voxel-level dosimetry. The multi-modality DIR tool of MIM can be used for SIRT to integrate information from multi-modality images.

# 7. CNN liver lesion segmentation

## 7.1 Abstract

**Purpose:** Liver lesion segmentation is critical for precise dose calculation in SIRT. This study aims at investigating the feasibility of automatic liver lesion segmentation for CT and T1-weighted MR using a convolutional neural network (CNN). **Methods:** Different CNN settings were experimented with for liver lesion segmentation on CT to find the optimal setting. The training (validation) datasets contained 91 (20) CT images from the LiTS17 challenge and 54 (15) CT images from SIRT patients. In the first comparison, the 3D U-net model described in Section 4.3.2 for CT and MR liver segmentation (the low-resolution model, isotropic input voxel size of 3 mm) was used to evaluate the performance difference for training with the unmasked images and training with the images masked by the manual liver segmentation. Since lesion segmentation requires higher contour precision than liver segmentation, the input image resolution was changed to $1 \times 1 \times 3$ mm$^3$, which created a much larger image size and could overload the GPUs. To solve the overloading problem, the masked input images were cropped by the bounding box of the liver segmentation, and a new 3D U-net model (the high-resolution model, input voxel size of $1 \times 1 \times 3$ mm$^3$) was proposed to process 3D patches randomly sampled from the masked and cropped high-resolution images. The low-resolution and high-resolution U-net models (CT-specialized) using the masked input images were compared on the CT datasets for the second setting comparison. After that, 238 T1-weighted MR images from 90 patients, 24 T1-weighted MR images from 10 patients, and 55 T1-weighted MR images from 22 patients were used to train, validate, and test the high-resolution U-net model (MR-specialized) using the masked input images. Multiple T1-weighted MR sequences were selected from the imaging study of each patient. Furthermore, a generalized high-resolution U-net using the masked input was trained, validated, and tested with the abovementioned CT and T1-weighted MR images. Its performance on the validation CT datasets and test T1-weighetd MR images was compared with the CT-specialized and MR-specialized high-resolution U-net models using the masked input, respectively. The CNN lesion segmentations and manual lesion segmentations were compared through the DSC and sensitivity. **Results:** In the first setting comparison, the CT-specialized low-resolution U-net model trained with the masked images achieved mean DSCs 0.08 and 0.03 higher than the CT-specialized low-resolution model trained with the unmasked images for the LITS17 and SIRT validation CT datasets, respectively. In the second setting comparison, the mean DSCs for the CT-specialized high-resolution U-net using the masked images were 0.03 and 0.05 higher than the mean DSCs for the CT-specialized low-resolution U-net using the masked images for the LiTS17 and SIRT challenge CT datasets. The MR-specialized high-resolution U-net model using the masked input achieved a mean DSC of 0.48 and a mean sensitivity of 0.82 for the test MR datasets. The generalized high-resolution U-net using the masked input achieved a mean DSC (sensitivity) of 0.70 (0.68) for the LiTS17 validation CT datasets, a mean DSC (sensitivity) of 0.62 (0.60) for the SIRT validation CT datasets, and a mean DSC (sensitivity) of 0.52 (0.86) for the test MR datasets. **Conclusion:** Masking the input image through the liver segmentation increased the performance of CNN lesion segmentation for CT images. Better liver lesion segmentation was achieved for CT by using the high-resolution U-net compared to the low-resolution U-net. The MR-specialized high-resolution U-net using the masked input achieved relatively poor performance on the test MR datasets, compared to the performance of the CT-specialized U-net using the masked input on the validation CT datasets. Lesions in MR appeared with more intensity and texture variations than lesions in CT, which could result in poorer lesion segmentation for MR than for CT. However, the MR-specialized high-resolution U-net using the masked images achieved good performance on lesion detection for MR, which is clinically usable. The generalized high-resolution U-net using the masked input achieved comparable performance to the CT and MR-specialized high-resolution U-net models

using the masked input, which proves the feasibility of using one U-net model for liver segmentation of both CT and MR images.

## 7.2 Introduction

SIRT is used to treat primary liver cancer and liver metastases non-invasively. Before SIRT is selected as a treatment option for a patient, some imaging studies are performed for tumor evaluation. HCC is usually diagnosed through contrast-enhanced MR sequences, including T1-weighted MR, T2-weighted MR, and diffusion-weighed MR. [$^{18}$F]FDG or [$^{68}$Ga]Ga-DOTA-TATE PET/CT, contrast-enhanced CT or MR scans are used to detect liver metastases. If liver metastases appear as hyper-intense regions in the [$^{18}$F]FDG or [$^{68}$Ga]Ga-DOTA-TATE PET, they can be segmented through thresholding. Besides, liver metastases can also be delineated on contrast-enhanced CT images. For the non-[$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE avid liver tumors, contrast-enhanced MRs are used for tumor segmentation. Recent years have seen an increase in the fraction of SIRT patients that are treated for HCC. HCC is most often non-[$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE avid and needs to be delineated on MR. Therefore, tumor segmentation on MR is of high importance for SIRT.

Before performing SIRT planning, the minimum absorbed dose for killing tumors and the maximum absorbed dose for preserving normal liver tissues are defined to calculate the injected activity. To achieve the best treatment, the lower bound of the injected activity calculated based on the minimum absorbed dose for tumors should be no larger than the upper bound of the injected activity calculated based on the maximum absorbed dose for normal tissues, and the interval between the lower and upper bounds should be as large as possible. The upper bound is usually selected as the injected activity during the treatment, which achieves the largest absorbed dose for killing tumors while preserving normal tissues. Therefore, tumor delineations need to be performed carefully, because they have direct influence on calculation of the lower and upper bounds of the injected activity. Manual tumor delineation is tedious and time-consuming, which hinders introduction in clinical routine. Therefore, automatic tumor segmentation is needed for optimisation of precise and personalized SIRT.

Different methods have been used for (semi-)automatic liver tumor segmentation in the past years, which are described in Section 1.7.2. Thresholding methods make use of the intensity difference between tumors and normal tissues. Different approaches were adopted to find the optimal threshold that best separates tumors and normal tissues [75, 76, 77]. Tumors can appear hyperintense and hypointense in CT and MR. It is difficult to find one or several thresholds to separate all tumors. Separating diffusely infiltrating tumors from normal tissues is very challenging because diffuse tumors might have similar intensities with normal tissues. Region growing methods [80, 81] make use of the intensity similarity within the tumor and the intensity difference with the surroundings. Each region gradually grows from a manually selected seed point by incorporating the surrounding voxels with close intensities to the region until it covers the whole tumor. Region growing methods will fail for heterogeneous and diffuse tumors, where a large intensity variation exists inside the tumor. CNN has been used for liver tumor segmentation in recent years. It can extract high-level features automatically by learning from the human's annotations. The LiTS17 challenge provided 131 training CT datasets and 70 test datasets, which allowed different segmentation methods to be evaluated on the same datasets. The best performing methods adopted two cascaded U-net structures [72]. The second U-net was trained with the image masked through the liver segmentation from the first U-net so that the second U-net can focus on learning features within the liver. The highest DSC for liver tumor segmentation was 0.70 in the LiTS17 challenge [72]. Jansen *et al* used a set of dynamic contrast-enhanced MR and diffusion-weighted MR registered to each other to train a fully convolutional network for liver metastases detection [88]. A sensitivity of 99.8% was reported for liver metastases detection.

In practice, the clinical images for patients with liver cancer contain not only tumors but also other abnormal tissues, such as cysts, focal nodular hyperplasia, and vascular structures including shunts and

iatrogenic changes. All these abnormalities are collectively named lesions. If only tumors are segmented on these images, the segmentation method needs to distinguish tumors from other types of lesions. The classification of lesions requires strong medical expertise, multiple images of different modality, and diagnosis information, which is much more challenging than lesion segmentation. In the current stage, lesion segmentation is sufficient for SIRT use, since the goal is to automate segmentation of tumors, and these are part of the wider category of lesions. Besides, abnormalities such as vascular shunts and large biliary cysts (which do not contribute to functional liver volume) are of interest in SIRT planning.

The aim of this study is to investigate the feasibility of CNN liver lesion segmentation for clinical CT and MR images. A public challenge provided 111 CT datasets with liver lesion delineations, however, no public MR datasets with liver lesion delineations were found. Therefore, the number of clinical CT datasets required for CNN development was much smaller than the required number of clinical MR datasets. To explore the optimal CNN structure and settings for liver lesion segmentation, the clinical CT datasets were firstly collected, delineated, and used for CNN development. After that, the selected structure and setting were applied to CNN lesion segmentation on MR images collected and delineated later. In the clinical routine, HCC diagnosis is based on multiple sequences from a diagnostic MR study, including T1-weighted MR (arterial phase, portal-venous phase, venous phase, etc), T2-weighted MR, and diffusion-weighted MR. It might be favourable to train a CNN with an input of multiple sequences. In practice, the lesions are not always visible in each sequence. The sequence without visible lesions might exert negative influence on the learning process of CNN. Besides, the MR imaging procedure is not standardized, which means a certain sequence is not always present for each patient. This increases the difficulty in selecting fixed types of sequences for CNN training with a single input volume combining multiple MR sequences. The CNN input combining multiple MR sequences also increases the workload of manual lesion delineation. In the current stage, the CNN is trained with a single input of contrast-enhanced T1-weighted MR. The T1-weighted MR shows good anatomical information, which can be used for both liver and lesion segmentation. The CNN trained with T1-weighted MR images can help produce liver masks used for lesion segmentation. In this study, we want to explore the possibility of CNN liver lesion segmentation for clinical contrast-enhanced CT and T1-weighted MR images by using the input of one single image to avoid the difficulty in selection of multiple MR sequences from one study and annotation of a large amount of images.

## 7.3 Materials and methods

### 7.3.1 Data

The CT datasets for CNN liver lesion segmentation contain 111 contrast-enhanced CT images from the LiTS17 challenge and 69 contrast-enhanced CT images from the SIRT patients in our hospital. These CT datasets are mainly from patients with liver metastases and other lesions. Each CT image was from a different imaging study. The LiTS17 challenge provided liver and lesion delineations from the experts. Delineation of the liver and lesions for the SIRT CT datasets was performed by a researcher, and verified or corrected by an experienced radiographer trained in lesion delineation. The CT datasets from the LiTS17 challenge and SIRT patients were divided into the training and validation datasets (See Table 7.1). Due to the lack of sufficient CT images with liver and lesion delineations, no test datasets were separated from the total CT datasets. Since a CNN for MR lesion segmentation was more needed in the SIRT routine than a CNN for CT lesion segmentation, the CT datasets were mainly used to determine the optimal CNN structure and setting, which can be applied to MR lesion segmentation. Therefore, no more test CT datasets were further collected to perform unbiased evaluation of CNN performance on CT datasets unseen by the CNN. The image in-plane resolution of the LiTS17 datasets ranged from 0.56 to 0.98 mm and the slice thickness ranged from 0.7 to 5.0 mm. The SIRT datasets had an image in-plane resolution from 0.65 to 1.37 mm and a slice thickness between 1.0 and 5.0 mm.

Table 7.1 Division of datasets for training, validating, and testing CNN models for liver lesion segmentation

| | Training | Validation | Test | Total |
|---|---|---|---|---|
| LiTS17 (CT) | 91 | 20 | 0 | 111 |
| SIRT (CT) | 54 | 15 | 0 | 69 |
| Hospital (T1 MR) | 238 (90 patients) | 24 (10 patients) | 55 (22 patients) | 317 (122 patients) |

All T1-weighted MR datasets were collected from the patients with HCC and other lesions in our hospital. Each patient had two to four T1-weighted MR sequences from one imaging study. Two researchers delineated the liver and lesions on these MR images and each image was delineated by only one researcher. The liver and lesion delineations were verified or corrected by an experienced radiographer and an experienced nuclear medicine physician, respectively. The MR datasets had an image in-plane resolution between 0.68 and 1.95 mm and a slice thickness from 2.0 to 10.0 mm. The MR datasets were divided into the training, validation, and test datasets (See Table 7.1).

## 7.3.2 CNN development

### 7.3.2.1 U-net model with low-resolution input images

The U-net model described in Section 4.3.2 for liver segmentation was modified by adding residual connections (skip connections) [133] at each resolution level. By using residual connections, the input of the previous convolutional layer is added to the output of the current layer. In our case, the residual connection is used every two convolutional layers (see Figure 7.1). The residual connections allow the information flow between layers without passing convolutional layers, by using direct connection, which has the potential of improving the CNN optimization process. In our experiment, adding residual connections did not improve CNN liver lesion segmentation for CT and also did not degrade segmentation performance. Therefore, the residual connections were retained in the current and following U-net models. The low-resolution U-net model was used to evaluate the influence of masking the input image through a liver mask on lesion segmentation performance, which is described in detail in Section 7.3.4.

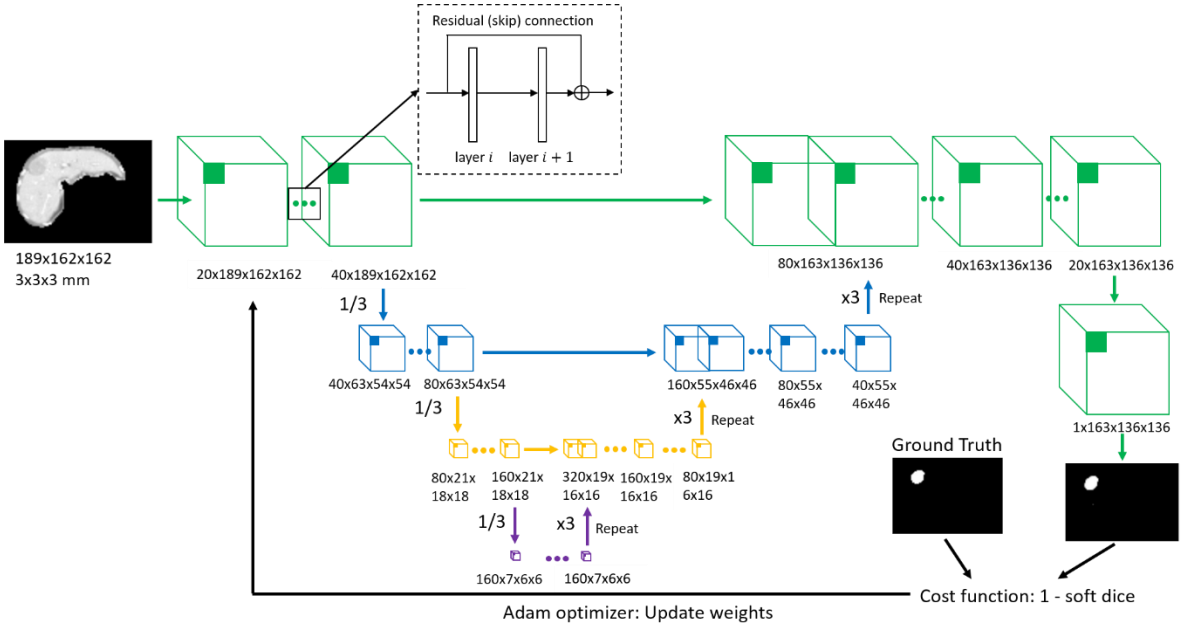



Figure 7.1 Overview of the low-resolution U-net model with residual connections. The model has four resolution levels and the down- and up-sampling factor between two neighbouring levels is 3. The input image resolution is $3 \times 3 \times 3$ mm$^3$ and the input size is $189 \times 162 \times 162$. The output size is $163 \times 136 \times 136$. The "⋯" in the figure represents two convolutional layers with one residual connection between the input of layer $i$ and the output of layer $i + 1$.

**Training**. The low-resolution U-net model was trained with contrast-enhanced CT images. The soft DSC was used as the loss function for the low-resolution U-net model, which was optimized through the Adam optimizer [118]. The initial learning rate was 0.001 and decreased by a factor of 0.5 when the mean DSC of the full segmentations of the validation datasets did not increase. The model quality was evaluated every 10 epochs on the full segmentations of the validation datasets using the DSC. The model parameters that achieved the highest DSC on the validation datasets were saved during the evaluation. The training process of the low-resolution U-net model took around 3 days using 2 GPUs of NVIDIA P100 with 16 GB DRAM.

**Data pre-processing**. The 3D CT images were cropped to cover the whole abdomen in each axial slice and the entire liver in the axial direction. The images were resampled to an isotropic voxel size of 3 mm. The CT image was clipped to HU values between -200 and 200 and normalized by a linear mapping to the range between -0.5 and 0.5. After that, the processed images were masked by the manual liver segmentations so that only the information inside the liver was preserved for CNN training. The unmasked images and the images masked by the liver masks were use as the CNN input, respectively.

**Data post-processing**. The CNN output was treated as a probability map, where each voxel value ranged from 0 to 1. The map was thresholded by 0.5 to obtain the lesion segmentation.

### 7.3.2.2 U-net model with high-resolution input images

A lesion usually has a much smaller volume than the liver. The $^{90}$Y activity can have high concentration in the lesions (especially tumors). A small change of the lesion contour might lead to a big change of dose calculation. Besides, resampling an image to a low resolution might result in information loss especially for a small lesion. This increases the difficulty for the CNN to detect the resampled lesions that almost disappear in the image. Therefore, lesion segmentation probably requires higher resolution than liver segmentation. Since the diagnostic images providing contours are registered to the $^{99m}$Tc-MAA CT with a slice thickness of 3 mm for dose calculation, the input resolution at the axial direction was kept at 3 mm for the U-net. The image in-plane resolution of the U-net input was increased to $1 \times 1$ mm$^3$. The increase of the input resolution results in a much larger image size, which

requires more computation resources and can cause overload of the GPU memory. To solve the overload problem with limited GPU resources, a new U-net structure was proposed.

Instead of inputting the whole 3D image, relatively smaller image patches sampled from the whole image are used as the U-net input to decrease the memory demand. The input patch passes through the whole U-net to generate an output patch, which contains liver segmentation on one part of the whole image. The size of the output patch is not larger than the size of the input patch size. If the output patch is generated on the boundary of the whole image, the input patch will exceed the range of the whole image. The region outside the whole image is padded with zero values (see Figure 7.2), which does not damage the information on the whole image. Each voxel in the output patch covers a sub-region of the input patch through multiple convolutions to classify the voxel as lesion or non-lesion. This sub-region is defined as the receptive field (see Figure 7.2). The U-net cannot learn the context information outside an input patch. If the receptive field of a voxel in the output patch is over the range of the input patch, the region of the receptive field outside the input patch is filled with zero intensities by the network (see Figure 7.2). The zero-padded region should have contained the voxel intensities that are in the original image but not included in the input patch. This leads to degradation of the original image information and forces the U-net to learn wrong information that never occurs in the image, which can cause false segmentation. Therefore, the receptive field is decreased to be within the input patch by using a down- and up-sampling factor of 2 between two resolution levels (the sampling factor of the low-resolution U-net is 3). Besides, the feature volume is not padded with zero values on the boundary before undergoing one convolution, which results in an output size smaller than the input size after one convolution. The design details of the high-resolution U-net model are presented in Figure 7.3. The high-resolution U-net model was trained with contrast-enhanced CT images and/or contrast-enhanced T1-weighted MR images to evaluate its performance on CT and MR images, which will be described in detail in Section 7.3.4.

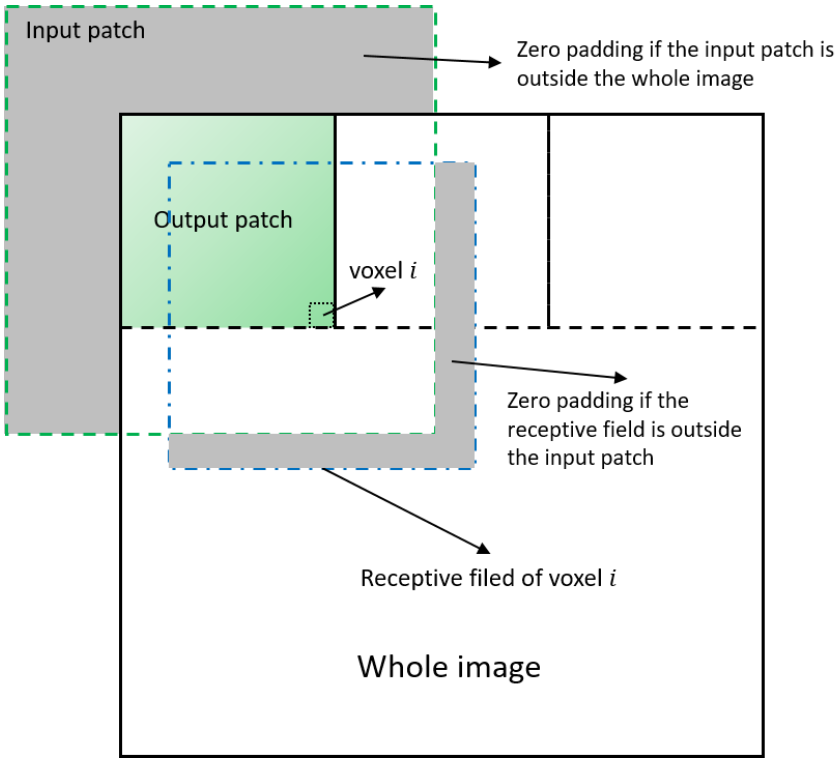



Figure 7.2 Demonstration of the sampling of input and output patches, zero padding, and the receptive filed for the high-resolution U-net.

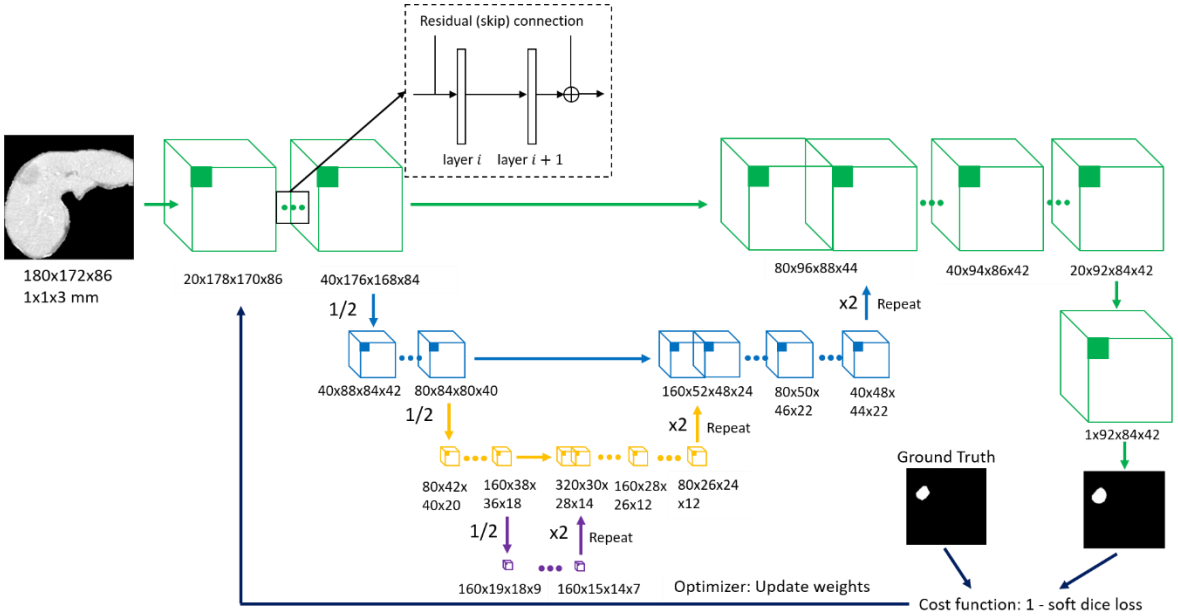



Figure 7.3 Overview of the high-resolution U-net model. The model has four resolution levels and a down- and up-sampling factor of 2 between two levels. The input image resolution is $1 \times 1 \times 3$ mm$^3$. The input patch size is $180 \times 172 \times 86$ and the output patch size is $92 \times 84 \times 42$. The residual connections are used at each resolution level.

**Training**. The loss function for the high-resolution U-net model was the soft DSC and was optimized through the Adam optimizer [118]. The initial learning rate was 0.001 and decreased by a factor of 0.5 when the mean DSC of the full segmentations of the validation datasets did not increase. The model quality was evaluated every 5 epochs on the full segmentations of the validation datasets using the DSC. The model parameters that achieved the highest DSC on the validation datasets were saved during evaluation. The training process of the high-resolution U-net model took around 3 days using 4 GPUs of NVIDIA P100 with 16 GB DRAM.

**Data pre-processing**. The 3D CT and MR images were masked with the manual liver segmentations and cropped by using the bounding box of the liver segmentations. In this way, the U-net model can focus on learning features inside the liver for lesion segmentation and the input image size can be decreased to reduce the demand for computation resources. The images were resampled to an anisotropic voxel size of $1 \times 1 \times 3$ mm$^3$. The CT image was clipped to HU values between -200 and 200 and normalized by a linear mapping to the range between -0.5 and 0.5. The MR intensities were subtracted by the median intensity of the voxels within the liver and normalized by a linear mapping to a range within [-0.5, 0.5].

**Data post-processing**. The CNN output was treated as a probability map, where each voxel value ranged from 0 to 1. The map was thresholded by 0.5 to obtain the lesion segmentation.

### 7.3.3 Evaluation metrics

In our experiments, the difference between the CNN lesion segmentation and the manual lesion segmentation was computed through the DSC. The DSC measures the overlap between two segmentations [112]. Its value is between 0 and 1. The larger the DSC is, the more overlap the two segmentations have.

In practice, false detection (false positive) or segmentation of liver lesions can be easily removed without consuming much time. Missing detections (false negative) and segmentations of liver lesions will cost doctors time to find and manually segment the undetected lesions. Therefore, it is important to measure the sensitivity, which is also called true positive rate (TPR), to evaluate if the network can

detect as many lesions as possible. The computation of the sensitivity is presented in Equation 7.1, where TP and FN represent true positive and false negative.

$$sensitivity = \frac{TP}{TP + FN} \tag{7.1}$$

Besides, the precision is measured to evaluate the ratio of the true positive detections to the total positive detections, as is shown in Equation 7.2. FP represents false positive.

$$precision = \frac{TP}{TP + FP} \tag{7.2}$$

The sum of TP and FN is the number of manually delineated lesions in the image. If one or multiple CNN lesion segmentations have an overlap with a manually delineated lesion, the TP detection number is counted as one. If one CNN lesion segmentation has an overlap with multiple manually delineated lesions, the TP detection number is equal to the number of manually delineated lesions. If one CNN lesion segmentation has no overlap with any delineated lesions, the FP detection number is counted as one.

When comparing the metrics for different CNN settings, the statistical tests mentioned in Section 4.3.4 were used. The level of significance was set to 0.05.

### 7.3.4 Experiments

**Comparison between the low-resolution U-net models using the unmasked and masked input of CT images**. The irrelevant information outside the liver might increase the difficulty for the network to learn features for liver lesion segmentation. Masking the whole image with the liver segmentation can help the network focus on learning features within the liver. In practice, the manual liver segmentation is slightly dilated by 6 mm to avoid excluding the voxels on the liver boundary. For convenience, the masked input always means that the input image is masked by the dilated manual liver segmentation in the following paragraphs of this chapter. To evaluate the influence of masking the image on lesion segmentation performance, the low-resolution U-net was trained with the unmasked input images and the masked input images, respectively. The U-net models were trained with 91 CT datasets from the LiTS17 challenge and 54 CT datasets from SIRT patients, and validated with 20 CT datasets from the LiTS17 challenge and 15 CT datasets from SIRT patients (see Table 7.1). The manual lesion segmentations of these datasets were used for CNN training and validation. The CNN lesion segmentations were compared with the manual lesion segmentations by computing the DSC. The CNN's performance on lesion detection was evaluated by computing the sensitivity.

**Evaluation of the random error of CNN**. The CNN contains a huge amount of parameters which are used to learn features for segmentation. These parameters are assigned with random values from a probabilistic distribution at the initial stage, which are updated during the training process. Different random parameters can be assigned for each training, which might result in a different segmentation performance (random error). When comparing the segmentation performance for different CNN settings, the performance difference might be caused by the random error instead of the CNN settings. Therefore, it is necessary to perform an experiment to evaluate the error caused by the random initialization of CNN parameters. The low-resolution U-net using the masked input images was trained and validated three times with the same CT datasets and their manual lesion segmentations used in the first experiment. The DSC for each validation CT dataset was computed between the CNN lesion segmentation and the manual lesion segmentation for each trained low-resolution U-net model.

**Comparison between the low-resolution and high-resolution U-net models using the masked input of CT images**. As is explained in Section 7.3.2.2, a lesion segmentation needs to have higher

resolution than a liver segmentation does for precise dose calculation within the lesion. Besides, resampling the original image to a low resolution can cause information loss, especially for small lesions. This might result in failed detection of small lesions. Therefore, a high-resolution U-net model was developed (described in Section 7.3.2.2). To compare the performance of the high-resolution and low-resolution U-net models, the two models were trained and validated with the same CT datasets and their manual lesion segmentations from the LiTS17 challenge and SIRT patients (see Table 7.1). The models trained with CT images are called CT-specialized models in the following paragraphs of this chapter. The segmentation performance was evaluated by computing the DSC and sensitivity between the CNN lesion segmentation and the manual lesion segmentation.

**High-resolution U-net model using the masked input for liver lesion segmentation on MR images**. After the above three experiments, the high-resolution U-net model using the masked input was selected for liver lesion segmentation. If masking the CT image with the liver segmentation can help the CNN focus on learning features inside the liver for better performance on lesion segmentation, this advantage is expected to remain unchanged when masking the MR image for CNN lesion segmentation. Similarly, if higher resolution is beneficial for lesion segmentation of CT, the same is expected to hold for MR. Therefore, the high-resolution U-net model using the masked input was also trained for liver lesion segmentation on MR images. The high-resolution model using the masked input was trained, validated, and tested with the T1-weighted MR images and their manual liver segmentations described in Table 7.1. The models trained with T1-weighted MR images are called MR-specialized models in the following paragraphs of this chapter. The performance on MR lesion segmentation was evaluated by computing the DSC and sensitivity between the CNN lesion segmentation and the manual lesion segmentation for the test datasets.

**Generalized high-resolution U-net model using the masked input for liver lesion segmentation on CT and MR images**. One generalized U-net that can produce lesion segmentations for both CT and MR images is simpler to be deployed in the clinical workflow. In this experiment, the high-resolution U-net model using the masked input was trained with CT and T1-weighted MR images described in Table 7.1. Its performance on the validation CT datasets (see Table 7.1) were compared with the performance of the CT-specialized high-resolution U-net, and its performance on the test MR datasets (see Table 7.1) were compared with the performance of the MR-specialized high-resolution U-net. The performance of different models on lesion segmentation were compared through the DSC and sensitivity between the CNN lesion segmentation and the manual lesion segmentation. The feasibility of replacing the specialized models with the generalized model for similar performance on CT and MR images was evaluated through this experiment.

### 7.3.5 Results

**Comparison between the low-resolution U-net models using the unmasked and masked input of CT images**. As is shown in Figure 7.4, the mean DSCs for the low-resolution U-net models using the masked input and the unmasked input were 0.66 and 0.58 for the LiTS17 validation CT datasets, respectively. For the SIRT validation CT datasets, the mean DSCs for the low-resolution U-net models using the masked input and the unmasked input were 0.57 and 0.54, respectively. For the LiTS17 CT datasets, the low-resolution U-net models using the masked input and the unmasked input achieved the same mean sensitivity of 0.63. For the SIRT CT datasets, the mean sensitivities for the low-resolution U-net models using the masked input and the unmasked input achieved were 0.48 and 0.42, respectively. As is shown in Figure 7.6, the low-resolution U-net models using the masked input and the unmasked input achieved mean precisions of 0.79 and 0.74 for the LiTS17 validation CT datasets. The mean precisions for the low-resolution models using the masked input and the unmasked input were 0.65 and 0.66 for the SIRT validation CT datasets, respectively.

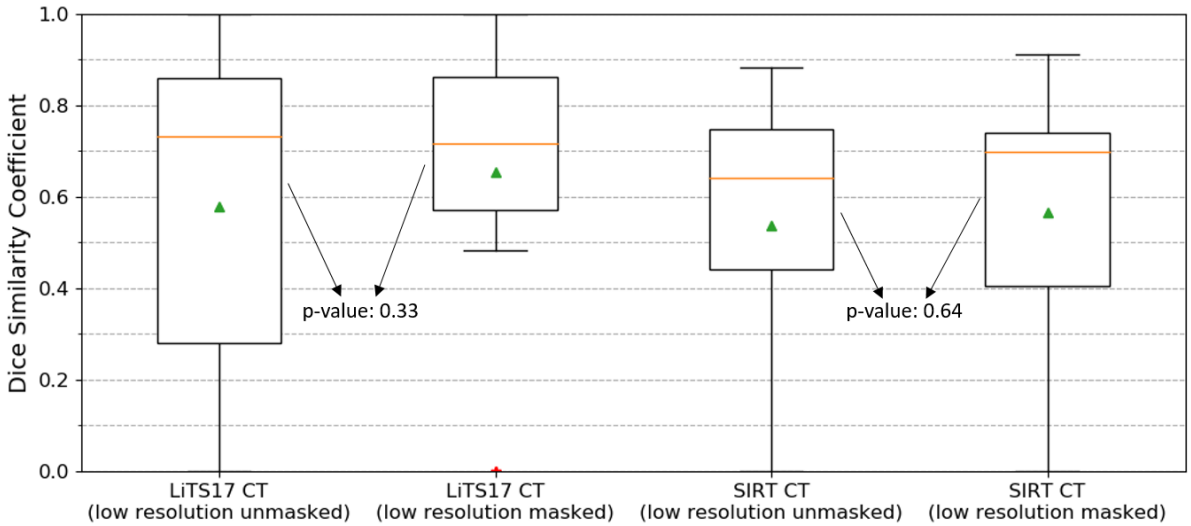



Figure 7.4 DSCs for the validation CT datasets from the LiTS17 challenge and SIRT patients by using the low-resolution U-net with the unmasked input (low resolution unmasked) and the low-resolution U-net with the masked input (low resolution masked). The orange line, the green triangle, and the red star represent the median, mean, and outlier.

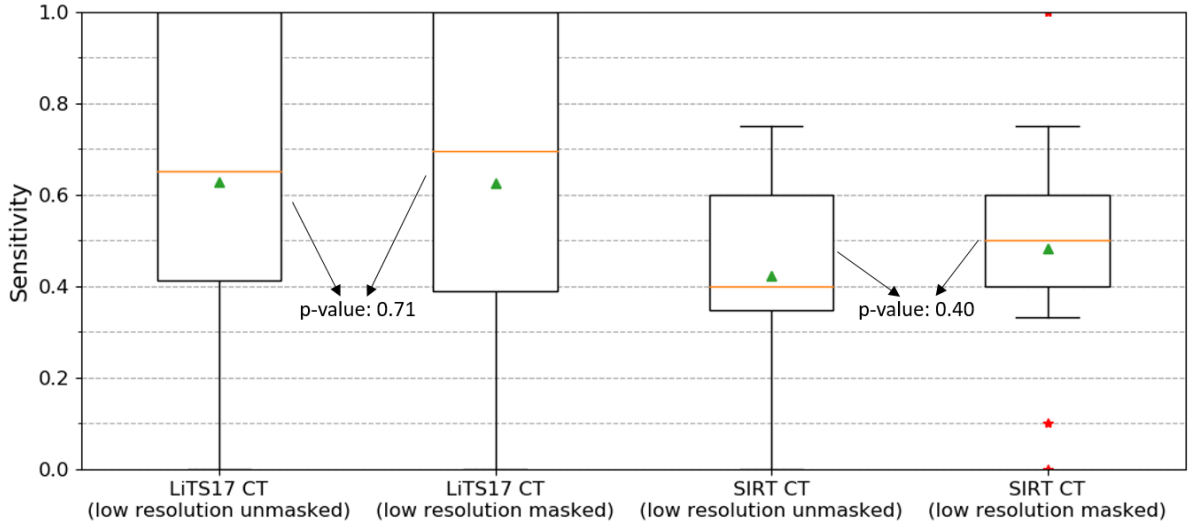



Figure 7.5 Sensitivities for the validation CT datasets from the LiTS17 challenge and SIRT patients by using the low-resolution U-net with the unmasked input (low resolution unmasked) and the low-resolution U-net with the masked input (low resolution masked). The orange line, the green triangle, and the red star represent the median, mean, and outlier.

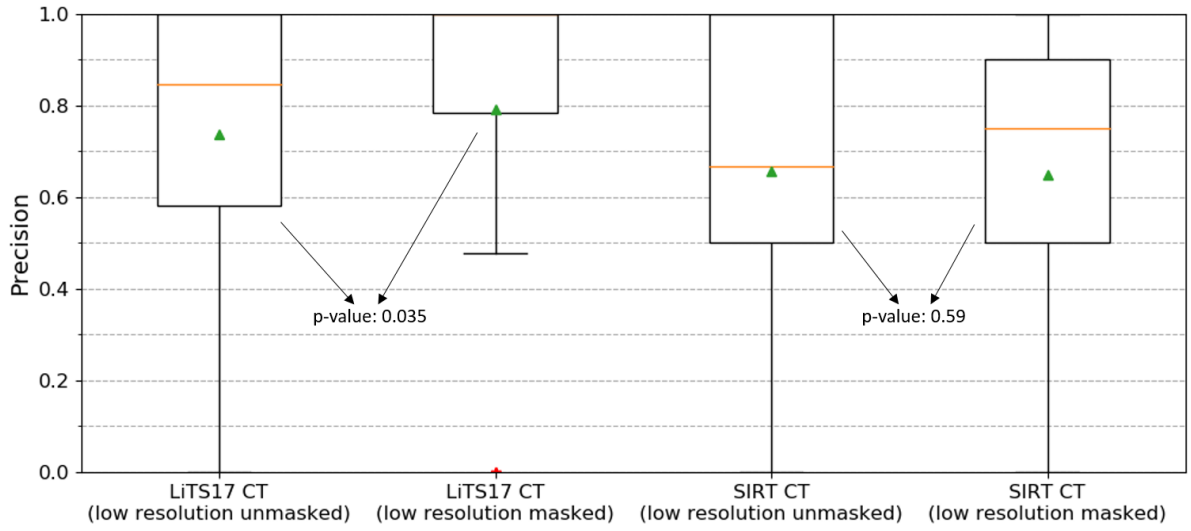



Figure 7.6 Precisions for the validation CT datasets from the LiTS17 challenge and SIRT patients by using the low-resolution U-net with the unmasked input (low resolution unmasked) and the low-resolution U-net with the masked input (low resolution masked). The orange line, the green triangle, and the red star represent the median, mean, and outlier.

**Evaluation for the random error of CNN**. The DSCs, sensitivities, and precisions for the total validation CT datasets (see Table 7.1) using the low-resolution U-net trained with the masked training CT datasets for three times were presented in Figure 7.7, Figure 7.8, and Figure 7.9. The difference between the three groups of DSCs (sensitivities/precisions) caused by the random initialization of CNN parameters was not statistically significant.

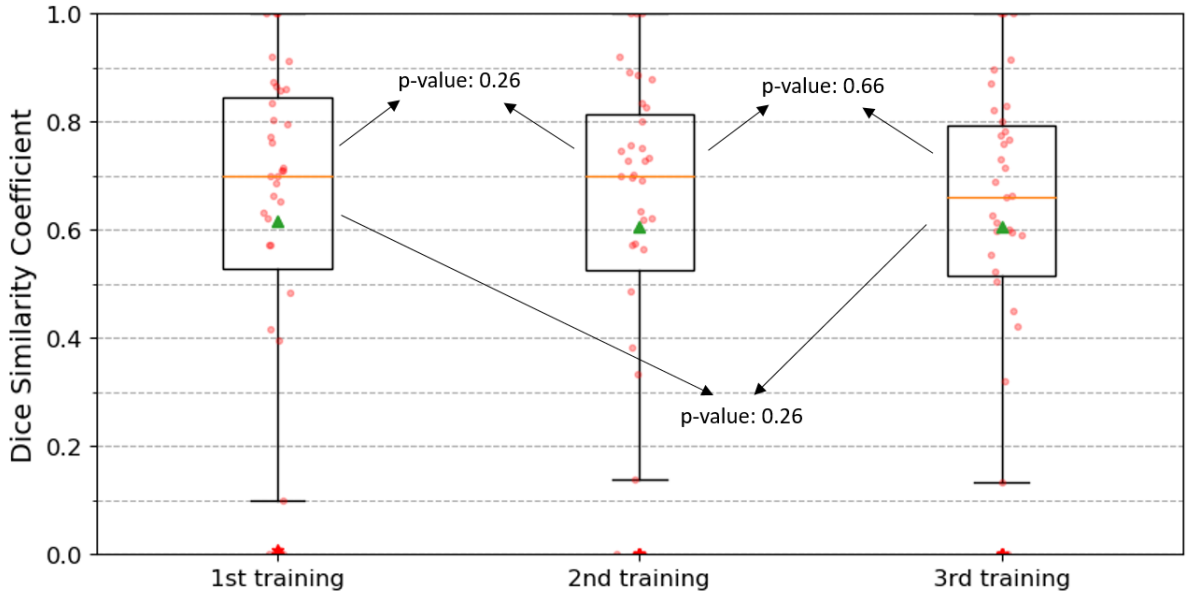



Figure 7.7 DSCs between CNN lesion segmentations and manual lesion segmentations for the total validation CT datasets, using the low-resolution U-net trained with the masked training CT datasets for three times. The Kruskal-Wallis H test [134] was used to evaluate the median difference of the DSCs from three times of training. Its p-value was 0.89. The orange line, the green triangle, and the red star represent the median, mean, and outlier.

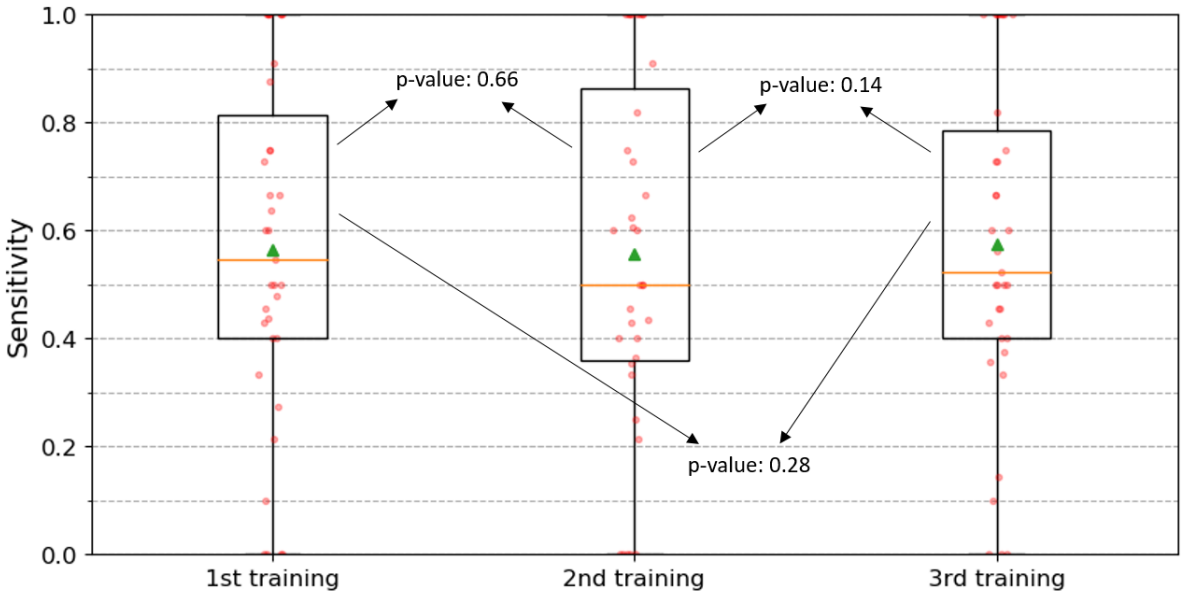



Figure 7.8 Sensitivities between CNN lesion segmentations and manual lesion segmentations for the total validation CT datasets, using the low-resolution U-net trained with the masked training CT datasets for three times. The Kruskal-Wallis H test was used to evaluate the median difference of the sensitivities from three times of training. Its p-value was 0.96. The orange line, the green triangle, and the red star represent the median, mean, and outlier.

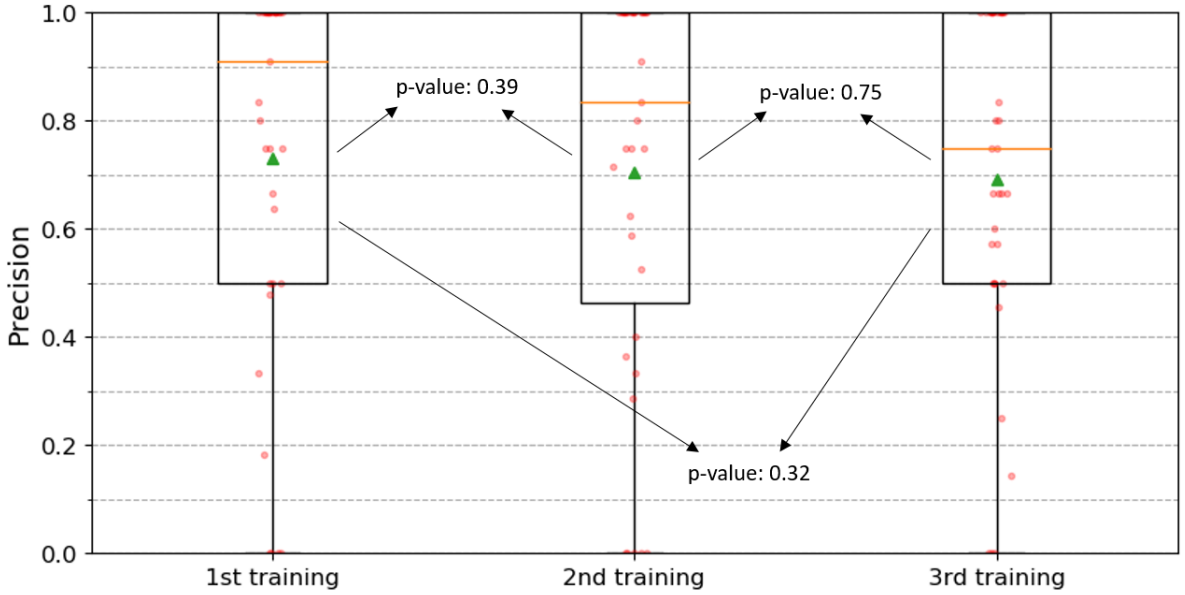



Figure 7.9 Precisions between CNN lesion segmentations and manual lesion segmentations for the total validation CT datasets, using the low-resolution U-net trained with the masked training CT datasets for three times. The Kruskal-Wallis H test was used to evaluate the median difference of the precisions from three times of training. Its p-value was 0.32. The orange line, the green triangle, and the red star represent the median, mean, and outlier.

**Comparison between the low-resolution and high-resolution U-net models using the masked input of CT images**. As is shown in Figure 7.10, the low-resolution and high-resolution U-net models with the masked input achieved mean DSCs of 0.66 and 0.69 for the LiTS17 validation CT datasets ($p = 0.04$), respectively. For the validation CT datasets from the SIRT patients, the mean DSCs for the low-resolution and high-resolution U-net models with the masked input were 0.57 and 0.62 ($p = 0.57$), respectively. For the LiTS17 validation CT datasets, the low-resolution and high-resolution U-net models with the masked input achieved mean sensitivities of 0.63 and 0.70 for the LiTS17 validation

CT datasets ($p = 0.05$), respectively (see Figure 7.11). For the SIRT validation CT datasets, the mean sensitivities for the low-resolution and high-resolution U-net models with the masked input were 0.48 and 0.64 ($p = 0.003$), respectively (see Figure 7.11). As is shown in Figure 7.12, the mean precisions for the low-resolution and high-resolution U-net models using the masked input were 0.79 and 0.52 for the LiTS17 validation CT datasets. For the SIRT validation CT datasets, the low-resolution and high-resolution models using the masked input achieved mean precisions of 0.65 and 0.53, respectively.

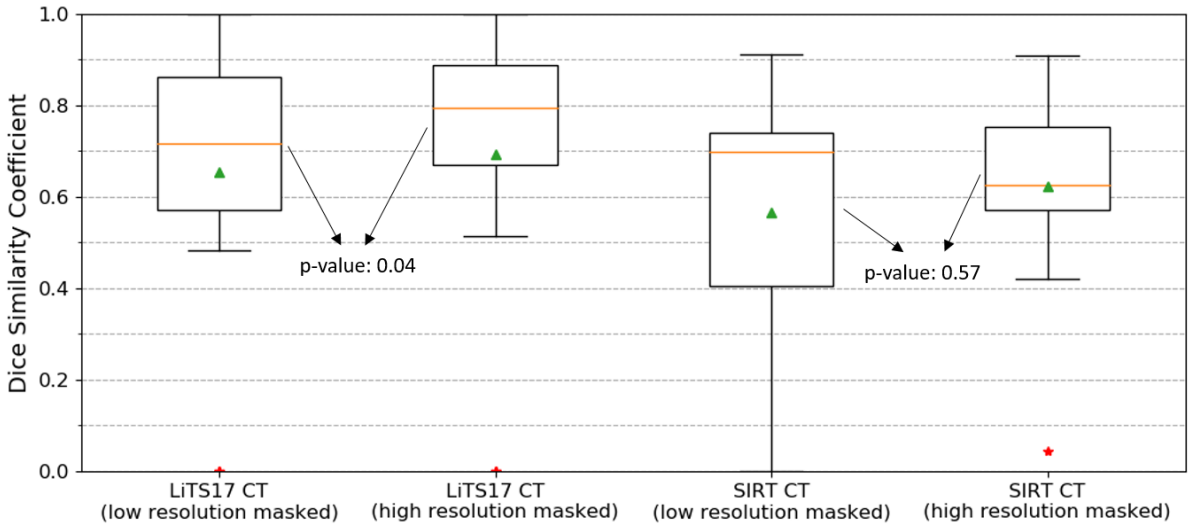



Figure 7.10 DSCs for the validation CT datasets from the LiTS17 challenge and SIRT patients by using the low-resolution U-net with the masked input (low resolution masked) and the high-resolution U-net with the masked input (high resolution masked). The orange line, the green triangle, and the red star represent the median, mean, and outlier.

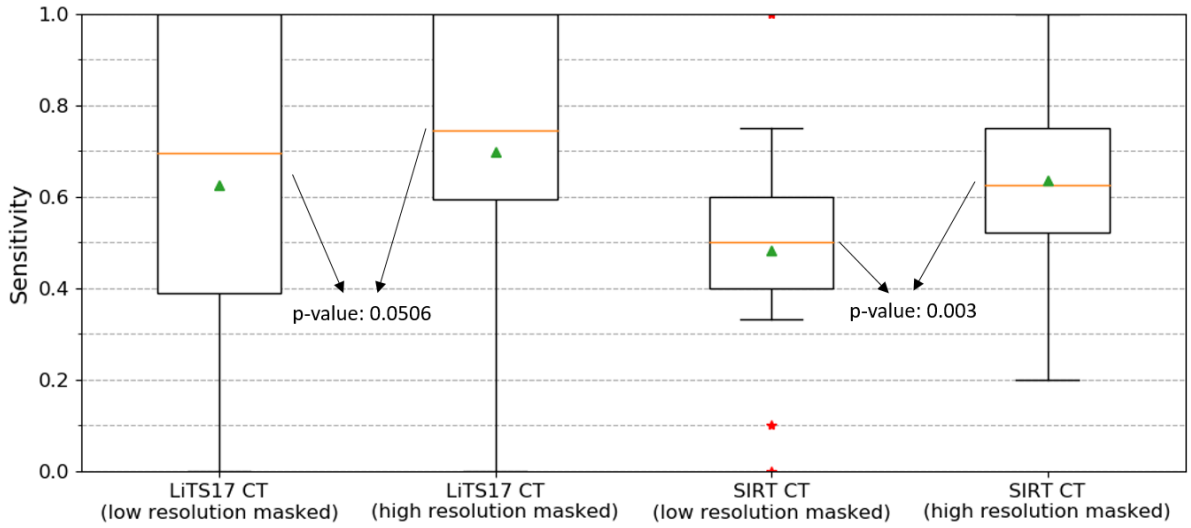



Figure 7.11 Sensitivities for the validation CT datasets from the LiTS17 challenge and SIRT patients by using the low-resolution U-net with the masked input (low resolution masked) and the high-resolution U-net with the masked input (high resolution masked). The orange line, the green triangle, and the red star represent the median, mean, and outlier.

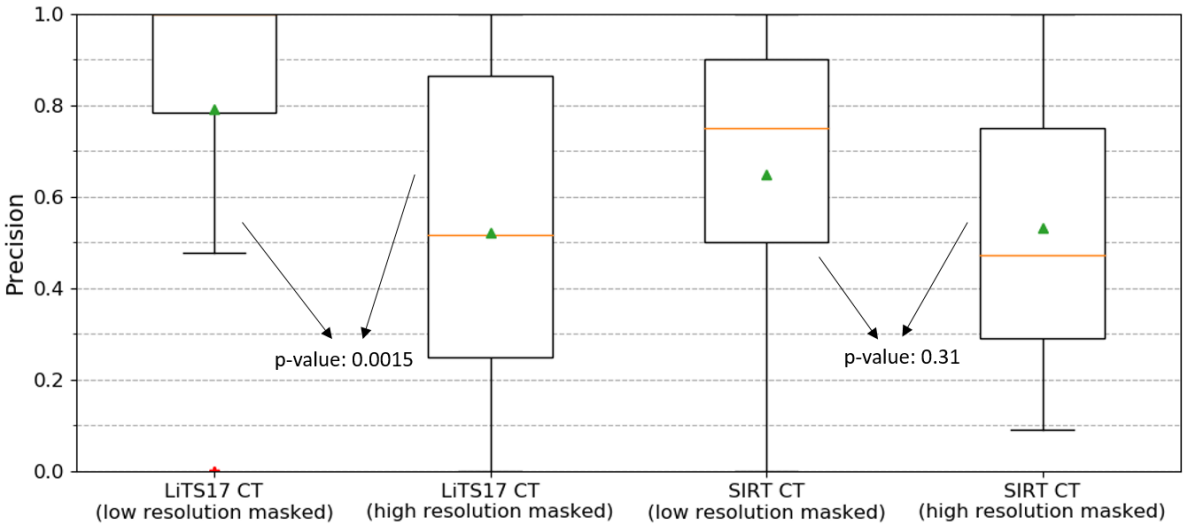



Figure 7.12 Precisions for the validation CT datasets from the LiTS17 challenge and SIRT patients by using the low-resolution U-net with the masked input (low resolution masked) and the high-resolution U-net with the masked input (high resolution masked). The orange line, the green triangle, and the red star represent the median, mean, and outlier.

**High-resolution U-net model using the masked input for liver lesion segmentation on MR images**. As is shown in Figure 7.13a, the high-resolution U-net model trained with the masked T1-weighted MR images achieved a mean DSC of 0.49 and a median DSC of 0.60 for the test MR datasets from our hospital. The median and mean sensitivity for the high-resolution U-net model was 1.0 and 0.82 for the test MR datasets (see Figure 7.13b). The high-resolution U-net model using the masked T1-weighted MR images achieved a median precision 0.33 of and a mean precision of 0.40 (see Figure 7.13c).

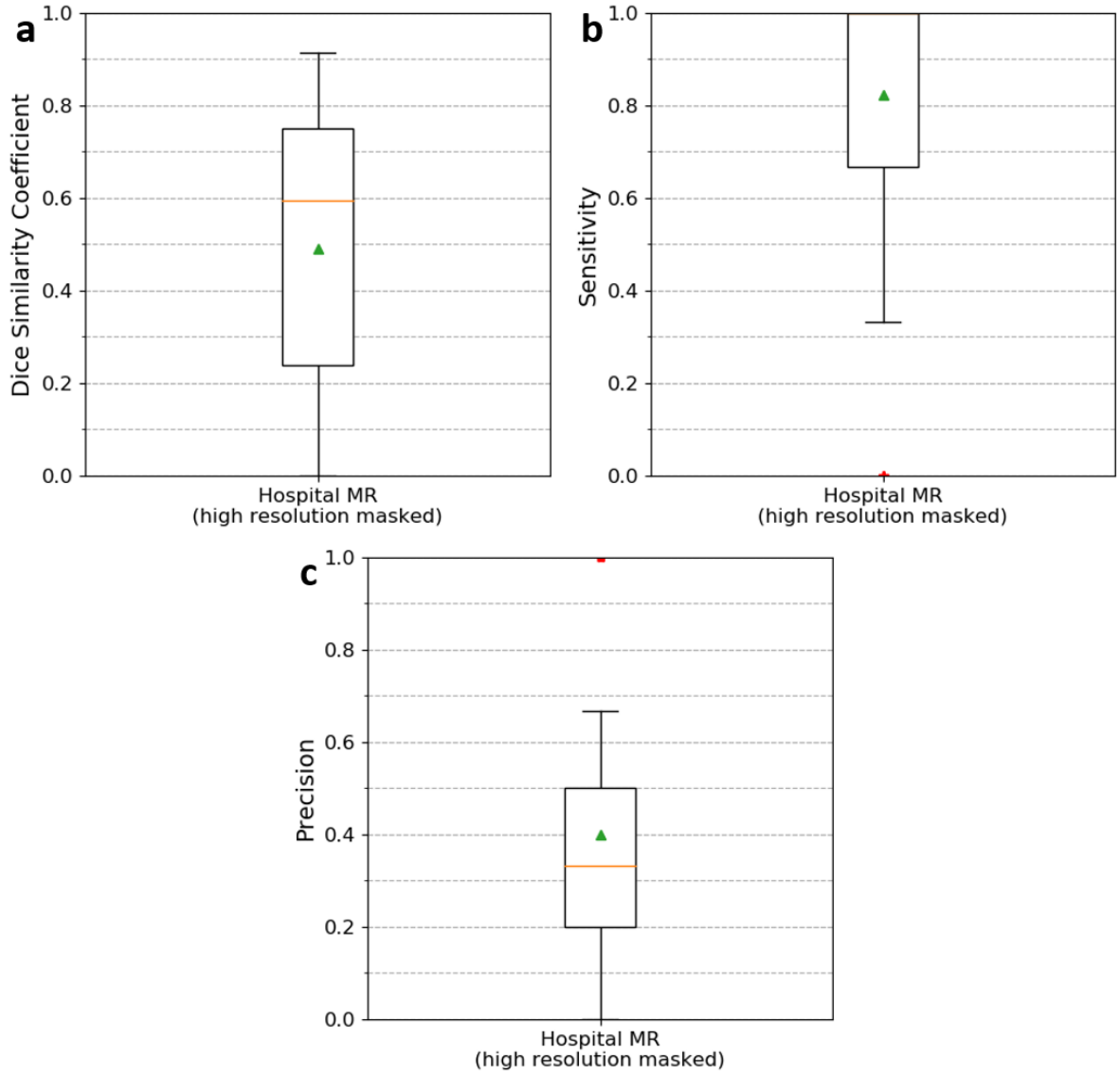



Figure 7.13 DSCs (a), sensitivities (b), and precisions (c) for the test MR datasets from our hospital by using the high-resolution U-net with the masked input (high resolution masked). The orange line, the green triangle, and the red star represent the median, mean, and outlier. The median in Figure 7.13b is 1.0, which is not visible.

**Generalized high-resolution U-net model using the masked input for liver lesion segmentation on CT and MR images**. The generalized high-resolution U-net achieved a mean DSC of 0.70 for the LiTS17 validation CT datasets and a mean DSC of 0.62 for the SIRT validation CT datasets. The CT-specialized U-net achieved a mean DSC of 0.69 for the LiTS17 validation CT datasets and a mean DSC of 0.62 for the SIRT validation CT datasets. For the test MR datasets from our hospital, the generalized and MR-specialized high-resolution U-net models achieved mean DSCs of 0.52 and 0.49, respectively. For the LiTS17 validation CT datasets, the mean sensitivities for the generalized and CT-specialized high-resolution U-net models were 0.68 and 0.70, respectively. For the SIRT validation CT datasets, the mean sensitivities for the generalized and CT-specialized high-resolution U-net models were 0.60 and 0.64, respectively. The generalized and MR-specialized high-resolution U-net models achieved mean DSCs of 0.86 and 0.82 for the test MR datasets, respectively. For the LiTS17 validation CT datasets, the generalized and CT-specialized high-resolution U-net models achieved mean precisions of 0.51 and 0.52, respectively. For the SIRT validation CT datasets, the mean precisions for the generalized and CT-specialized high-resolution models were 0.52 and 0.53, respectively. For the test MR datasets,

the generalized and MR-specialized high-resolution models achieved mean precisions of 0.44 and 0.40, respectively.

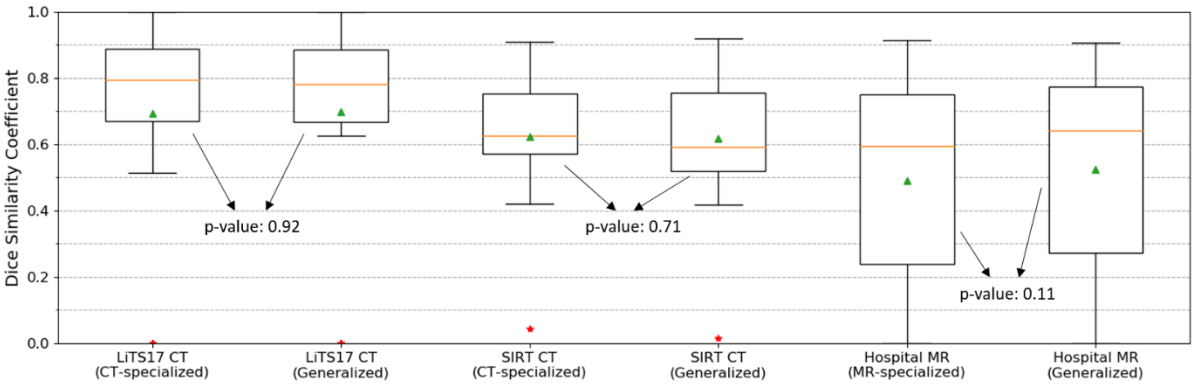



Figure 7.14 DSCs for the LiTS17 and SIRT validation CT datasets by using the CT-specialized and generalized high-resolution U-net models with the masked input and DSCs for the test MR datasets from our hospital by using the MR-specialized and generalized high-resolution U-net models with the masked input. The orange line, the green triangle, and the red star represent the median, mean, and outlier.

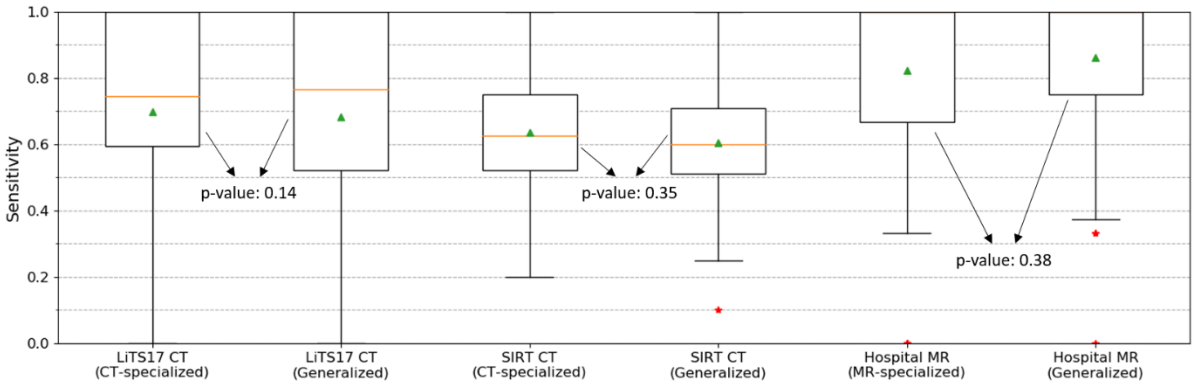



Figure 7.15 Sensitivities for the LiTS17 and SIRT validation CT datasets by using the CT-specialized and generalized high-resolution U-net models with the masked input and sensitivities for the test MR datasets from our hospital by using the MR-specialized and generalized high-resolution U-net models with the masked input. The orange line, the green triangle, and the red star represent the median, mean, and outlier.

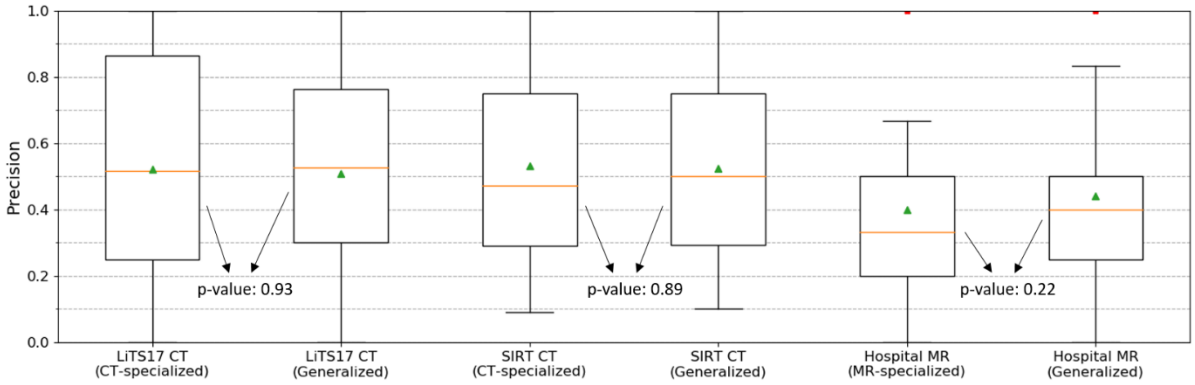



Figure 7.16 Precisions for the LiTS17 and SIRT validation CT datasets by using the CT-specialized and generalized high-resolution U-net models with the masked input and precisions for the test MR datasets from our hospital by using the MR-specialized and generalized high-resolution U-net models with the masked input. The orange line, the green triangle, and the red star represent the median, mean, and outlier.

## 7.4 Discussion

As is shown in Figure 7.4, the low-resolution U-net with the masked input improved the segmentation performance for the LiTS17 and SIRT validation CT datasets compared to the low-resolution U-net with the unmasked input, although this improvement was not statistically significant. The liver mask was also used for CNN liver lesion segmentation by Chlebus *et al* [87] and Jansen *et al* [88]. Chlebus *et al* computed the loss function of the CNN only in a liver mask dilated by 10 mm [87]. The output of the CNN was restricted within a liver mask to remove false positives outside the liver [87]. Jansen *et al* masked the CNN output with a liver mask [88]. Masking the input image through the liver segmentation forces the CNN to focus on learning features inside the liver, which prevents the disturbance from the organs and tissues outside the liver. This strategy simplifies the lesion segmentation task for the CNN by imposing prior information about the liver and can help improve the performance on liver lesion segmentation. Although the liver segmentation for masking the input was manually delineated, the CNN liver segmentation can also be used provided that it does not miss liver lesions. In case that some lesions are excluded in the CNN liver segmentation, manual correction can be applied to include these lesions in a short time. The liver mask does not need to precisely follow the liver boundary. Therefore, the CNN liver segmentation can be used for masking the input, as long as it does not miss a large part of the whole liver.

For the LiTS17 and SIRT validation CT datasets, the mean DSCs and sensitivities for the high-resolution U-net using the masked input were improved compared to the mean DSCs and sensitivities for the low-resolution U-net using the masked input, respectively (see Figure 7.10 and Figure 7.11). For the low-resolution U-net, the original image is resampled to a relatively low resolution for being used as the CNN input. The small lesions are barely visible in these resampled images. The details of small lesions are better preserved in the input image of the high-resolution U-net, which has a smaller voxel size. This enables the CNN to learn features of these small lesions for better detection and segmentation.

The high-resolution U-net using the masked input achieved a median DSC of 0.60 and a mean DSC of 0.49 for the test MR datasets from our hospital (see Figure 7.13a). The performance of the CNN on MR lesion segmentation was poorer than its performance on CT lesion segmentation. MR images measure the water content of tissues while CT images measure the x-ray attenuation of tissues. MR images have more intensity, contrast, and texture variations than CT images. Lesions can appear with more heterogeneity in MR images than in CT images. In SIRT, MR images are mainly used for HCC detection, while CT images are used for liver metastases detection. HCCs can be highly diffuse in MR images for some patients, while liver metastases generally appear with homogeneous intensity in CT images. The above factors increase the difficulty in MR lesion segmentation for the CNN.

As is shown in Figure 7.13b, the median sensitivity for MR lesion detection using the high-resolution U-net was 1.0, which was much higher than that for CT lesion detection using the same U-net structure. The number of liver metastases in a patient can be much larger than the number of HCCs in a patient. As is shown in Figure 7.17, liver metastases can distribute in many spots of a liver and each metastasis is usually not large and might have small intensity difference with the healthy tissue. It is difficult for the CNN to detect very small lesions, which often occur in CT images for patients with liver metastases. The number of HCCs in a liver can be more limited as this is a primary liver tumor and the volume of a HCC can be relatively large, which makes it easy to be detected by the CNN. This resulted in a higher sensitivity for MR lesion detection than for CT lesion detection. Since the CNN segmented mainly liver metastases with relatively large volume in CT, the DSC for the liver segmentation in the whole liver could still be high. In SIRT, a very small tumor is usually not used for calculation of the activity to inject. Therefore, missing these small tumors in CNN lesion segmentation may be acceptable for SIRT.

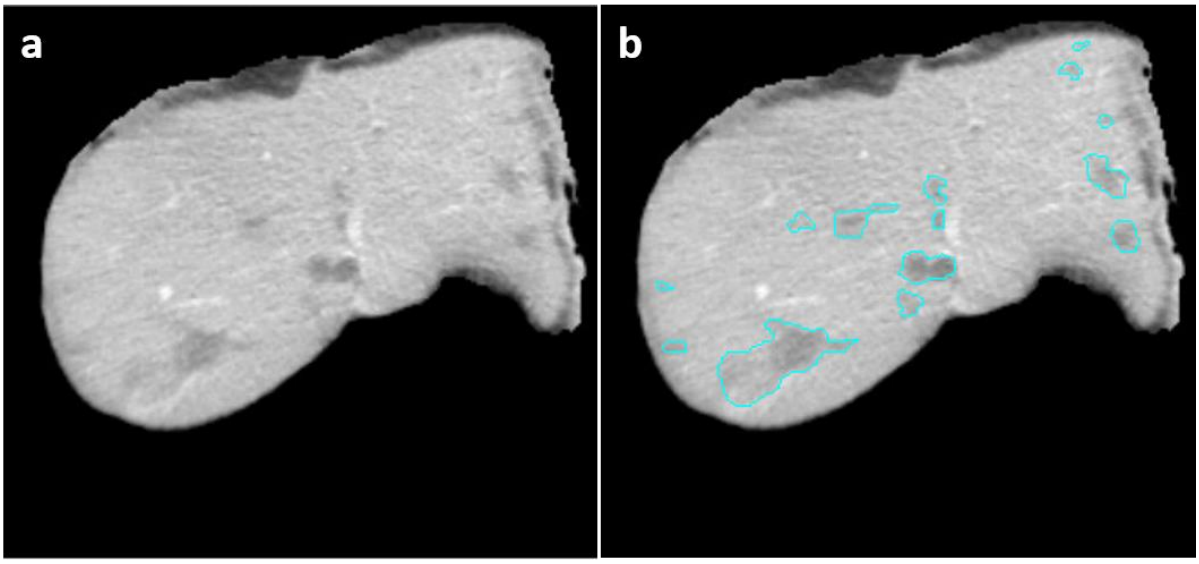



Figure 7.17 (a) The input CT image masked by the liver segmentation for a patient with liver metastases; (b) the same input CT image overlayed with manual lesion segmentations (blue contours).

As is shown in Figure 7.12, the high-resolution U-net model using the masked input had lower precision than the low-resolution U-net model using the unmasked input did for the validation CT datasets. The high-resolution input image retained more details of intensities than the low-resolution input image did, which resulted in more false positive detections of lesions. As is shown in Figure 7.16, the precision for the MR images is lower than that for the CT images. The number of lesions (mainly HCCs) in the MR images is generally lower than the number of lesions (mainly liver metastases) in the CT images. Therefore, the same number of false positive detections can result in lower precision for the MR images than for the CT images. The mean (median) false positive detections for the LiTS17 validation CT datasets, the SIRT validation CT datasets, and the test MR datasets using the generalized high-resolution U-net model were 4 (2), 4 (3), and 3 (3), respectively. The number of false detections that need to be removed by the physician is within the acceptable range, which will not consume much time.

As is shown in Figure 7.14 and Figure 7.15, the generalized high-resolution U-net using the masked input achieved comparable performance on CT lesion segmentation to the CT-specialized U-net using the masked input, and comparable performance on MR lesion segmentation to the MR-specialized U-net using the masked input. Training with multiple modalities did not help the CNN learn more useful information to improve its performance, compared to the CNN trained with a single modality, but it did not degrade its performance either. However, this allows the use of one single CNN model for lesion segmentation on both CT and MR images. The generalized U-net model achieved the same DSC for the validation CT images from the LiTS17 challenge as the highest DSC of 0.70 [72] for tumor segmentation achieved in the LiTS17 challenge.

It was found that the CT-specialized and MR-specialized high-resolution U-net models using the masked input achieved the same high DSC of 0.83 for the training CT and MR datasets, respectively. This proves that the high-resolution U-net structure can learn features to achieve a good performance on the training datasets. However, when the models were used to produce lesion segmentations for the unseen datasets, the MR-specialized model had poorer performance on MR images than the CT-specialized model did on CT images. This demonstrates that the MR-specialized model was more poorly generalized on MR images, compared to the generalization of CT-specialized model on CT images. The variations contained in the training MR datasets were not sufficient to prevent the CNN

from overfitting[6] on the seen datasets. Although there were 238 T1-weighted MR images for training, these images of different phases (the arterial, portal, or venous phase) were from the MR imaging studies of only 90 patients. Each CT image in the training CT datasets was from the imaging study of a different patient. Therefore, the training CT images contain more anatomical variations from different patients than the training MR images. The number of patients for the MR training datasets might not be able to provide sufficient variations for better CNN generalization on MR images. Therefore, CNN lesion segmentation on MR images is expected to be improved by collecting representative MR images from more patients for CNN training.

Due to the lack of sufficient datasets, no test datasets were separated from the total CT datasets. The CNN performed the learning process on the training datasets and the validation datasets were only used to find the optimal CNN parameters for unseen datasets. Therefore, the comparison of different CNN settings on the validation CT datasets can be used to evaluate the influence of these settings on the segmentation performance for unseen datasets, despite of a limited bias caused by model tuning with the validation datasets. Since liver metastases can also be segmented by thresholding the [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET and HCCs are mainly segmented on diagnostic MR images, lesion segmentation for MR images is more needed than for CT images in clinical routine. Therefore, no more new CT datasets were collected and annotated to evaluate CNN's performance on lesion segmentation for test CT datasets.

In this study, only T1-weighted MR images were used to train the CNN for lesion segmentation. In clinical routine, HCC detection is based on multi-modality MR images, including T1-weighted MR, T2-weighted MR, and diffusion-weighted MR. It might be helpful to concatenate multiple aligned MR modalities from one imaging study into one input volume for CNN training, which allows the CNN to integrate information from multiple modalities. The multi-modality MR images in one imaging study are usually automatically aligned during the scan acquisition, so no extra registration needs to be performed. However, the MR imaging protocol has much more parameters than the CT imaging protocol, which have big influence on MR appearance. The MR imaging protocol is not standardized for all patients. Besides, there are usually multiple sequences for each MR modality and the sequence types are not fixed for the imaging study of each patient. Therefore, it is challenging to select the sequence for each MR modality optimal for lesion segmentation, which needs more input from radiologists.

## 7.5 Conclusion

Masking the input image with the liver segmentation improves the performance of CNN liver lesion segmentation by helping the CNN to focus on learning features inside the liver. The input image at higher resolution preserves the information of small lesions, which helps improve CNN's performance on lesion segmentation and detection. The high-resolution U-net using the masked input achieved good performance on lesion segmentation for CT images, while its performance on MR lesion segmentation remains to be improved. HCCs in MR images appear with more intensity and texture variations than liver metastases in CT images, which makes CNN lesion segmentation for MR images more challenging. However, the high-resolution U-net using the masked input achieved very good performance on MR lesion detection (median sensitivity of 1.0), which shows its clinical value. The generalized U-net produces lesion segmentation on both CT and MR images, with a performance comparable to that of the U-net models specialized for a single modality. This proves the feasibility of

[6] Overfitting means that the CNN model achieved very good performance on the training datasets but very poor performance on unseen datasets. Overfitting occurs when the CNN model only learns features from specific datasets instead of general features.

using one single CNN to produce lesion segmentations for both CT and MR images, which is easier to be deployed in clinical routine.

# 8. Clinical workflows for SIRT planning and verification assisted with image segmentation and registration methods

## 8.1 Abstract

This chapter demonstrates the clinical workflow for SIRT planning and verification that is assisted by (semi-)automatic methods for liver (described in Chapter 3 and 4) and lesion segmentation (described in Chapter 7) and multi-modality image registration (Chapter 5 and 6). The workflow for treatment planning consists of seven steps: (1) image selection, (2) CNN liver segmentation and its verification, (3) CNN lesion segmentation and its verification, (4) Semi-automatic liver perfusion territory (LPT) delineation, (5) multi-modality deformable image registration, (6) mapping of liver and lesion contours and LPTs to the $^{99m}$Tc-MAA SPECT/CT, (7) prescription of injected activity. The workflow for treatment verification consists of five steps: (1) image selection, (2) CNN liver segmentation and its verification, (3) multi-modality deformable image registration, (4) mapping of liver and lesion contours and LPTs to the $^{90}$Y PET/CT or PET/MR, (5) absorbed dose calculation. The workflows are divided into different modules (steps), which can be called from MIM (MIM software Cleveland, USA) by the doctor based on their needs. This ensures enough flexibility in the workflows to deal with various situations in clinical routine. The order of CNN liver and lesion segmentation, LPT delineation, and multi-modality image registration can be swapped, depending on the doctor's preference for the individual patient. CNN liver segmentations on $^{99m}$Tc-MAA CT images from SIRT patients since 2021 were evaluated by looking into the number of the uncorrected CNN segmentations approved by the physician and the difference between the CNN liver segmentations and the verified segmentations. Around 76% (28/37) of the CNN liver segmentations were used for SIRT planning without any correction. The median and mean DSCs between the CNN liver segmentations and the verified segmentations are 1.0 and 0.98, respectively. Some representative examples of CNN liver and lesion segmentations are presented in this chapter. The application of image segmentation and registration methods allows the nuclear medicine physician to perform quantified and precise SIRT planning and verification by integrating precise VOIs (the liver, lesions, and LPTs) and activity distribution into a common space, requiring few manual interactions and workloads.

## 8.2 Introduction

SIRT relies heavily on multi-modality imaging studies for activity prescription and absorbed dose calculation, as is described in Section 1.5. The pre-treatment simulation of the $^{90}$Y activity distribution and the verification of the actual activity distribution are performed with $^{99m}$Tc-MAA SPECT and $^{90}$Y PET, respectively. The liver segmentation can be obtained from the $^{99m}$Tc-MAA CT, diagnostic CT, diagnostic MR, [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE CT, $^{90}$Y CT[7], or $^{90}$Y MR[8]. The lesion segmentation can be acquired from the diagnostic CT, diagnostic MR, or [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET, depending on the tumor type (see Section 1.5.4). The above multi-modal images need to be registered so that the segmentations and the activity map can be mapped to a common space for quantified

[7] The term "$^{90}$Y CT" is a shortcut for "the CT image acquired during the $^{90}$Y PET/CT imaging session", and similar terms for the other multi-modal systems. This slight abuse of notation will be used throughout this dissertation to improve the readability.

[8] The term "$^{90}$Y MR" is a shortcut for "the MR image acquired during the $^{90}$Y PET/MR imaging session". This slight abuse of notation will be used throughout this dissertation to improve the readability.

calculation. Manual liver and lesion delineation for precise segmentations is tedious and time-consuming, which is not applicable in clinical routine. Therefore, (semi-)automatic segmentation methods are needed. Besides, (semi-)automatic multi-modality image registration methods are needed for information integration.

A series of studies were performed in chapter 3, 4, 5, 6, and 7 to investigate the (semi-)automatic methods for liver and lesion segmentation and multi-modality image registration. These methods were integrated into the clinical workflows for SIRT planning and verification to reduce manual interactions as many as possible. The algorithms for CNN liver segmentation, CNN lesion segmentation, and multi-modality image registration were designed as individual modules. These modules can be assembled in different orders in the workflows based on the situations of different patients. In this way, the tasks for liver and lesion segmentation and image registration are standardized by using the (semi-)automatic methods, while enough flexibility is ensured to deal with various clinical needs. The physician can call different modules from MIM (MIM software Cleveland, USA), which is used as the working platform in clinical routine and provides good user interfaces.

In this chapter, the general clinical workflows for SIRT planning and verification assisted with image segmentation and registration methods are presented and discussed in detail. In practice, the workflows might vary slightly for each patient, depending on the available imaging studies.

## 8.3 Clinical workflow for SIRT planning

As is shown in Figure 8.1, the clinical workflow for SIRT planning consists of seven steps.

The first step is to select the appropriate image series from each imaging study ($^{99m}$Tc-MAA SPECT/CT, diagnostic CT or MR, [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET/CT). Diagnostic MR imaging is usually performed for HCC detection. Diagnostic CT or [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET/CT is usually used for liver metastases detection. The first step needs to be performed by the nuclear medicine physician manually in MIM.

In the second step, automatic liver segmentation can be performed on $^{99m}$Tc-MAA CT and/or diagnostic images (diagnostic CT/MR or [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE CT) by using the generalized U-net model in Chapter 4. CNN liver segmentation on $^{99m}$Tc-MAA CT is usually used to obtain the liver contour for SIRT planning. Although $^{99m}$Tc-MAA CT is generally non-contrast-enhanced, the quality of CNN liver segmentation on it is sufficient for clinical use. This can avoid the error caused by registering the liver segmentation on the diagnostic image to the $^{99m}$Tc-MAA CT. The CNN liver segmentation is always checked and corrected if deemed necessary by the physician for clinical use.

The generalized high-resolution U-net using the masked input (described in Section 7.3.2.2) is used for lesion segmentation on CT and T1-weighted MR images in the third step. CNN liver lesion segmentation on the diagnostic MR is performed for patients with HCC. For patients with liver metastases, an automatic thresholding tool for [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET, provided by MIM, is used to obtain the segmentation of [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE-avid tumors. The generalized high-resolution U-net can be performed to obtain the liver segmentation for all visible lesions on [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE CT. The lesions visible in the CT but invisible in the PET can be non-viable. If there is no [$^{18}$F]FDG or [$^{68}$Ga]Ga-DOTA-TATE PET/CT study for a patient, the diagnostic CT can be used for automatic liver lesion segmentation using CNN. The physician decides which imaging study is used for lesion segmentation based on the performed studies and the tumor type for a patient. The CNN lesion segmentation is checked or corrected by the physician for SIRT use.

Semi-automatic liver perfusion territory (LPT) delineation is performed using MIM's tool in the fourth step. MIM's LPT delineation tool requires the user to manually define a few lines dividing the

left and right LPTs at several axial slices, using cone-beam CT perfusion volumes [126]. After that, the tool can automatically fit a surface to divide the whole liver into the left and right LPTs based on the manually drawn lines. The $^{90}$Y microspheres are separately injected into different LPTs during treatment. The physician needs the LPTs to prescribe the injected activity for each LPT.

In the fifth step, all diagnostic images (diagnostic CT or MR, [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE CT) are registered to the $^{99m}$Tc-MAA CT by using MIM's multi-modality DIR tool mentioned in Chapter 6. Since the $^{99m}$Tc-MAA SPECT is automatically aligned to the CT during the imaging study, these registered diagnostic images are also automatically aligned to the $^{99m}$Tc-MAA SPECT.

In the sixth step, the verified VOIs of the liver and lesions segmented on the diagnostic images and the LPTs delineated on the CBCT are mapped to the $^{99m}$Tc-MAA SPECT by using the registration parameters generated for registering the diagnostic images to the $^{99m}$Tc-MAA CT. The mapping process is performed in MIM.

After the above six steps, the verified liver and lesion segmentations and the LPTs are mapped to the $^{99m}$Tc-MAA SPECT. The multi-compartment (described in 1.5.3.2) or voxel-level (described in 1.5.3.3) method can be used to prescribe the injected activity for each LPT based on the volume-level or voxel-level metrics for absorbed dose. The lower and upper bounds of the injected activity are computed based on the minimum absorbed dose for killing tumors and the maximum absorbed dose for preserving normal liver tissues, respectively. To achieve the best treatment, the lower bound should be no larger than the upper bound and the interval between two bounds should be as large as possible. The upper bound is usually determined as the injected activity, which achieves the highest absorbed dose to kill tumors while not damaging normal liver tissues. MIM automatically generates a dose report according to the given target absorbed dose.

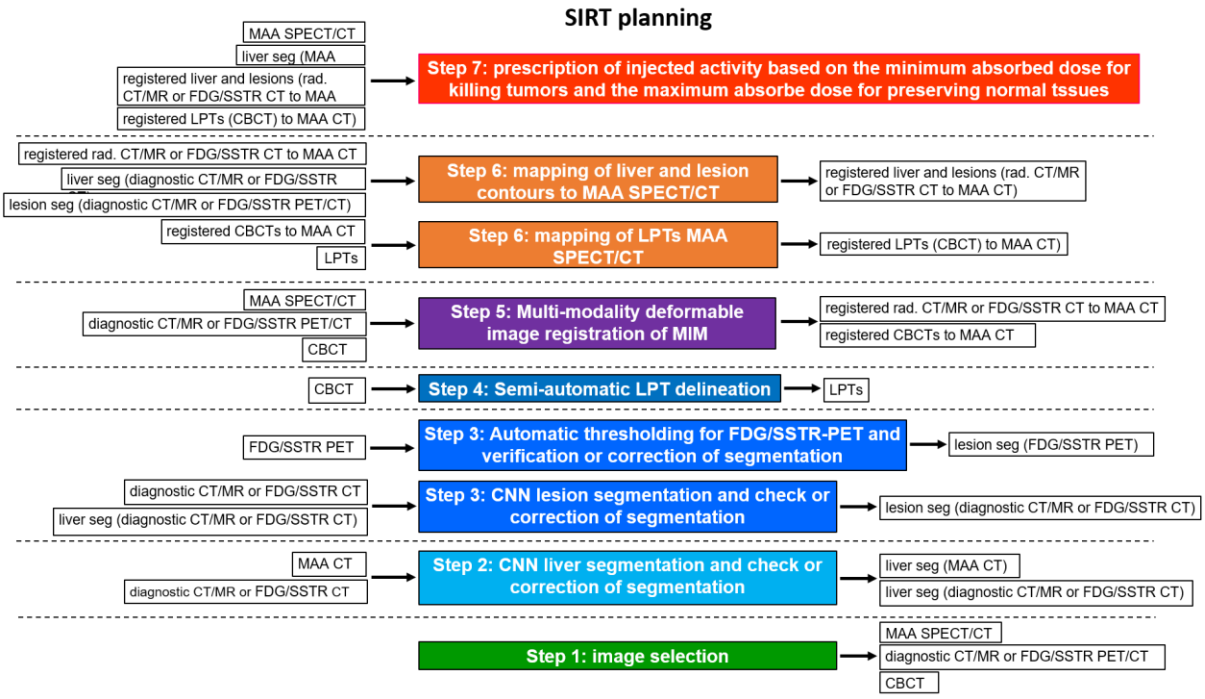



Figure 8.1 Overview of the clinical workflow for SIRT planning. "MAA", "FDG", and "SSTR" represent $^{99m}$Tc-MAA, [$^{18}$F]FDG, and [$^{68}$Ga]Ga-DOTA-TATE (a selective somatostatin analogue ligand, which has affinity for somatostatin receptor (SSTR)), respectively. "seg" represents segmentation. "LPT" represents liver perfusion territory.

## 8.4 Clinical workflow for SIRT verification

The clinical workflow for SIRT verification consists of five steps, as is shown in Figure 8.2.

In the first step, the appropriate series from the imaging study of $^{90}$Y PET/CT or PET/MR are selected. The selected series from the diagnostic imaging studies and CBCT study and their VOI delineations for SIRT planning are used for SIRT verification.

The generalized U-net model (described in Chapter 4) is used to produce liver segmentation on the $^{90}$Y CT or MR in the second step. The liver segmentation is checked and corrected if deemed necessary by the physician. The verified liver segmentation can be used for absorbed dose calculation. Besides, the verified CNN liver segmentation on the diagnostic images registered to the $^{90}$Y CT or MR can also be used for absorbed dose calculation. However, using the liver segmentation on the $^{90}$Y CT or MR can avoid the error caused by registration.

MIM's multi-modality DIR tool registers all diagnostic images and CBCTs to $^{90}$Y CT or MR in the third step. Since the CT or MR is automatically aligned to the PET during the imaging study, the registered diagnostic images and CBCTs can be automatically aligned to the $^{90}$Y PET.

In the fourth step, the verified segmentations of the liver and lesions and the LPTs are mapped to the $^{90}$Y PET by using the registration parameters for registering the diagnostic images and CBCTs to the $^{90}$Y CT or MR, respectively.

Since all verified segmentations and LPTs are mapped to the $^{90}$Y PET after the above four steps, the mean absorbed dose and cDVHs in the healthy liver tissues and the tumors can be computed in MIM by using the multi-compartment (described in 1.5.3.2) and voxel-level (described in 1.5.3.3) methods, respectively.

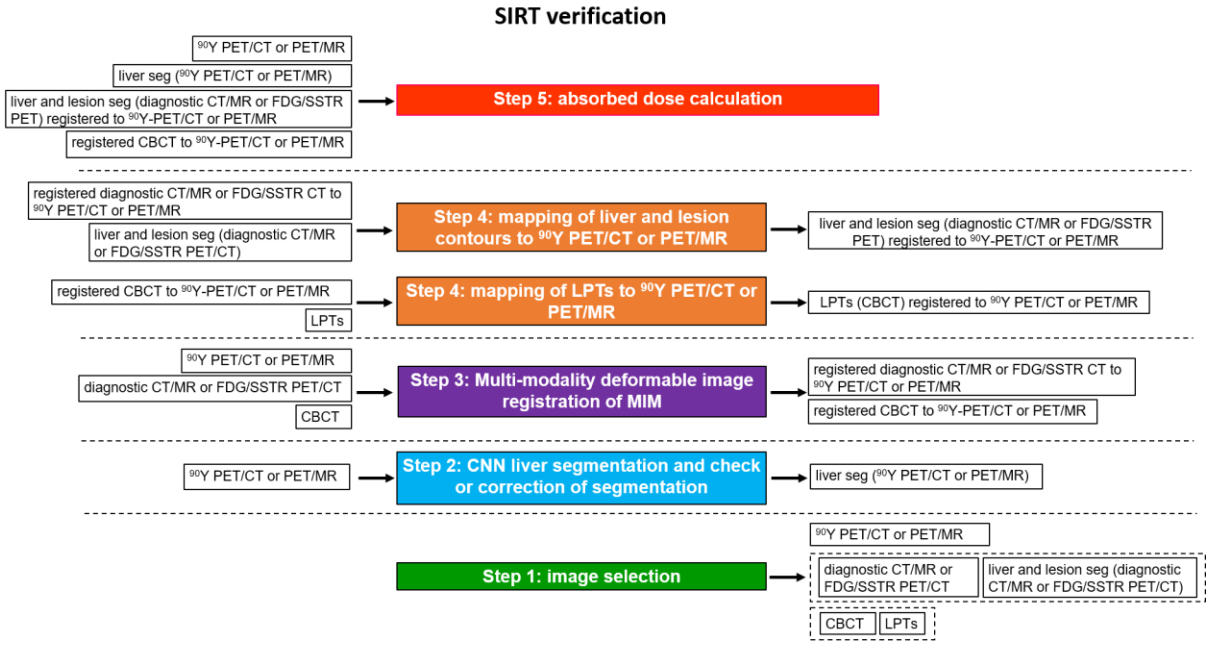



Figure 8.2 Overview of the clinical workflow for SIRT verification. "FDG" and "DOT" represent [$^{18}$F]FDG and [$^{68}$Ga]Ga-DOTA-TATE (a selective somatostatin analogue ligand, which has affinity for somatostatin receptor (SSTR)), respectively. "seg" represents segmentation. "LPT" represents liver perfusion territory.

## 8.5 Examples of CNN liver segmentation, CNN lesion segmentation, and multi-modality image registration

### 8.5.1 CNN liver segmentation in the clinical workflow

The generalized U-net model mentioned in Chapter 4 has been used for automatic liver segmentation in the hospital. The CNN liver segmentations on the $^{99m}$Tc-MAA CT images from the SIRT patients were checked or corrected by the nuclear medicine physician for SIRT planning. The $^{99m}$Tc-MAA CT is usually scanned with low dose and without contrast agents. The image quality is inferior to the diagnostic CT. It is challenging to segment these low dose and low contrast CT images. There were 37 $^{99m}$Tc-MAA CT images from 37 SIRT patients with CNN liver segmentations being collected since 2021. Around 76% of CNN liver segmentations were directly used for SIRT planning without any correction. The DSCs, MSDs, and HDs between the CNN liver segmentations and the verified CNN liver segmentations for the 37 CT images (either corrected or uncorrected) are presented in Figure 8.3. Two CNN liver segmentations on two $^{99m}$Tc-MAA CT images without correction needed and with correction are shown in Figure 8.4a and Figure 8.4b, respectively. The liver in Figure 8.4b was surrounded by ascites that have similar intensity and shape to the liver. It increases the difficulty for the CNN to distinguish the ascites from the liver.

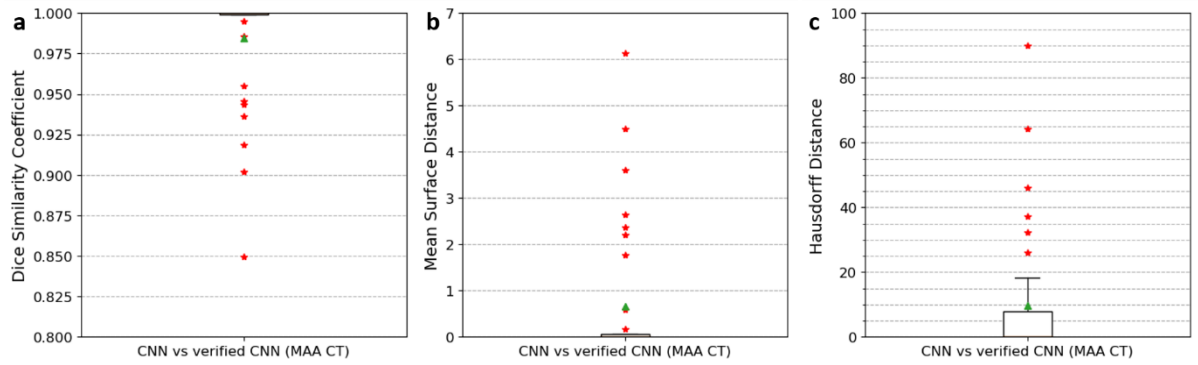



Figure 8.3 DSCs (a), MSDs (b), and HDs (c) between the CNN liver segmentations and the verified CNN liver segmentations for the 37 $^{99m}$Tc-MAA CT images. The green triangle, the yellow line, and the red stars represent the mean, median, and outliers. The yellow lines are invisible in the three pictures, because the median DSC is 1.0 and the median MSD and HD are both 0 mm.

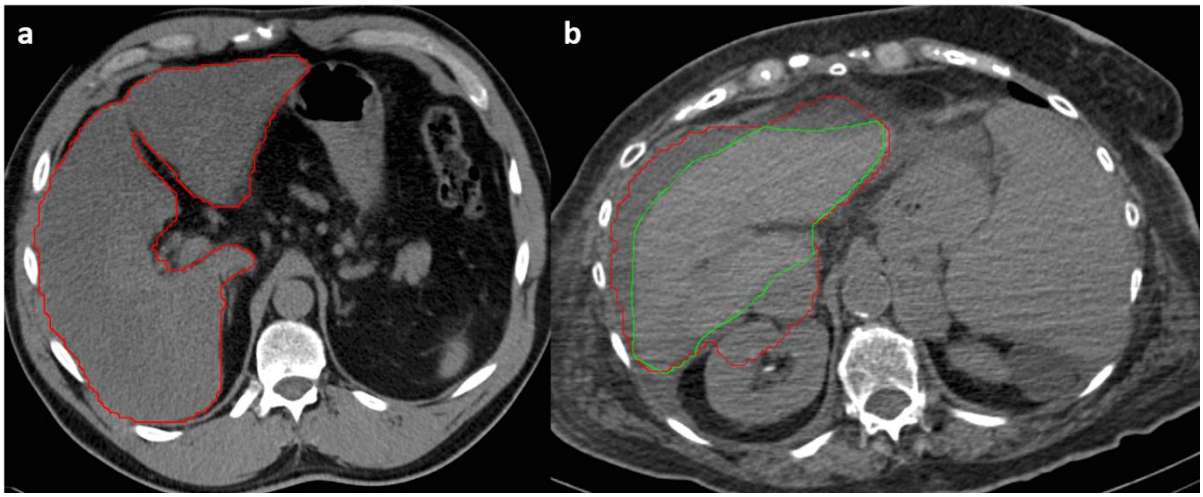



Figure 8.4 Examples of a CNN liver segmentation (DSC: 1.0) on a $^{99m}$Tc-MAA CT where no correction was deemed necessary (a) and a CNN liver segmentation (DSC: 0.85) on a $^{99m}$Tc-MAA CT that was corrected by the physician (b). The red and green contours represent the CNN liver segmentation and the corrected CNN segmentation, respectively.

### 8.5.2 CNN liver lesion segmentation on diagnostic CT and MR

The high-resolution U-net model described in Section 7.3.2.2 is used to generate liver lesion segmentations for diagnostic CT and T1-weighted MR images. Some examples of CNN lesion segmentations on CT and MR images are presented in Figure 8.5.

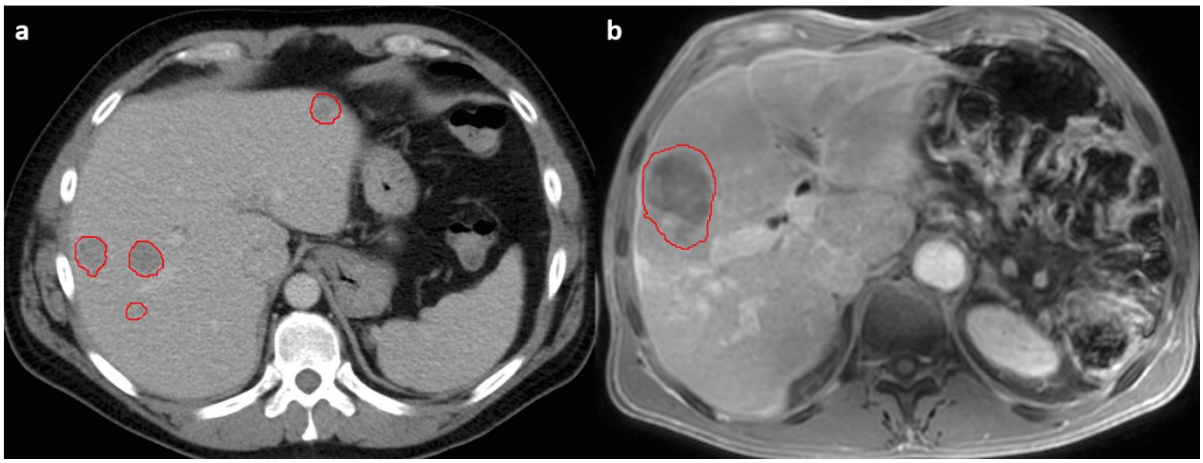



Figure 8.5 Examples of CNN liver lesion segmentations on diagnostic CT (a) and MR (b) images.

### 8.5.3 Multi-modality image registration

Two examples of registration of a diagnostic MR to a $^{99m}$Tc-MAA CT by using the multi-modality DIR of MIM (described in Chapter 6) and the CNN&LM-guided registration (described in Chapter 5) are presented in Figure 8.6 and Figure 8.7, respectively.

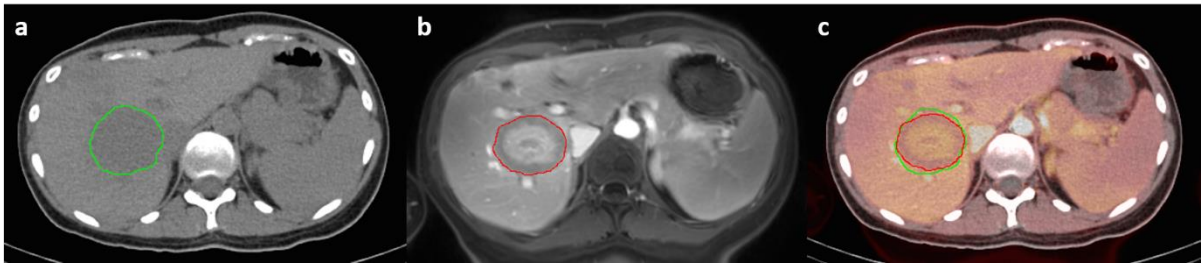



Figure 8.6 Example of a $^{99m}$Tc-MAA CT (a), a diagnostic MR registered to the $^{99m}$Tc-MAA CT by using the multi-modality DIR of MIM (b), and the fusion of the two images (c). The green and red lesion contours are delineated on the $^{99m}$Tc-MAA CT and the diagnostic MR, respectively.

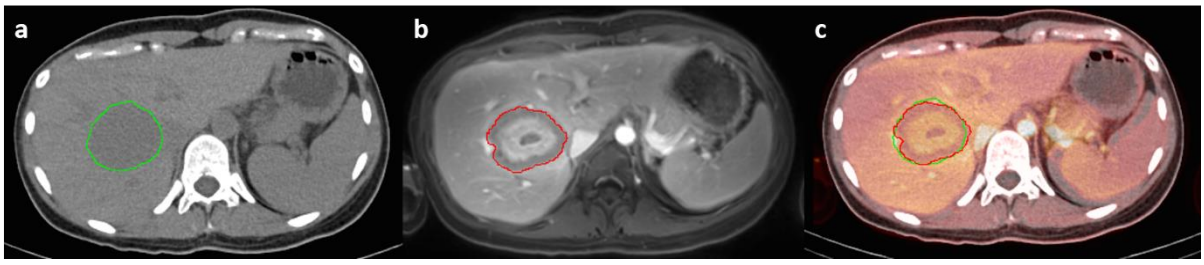



Figure 8.7 Example of a $^{99m}$Tc-MAA CT (a), a diagnostic MR registered to the $^{99m}$Tc-MAA CT by using the CNN&LM-guided registration method (b), and the fusion of the two images (c). The green and red lesion contours are delineated on the $^{99m}$Tc-MAA CT and the diagnostic MR, respectively.

## 8.6 Discussion

The clinical workflows for SIRT planning and verification are complicated and diverse. Before the dosimetry computation, several steps of medical image analysis need to be performed to acquire the VOIs (the liver, lesions) from multi-modality images and map them to the activity distribution map. Diagnostic MR images are performed to obtain lesion segmentations for patients with HCC, while diagnostic CT images or [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET/CT images are used for patients with liver metastases. The liver segmentation can be obtained from the diagnostic images, from the $^{99m}$Tc-MAA SPECT/CT image or from $^{90}$Y PET/CT or PET/MR image. Based on each patient's situation and the image quality, the physician needs to decide which images are selected for VOI delineation. Therefore, the workflows contain modules for (semi-)automatic VOI segmentation on different images (diagnostic CT or MR, [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET/CT, $^{99m}$Tc-MAA SPECT/CT, $^{90}$Y PET/CT or PET/MR), which can be called by the physician according to their needs. Besides, a good co-registration of multi-modality images is needed to integrate information from multiple images. The multi-modality image registration method is designed as an individual module. It has the flexibility to be performed before or after VOI segmentation. When the registration is performed before VOI segmentation, the physician can have pre-inspection of multi-modality images registered to each other, which can help the physician better decide the position and shape of tumors. By automating all steps of the medical image analysis, the VOI segmentations and multi-modality image registration can be used in the clinical workflows with few manual interactions and workloads. This also helps the standardization of the workflows.

The CNN liver segmentation has been made available to the clinical workflows as a research project for around two years. As is mentioned in Section 8.5.1, it constantly produces liver segmentations with reasonable quality, even on the low-contrast and low-dose $^{99m}$Tc-MAA CT images. Around 76% of CNN liver segmentations on $^{99m}$Tc-MAA CT images did not require any correction for SIRT use. CNN liver segmentation on the $^{99m}$Tc-MAA CT avoids the error from registering the liver segmentation on diagnostic images to the $^{99m}$Tc-MAA CT. Although the liver contours generated by CNN are not very smooth (see Figure 8.4) due to the up-sampling of the CNN output to the original image resolution, the contour quality is sufficient for SIRT use. If smoother contours are required for the clinical applications in the future, the contours can be improved by using a CNN with high-resolution inputs similar to the model described in Section 7.3.2.2.

The generalized high-resolution U-net using the masked input (described in Section 7.3.4) was developed for automatic liver lesion segmentation in the past year. It achieved a mean DSC of 0.70 for the validation CT datasets of the LiTS17 challenge, which was equal to the highest mean DSC achieved on the test CT datasets during the LiTS17 challenge [72]. Our high-resolution U-net was not evaluated on the test CT datasets of the LiTS17 challenge because the ground-truth lesion delineations from the expert for the test datasets were not published. The generalized high-resolution U-net should be able to produce liver lesion segmentations with reasonable quality for contrast-enhanced CTs. The segmentation quality for the U-net and the correction workload for clinical use remain to be further evaluated on clinical CT images. The generalized high-resolution U-net described in Section 7.3.4 achieved high sensitivities (median sensitivity: 1.0, mean sensitivity: 0.86) for lesion detection on the test T1-weighted MR datasets from our hospital, as is shown in Figure 7.13. However, the segmentation quality (a mean DSC of 0.52 on the test T1-weighted MR datasets) needs to be further improved. The lesions appear with more intensity and texture variations in the T1-weighted MR than in the CT. Therefore, training the U-net with T1-weighted MR sequences from more patients is expected to improve the performance on lesion segmentation. Besides, the combination of T1-weighted MR images with other MR modalities (diffusion-weighted MR images) can be investigated for potential improvement of lesion segmentation in the future.

The multi-modality DIR tool of MIM was investigated on the SIRT datasets for its applicability in SIRT, as is described in Chapter 6. It was proven to be clinically usable. Since guidance of liver and lesion segmentations was proven to improve the registration of multi-modality images in Chapter 5, it might be helpful to incorporate guidance of liver and lesion segmentations into the multi-modality DIR tool of MIM. Its feasibility and registration performance can be studied in the future.

## 8.7 Conclusion

The automatic liver and lesion segmentation methods using CNN and the multi-modality image registration method were designed as individual modules, which are integrated into the clinical workflows for SIRT planning and verification. The physician can call these modules from MIM in different orders according to the specific situation of each patient. The use of these modules standardize the process of VOI delineation and image registration. The delineated VOIs and registered images allow the physician to perform precise and personalized treatment planning and verification with as few manual interactions and an as rapid workflow as possible.

# 9. Final discussion

## 9.1 CNN liver segmentation

Automatic liver segmentation of CT and MR images is needed for precise liver contours to be used for dose calculation in personalized SIRT planning and verification. It is challenging due to the large variation in the liver shapes and textures for different individuals. This task becomes more challenging for clinical CT and MR images with pathological livers. Compared to the healthy liver, the pathological liver may contain many different kinds of lesions, including tumors, cysts, ablation sites, etc., which increase the shape, intensity, and texture variations in images. Besides, clinical images can be obtained with low dose protocols and hence lower quality, which further increases the difficulty in automatic liver segmentation. To solve this problem, a convolutional neural network (CNN) has been used in our studies. CNN consists of a stack of convolutional layers, which mimics the connected patterns of neurons. It can automatically learn the features for liver segmentation by being trained with images and ground-truth liver segmentations provided by experts.

### 9.1.1 Automatic liver segmentation using DeepMedic

In chapter 3, the first CNN structure, DeepMedic, was introduced for automatic liver segmentation of CT images. Its performance was evaluated with 40 SIRT CT datasets, by comparing the CNN liver segmentation with the manual liver segmentation and the adjusted CNN liver segmentation from the first expert. After that, the second expert was asked to perform manual segmentation and adjustment of CNN segmentations on 20 datasets randomly selected from the 40 SIRT CT datasets. The difference between the adjusted CNN segmentations from two experts was compared with the manual segmentations for the 20 datasets to evaluate DeepMedic's capability of reducing the inter-observer variability in liver segmentation.

The DeepMedic model achieved a good performance for liver segmentation of the SIRT CT datasets. The CNN liver segmentation achieved a median DSC of 0.94 with the manual segmentation and of 0.98 with the adjusted segmentation, respectively. There were only 2 out of the 40 SIRT datasets having a RV (described in Section 3.3.4) outside the range from 0.9 to 1.1 between the CNN segmentation and adjusted segmentation. When using the mono-compartment method (described in Section 1.5.3.1), there were 95% of the 40 datasets having a difference in the administered activity within 10% by using either the CNN liver segmentations or the adjusted segmentations. Only limited adjustment from the expert was needed for 35 out of the 40 CNN liver segmentations to be used in clinical routine SIRT planning.

The mean DSC (described in Section 3.3.4) between the adjusted segmentations was around 0.04 higher than the mean DSC between the manual segmentations from the two experts. The RVs between the adjusted segmentations were always between 0.9 and 1.1, while 20% of the RVs between the manual segmentations were outside the range of 0.9 and 1.1. This indicates that the difference in the adjusted segmentations from the two experts did not cause a difference in the administered activity over 10%, when using the mono-compartment model. CNN liver segmentation proves its ability to reduce the inter-observer variability of liver segmentation. It can be used as a starting point for adjustment to reduce the random and subjective errors in manual liver segmentations from different experts, which helps reduce the error in absorbed dose calculation caused by inconsistent liver segmentations from different observers.

The time for manual liver segmentation and adjustment of CNN liver segmentations by the two experts is shown in Figure 3.7. It is remarkable that the first expert spent more time on adjustment than on manual segmentation, whereas the second expert spent more time on manual segmentation. The first expert is a radiographer who has tens of years of experience in manual liver segmentation. He chose

the software clinically used for segmentation: the Siemens' Syngo MMWP Volume software. This software is not suitable for the adjustment of CNN segmentations. Therefore, he chose the MIM software (MIM software Cleveland, USA) for adjustment, which he had not used before. The second expert is a nuclear medicine physician who is not used to manual liver segmentation. He chose the MIM software for both manual segmentation and adjustment. We cannot claim that adjusting a CNN liver segmentation is always faster than manual liver segmentation for all experts. The time for manual segmentation and adjustment depends heavily on the software used for segmentation, the operator's experience in liver delineation, and the operator's talent for efficient use of the software. However, the second expert, as a beginner in providing manual organ segmentation, saved over half of the time for obtaining the liver segmentation by adjusting the CNN segmentation rather than segmenting the liver from the beginning. This proves the CNN's ability to enable the beginner to achieve good segmentation in a shorter time without much training, which reduces the dependency on well trained people for organ segmentation. We also believe that a skilled operator should be able to adjust a fairly good segmentation in a shorter time than needed for manual segmentation when the software is optimized for adjustment, because adjustment of CNN segmentation is simpler in principle.

However, the DeepMedic model tends to produce liver segmentations with right-angled edges, especially for CT images with low dose and low contrast (see Figure 4.3). Before being put into the CNN, the input image is down-sampled by three sampling factors (1, 5, 15) and small patches are randomly sampled from the images at three different resolutions (see Figure 3.1). The patches at three resolutions pass through multiple convolutional layers at three pathways to extract the local and global liver feature volumes, respectively. After that, the feature volumes at the two low resolutions are up-sampled by 5 and 15 to recover the original sampling, respectively. The up-sampled feature volumes from three pathways are fused by two convolutional layers to combine the local and global feature volumes, which might be insufficient to integrate features from the three pathways in some situations. When the CNN has difficulty in fusing the feature volumes from different pathways, the CNN output might be dominated by the up-sampled feature volumes from the low-resolution pathway. This can cause a mosaic effect in the CNN output, as is shown in Figure 4.3. After thresholding the CNN output, the liver segmentation will have right-angled edges on the boundary. To solve this problem, a better way of integrating features at the high- and low-resolutions needs to be found.

### 9.1.2 Automatic liver segmentation using U-net

To obtain smooth liver contours without sharp edges, a U-net model was developed and compared with the DeepMedic model for the performance on CT liver segmentation in Chapter 4. Besides, the U-net model was trained with CT images, T1-weighted MR images, and both CT and T1-weighted MR images to obtain the CT-specialized, MR-specialized, and generalized U-net models (described in Section 4.3.3). The performance of the three U-net models on liver segmentation for CT and T1-weighted MR images was compared to investigate the feasibility of producing both CT and MR liver segmentations using a single generalized model with comparable performance to the specialized models. The CT and MR datasets used for CNN training and validation are either from the public challenges or our hospital, as is shown in Table 4.1.

As is shown in Figure 4.2, the CT-specialized U-net and DeepMedic models achieved comparable mean DSC, MSD, and HD for the validation challenge CT datasets. For the validation CT datasets from our hospital, the CT-specialized U-net model improved the mean DSC, MSD, and HD by 0.01, 0.43 mm, and 7.6 mm compared with the CT-specialized DeepMedic model. The U-net model achieved liver segmentations with better quality than the DeepMedic model did for the CT images from our hospital. The clinical CT datasets from the hospital are usually images with low dose, low contrast, abnormal liver shapes, and high disease burden, which increase the difficulty in CNN liver segmentation. As is discussed in the last paragraph of Section 9.1.1, the DeepMedic model tends to produce liver segmentations with right-angled edges (see Figure 4.3) due to its drawback in feature integration. The

U-net model solves this problem in an elegant way. As is shown in Figure 4.1, the input image at the original resolution passes through two convolutional layers and the extracted feature volume is down-sampled by a relatively small sampling factor of 3. This process is repeated at the four resolution levels in the encoder (the left half of the U-net). The global liver features are extracted through multiple down-sampling and convolutions. After that, the global feature volume at the lowest resolution level is up-sampled by a sampling factor of 3. The up-sampled feature volume is integrated with the more local feature volume from the encoder at the current resolution level. This process is repeated at the four resolution levels in the decoder (the right half of the U-net). The feature volumes at different resolutions are combined gradually through multiple convolutions at each resolution level. This design ensures sufficient feature integration, which avoids the domination of the up-sampled low-resolution feature volumes in the CNN output. Therefore, the U-net structure was selected for the segmentation tasks in the following studies.

As is shown in Figure 4.4 and Figure 4.5, the CT- and MR-specialized U-net models had poor performance of liver segmentation on the image modalities that they were not trained with. The MR-specialized model achieved reasonable liver segmentations on CT images with a mean DSC of 0.90 for the validation challenge CT datasets and a mean DSC of 0.88 for the validation hospital CT datasets. However, the CT-specialized model often failed to produce acceptable liver segmentations for T1-weighted MR images with a mean DSC of 0.2 for the validation challenge MR datasets and a mean DSC of 0.48 for the validation hospital MR datasets. In general, MR images have more variations of intensity and texture than CT images. The CNN can learn features from MR images that can be transferred to CT images, while it is not true vice versa. Therefore, the MR-specialized model can achieve reasonable liver segmentation on CT images. According to Figure 4.6, the generalized U-net model achieved comparable performance on liver segmentation with the specialized U-net models for the image modalities they are specialized in. Generalizing the specialized models with another image modality does not degrade the performance for liver segmentation. Therefore, a generalized U-net model can be used to produce liver segmentations for both CT and T1-weighted MR images.

The generalized model achieved a mean DSC of 0.97 for the validation challenge CT datasets and a mean DSC of 0.96 for the validation hospital CT datasets, with all DSCs over 0.92. It achieved a mean DSC of 0.93 for the validation challenge and hospital MR datasets with the smallest DSC at around 0.80. CNN liver segmentation on MR images is more difficult than on CT images, because more intensity and texture variations exist in MR images. We used 169 CT and 140 MR image datasets, which are comparable amounts of datasets. The CNN might need more representative MR images to achieve liver segmentation quality for MR images similar to the segmentation quality for CT images.

It was difficult to collect sufficient CT and MR images with manual liver delineations for CNN training, because manual liver delineation is tedious and time consuming. Therefore, the total datasets were divided into the training and validation datasets, without test datasets. The training datasets were used to optimize the CNN parameters so that the CNN can learn liver features for good liver segmentation. In our experiments, the validation datasets were only used to evaluate the performance of different CNN structures and parameters on unseen datasets. The liver features were not directly learned from the validation datasets by the CNN. Therefore, the bias caused by CNN tuning with the validation datasets was limited. The main goal of our study was not to fully investigate the influence of different image modalities and CNN structures on CNN performance. Therefore, we did not continue the experiments by collecting and annotating more datasets for testing.

Despite the fact that the generalized U-net model was not evaluated on independent test datasets, its performance on clinical CT images was verified by the physician during SIRT planning and verification. The generalized model was used for automatic liver segmentation on the $^{99m}$Tc-MAA CT images from 37 SIRT patients since 2021 (Section 8.5.1). The nuclear medicine physician checked or corrected the CNN liver segmentations for use in the SIRT workflow. As is shown in Figure 8.3, around 76% of CNN liver segmentations were not corrected by the physician and directly used for SIRT

planning, with a de facto DSC at 1.0 between the CNN liver segmentation and the verified segmentation. The CNN liver segmentations and the verified segmentations had a mean DSC of 0.98 and the lowest DSC at 0.85. The $^{99m}$Tc-MAA CT images were scanned with low dose and without IV contrast, which was challenging for liver segmentation. The generalized U-net model demonstrated its good performance on clinical CT images with relatively poor quality.

In summary, CNN liver segmentation provides the physician a good starting point for segmentation correction, which helps to reduce the inter-observer variability caused by manual delineations from different observers. The U-net model can produce smoother liver contours than the DeepMedic model, which is more favoured for clinical application. It is feasible to train a generalized U-net model, which achieves comparable performance on liver segmentation to the specialized U-net models. The generalized U-net has been used for automatic liver segmentation in the clinical workflow. It achieved good liver segmentation on $^{99m}$Tc-MAA CT images with low dose and without intravenous contrast. Most of the generated segmentations did not need extra correction from the doctor.

## 9.2 Multi-modality image registration

In the SIRT workflow, the liver and tumors are usually delineated on the diagnostic CT image, diagnostic MR image, or [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET/CT, depending on the tumor type. The segmentations of the liver and tumors need to be mapped to the $^{99m}$Tc-MAA SPECT/CT for SIRT planning. The $^{90}$Y PET/CT or PET/MR can be registered to the $^{99m}$Tc-MAA SPECT/CT to investigate if the $^{99m}$Tc-MAA SPECT is a good estimation of the $^{90}$Y activity distribution. This needs to be achieved by registering the anatomical images (CT or MR) from the diagnostic imaging study and the $^{90}$Y PET/CT or PET/MR study to the anatomical image from the $^{99m}$Tc-MAA SPECT/CT study. Therefore, a (semi)automatic method for registration of CT and MR images is needed for SIRT use. The CT image from the $^{99m}$Tc-MAA SPECT/CT or the $^{90}$Y PET/CT study usually has low dose and low contrast. The MR from the $^{90}$Y PET/MR study is usually scanned without contrast agents and can have artefacts. It is challenging to perform multi-modality image registration for these clinical images.

In Chapter 5 and 6, the in-house image registration method, the registration methods guided by only CNN liver segmentations (CNN-guided) and both CNN liver segmentations and manually delineated landmarks (CNN&LM-guided), and the commercial multi-modality deformable image registration (DIR) tool of MIM were evaluated for registration of CT and MR images. All registration methods consisted of affine registration followed by non-rigid registration. These methods were tuned and evaluated with 49 CT and 26 MR images from 20 SIRT patients (described in Section 5.3.1). Each patient case contained one $^{99m}$Tc-MAA CT as reference and at least one CT and one MR (the floating image, registered to the reference image) from the diagnostic CT or MR study, the [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET/CT study, or the $^{90}$Y PET/CT or PET/MR study. Manually delineated liver and landmarks (including lesions) on each image were used for registration evaluation. Each liver registration was evaluated with RMSD and all registrations for each method were evaluated with RRMSD (described in Section 5.3.3). The lower the RMSD and RRMSD are, the less registration error is. The dosimetry metrics, including mean absorbed dose, V70, and V100, were used to evaluate the influence of registration errors on dose estimation (describe in Section 5.3.4). The dosimetry metrics were calculated on the $^{99m}$Tc-MAA SPECT by using the lesion contours delineated on the $^{99m}$Tc-MAA CT (reference lesions) and the lesion contours delineated on the floating CT and MR and registered by different methods (registered floating lesions). The dosimetry metrics computed by using the registered floating lesions were compared with the metrics computed by using the reference lesions to evaluate the dose estimation difference caused by different registration methods.

### 9.2.1 Evaluation of registration errors

Using CNN liver segmentations for the affine registration greatly improved registration performance compared with the affine registration of images. As is shown in Figure 5.6 and Figure 5.7, the CNN-based affine registration improved the RRMSD by 3.1 mm (27%) and 3.4 mm (34%) compared with the image-based affine registration for the floating CT and MR images, respectively. The affine registration of CNN liver segmentations excludes the negative influence of organs and tissues outside the liver, which gives the liver in the floating image flexibility to match the reference liver as a whole object through affine transformation. The CNN-based affine registration provides the following non-rigid registration a good starting point, which is critical for the non-rigid registration. A good affine registration can prevent large deformations during the non-rigid registration.

For CT to CT registration, the CNN-guided non-rigid registration decreased the RRMSD by 0.1 mm (1%) and the image-based non-rigid registration increased the RRMSD by 0.4 mm (5%), compared to the CNN-based affine registration (see Figure 5.6). The image intensity information does not help to improve the CNN-based affine registration. Given a good CNN-based affine registration, the improvement by the CNN-guided non-rigid registration is limited. This demonstrates that the segmentations provide more efficient information for good registration than the image intensities. For MR to CT registration, the use of images and CNN liver segmentations for the non-rigid registration did not improve the performance of the CNN-based affine registration. As a result, the optimal registration procedure did only use the CNN-guided registration[9]. This might be caused by the relatively poorer CNN liver segmentation on MR images than on CT images (see Figure 5.5). Some CNN liver segmentations on MR images were found to miss low-intensity lesion regions in the liver. During the CNN-guided non-rigid registration, the surface of the CNN segmentation on the floating MR was deformed to match the surface of the CNN segmentation on the reference CT. This resulted in severe spurious deformation of the floating lesion, which increased the RRMSD.

The CNN-guided registration improved the RRMSD by 1.0 mm (11%) and 3.4 mm (34%) compared with the image-based registration[10] for the CT to CT and MR to CT registrations, respectively (see Figure 5.6 and Figure 5.7). Although the CNN liver segmentations were not perfect, these unedited segmentations proved their ability to guide the registration for better performance. This allows the full automation of the CNN-guided registration without the need for manual correction of CNN liver segmentations.

As is shown in Figure 5.6 and Figure 5.7, landmark guidance improved the RRMSDs of the CNN-guided CT to CT and MR to CT registrations by 2.1 mm (26%) and 1.4 mm (21%), respectively. It demonstrates that landmark guidance can improve the registration performance. The landmarks used for registration were manually delineated due to the lack of automatic lesion segmentation at that time. This might cause a self-fulfilling effect for registration evaluation. However, these manually delineated landmarks can be regarded as perfect automatic lesion segmentations. This indicates the feasibility of developing automatic lesion segmentation for image registration guidance. Besides, manual lesion segmentations verified by the physician are clinically usable.

The multi-modality DIR tool of MIM achieved comparable performance to the CNN&LM-guided registrations with regard to registration errors. As is shown in Figure 6.1, the absolute RRMSD difference between the multi-modality DIR and the CNN&LM-guided registration was 0.3 mm for both CT to CT ($p = 0.89$) and MR to CT ($p = 0.41$) registrations.

[9] The CNN-based affine registration and CNN-guided non-rigid registration.
[10] The image-based affine and non-rigid registration.

## 9.2.2 Dose Estimation

As is shown in Figure 6.3, the mean doses between the reference lesions and the registered floating CT and MR lesions had a strong correlation ($r > 0.9$) for the CNN-guided registration method, the CNN&LM-guided registration method, and the multi-modality DIR of MIM. Mean dose measures the average absorbed dose within a volume, which is less sensitive to the contour changes than the voxel-level dosimetry.

For the voxel-level dosimetry, 83% (69%), 79% (76%), and 83% (69%) of the reference lesions had an absolute V70 (V100) difference smaller than 10% with the corresponding floating CT lesions registered by the CNN-guided method, the CNN&LM-guided method, and the multi-modality DIR of MIM, respectively. The reference and the registered floating CT lesions had a strong correlation ($r \geq 0.94$) of V70 and V100 for the three registration methods. The three methods achieved similar performance on CT to CT registration. For MR to CT registration, the CNN-guided method, the CNN&LM-guided method, and the multi-modality DIR of MIM had around 61% (61%), 70% (70%), and 70% (74%) of lesions with an absolute V70 (V100) difference smaller than 10% between the reference and registered floating lesions, respectively. The CNN-guided registration had a weaker correlation of V70 and V100 ($r < 0.82$) between the reference and the registered floating MR lesions than the other two methods. For MR to CT registration, landmark guidance helped decrease the dose estimation difference caused by registration errors from the CNN-guided method. The multi-modality DIR tool of MIM achieved a similar dose estimation difference with the CNN&LM-guided registration.

The lesion volume can mitigate or exaggerate the dose estimation difference caused by registration errors. It was found that most lesions with a volume larger than 50 cc had a relative mean dose difference and an absolute V70 and V100 difference smaller than 10%. Larger lesion volumes have higher tolerance of registration errors than smaller lesion volumes with regard to dose estimation difference. A small shift in a few voxels can become relatively large in proportion to a small lesion size, which can lead to a large change in the activities included by the lesion contour. Besides, a lesion can appear with various shapes and volumes in images of different modalities, because of different lesion information reflected by different images or tumor growth. Therefore, the registration error is not the only cause of dose estimation difference. Despite of good lesion registration, the dose estimation difference can be large due to the small lesion size and large shape and volume difference of lesions in two images. Each image modality exhibits a different characteristics of lesion appearance, which makes it difficult to eliminate the shape and volume difference of a lesion between two images. Therefore, a joint lesion delineation on co-registered multi-modality images is helpful to approach the ground-truth delineation by considering all lesion characteristics reflected by different images.

Small dose estimation difference does not necessarily mean good lesion registration. As is shown in Figure 5.13c and Figure 5.13d, the floating MR lesion (green) registered by the CNN-guided method did not have a good overlap with the reference lesion (red). The cDVHs for the reference lesions and the registered floating MR lesions are similar, because both lesion contours include most part of the high-uptake region. The reference and registered floating lesions have a mean dose difference of -5.2%, a V70 difference of -1.0%, and a V100 difference of -8.4%. This case shows that a small dose estimation difference can occur despite of poor lesion registration, as long as a similar volume of low- and high-uptake regions is included in the reference and registered floating lesions. The true energy deposition for each voxel in the reference and registered floating lesions is not necessarily reflected by the dosimetry metrics (mean dose, V70, and V100).

## 9.2.3 Registration workflow

In the standard clinical workflow for SIRT planning, the tumors should be delineated on the diagnostic CT, the diagnostic MR, and the [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET/CT. After that, they are mapped to the $^{99m}$Tc-MAA SPECT/CT through manual or semi-automatic registration tools. In

practice, the nuclear medicine physician may prefer to delineate lesions directly on the $^{99m}$Tc-MAA SPECT through thresholding to shorten the processing time and limit the risk of inter-modality mis-registration, without the need for registration. However, good registration of diagnostic images to the $^{99m}$Tc-MAA SPECT/CT can promote lesion delineation on diagnostic images and correlation to the findings in the $^{99m}$Tc-MAA SPECT, which are recommended in recent international SIRT guidelines [128]. The physician's interaction is needed for the above standard workflow during the whole process. The standard workflow takes around 30 to 45 min in general.

The CNN-guided registration workflow consists of automatic liver segmentation using CNN and the CNN-guided registration. CNN liver segmentation needs to be performed on the reference and floating images. The CNN model takes no more than 5 min to produce one liver segmentation using a CPU-based computation server. The generated CNN liver segmentations require the physician's check-up or correction, which needs 1 to 5 min. A CPU-based server is used to run the CNN-guided registration algorithm without parallel computation, which consumes around 15 min for each registration. The whole CNN-guided registration workflow, which takes around 20 min in total, is automated except for the physician's verification on CNN liver segmentations. The time for manual interaction is no more than 5 min, which is acceptable for clinical use. However, the CNN-guided method had poorer performance on MR to CT registration than the CNN&LM-guided method and the multi-modality DIR of MIM, especially when performing the voxel-level dosimetry.

The CNN&LM-guided registration workflow requires lesion segmentations before registration. Lesion delineation is manually performed by using MIM's delineation tool, which costs around 10 min and needs to be automated in the future. The other steps in the CNN&LM-guided registration workflow are the same as the steps in the CNN-guided registration workflow. In total, the CNN&LM-guided registration workflow takes around 30 min. Manual interaction costs around 15 min in the workflow.

It takes around 5 min for the multi-modality DIR tool of MIM to produce a registration of two images, which is fully automated. The multi-modality DIR achieved comparable performance on CT to CT and MR to CT registration to the CNN&LM-guided registration. No segmentations of the liver and lesions are needed in advance to guide the DIR registration, which saves the time for VOI segmentation on multiple images. It is easier to deploy the multi-modality DIR of MIM in the clinical workflow than the CNN&LM-guided registration.

In summary, the CNN&LM-guided registration method and the multi-modality DIR of MIM achieved comparable performance with regard to registration errors, while the CNN-guided registration had larger registration errors than the other two methods. The three methods had comparable performance with respect to dose estimation difference caused by registration errors for the CT to CT registration. For MR to CT registration, the multi-modality DIR achieved comparable performance on dose estimation difference to the CNN&LM-guided registration and better performance than the CNN-guided registration. The three registration methods are applicable in the volume-level and voxel-level dosimetry for registration of CT images. For MR to CT registration, the CNN&LM-guided method and the multi-modality DIR can be used for the volume-level and voxel-level dosimetry. In case that the CNN&LM-guided method and the multi-modality DIR have comparable performance, the latter has advantage of being applied to the clinical workflow without the need for VOI segmentation.

## 9.3 CNN liver lesion segmentation

Automatic liver lesion segmentation is critical for accurate lesion delineations to be used for precise and personalized SIRT planning and verification. It is challenging due to the large intensity, shape, and texture variation of lesions. Lesions in MR images exhibit more intensity and texture variations than lesions in CT images, which makes automatic lesion segmentation for MR images more difficult than for CT images.

In chapter 7, different CNN structures and input settings were evaluated on CT images to find the optimal CNN structure (high-resolution and low-resolution U-net models) and input setting (masking the input image through its liver segmentation or not) for lesion segmentation. The optimal CNN structure (the high-resolution U-net) and input setting (masking the input image through its liver segmentation) were applied to lesion segmentation for T1-weighted MR images. After that, the CNN model was generalized by training it with both CT and T1-weighted MR images. The performance of the generalized model was compared with the performance of the specialized models to investigate the feasibility of using one single model to produce lesion segmentations for both CT and MR images without degrading segmentation performance. The CT datasets from the LiTS17 challenge and SIRT patients in our hospital were used for CNN training and validation, and the MR datasets from our hospital were used for CNN training, validation, and test (see Table 7.1).

The low-resolution U-net described in Section 7.3.2.1 was trained with either the CT images masked by the dilated manual liver segmentations or the unmasked CT images to evaluate the influence of masking the input image on lesion segmentation performance. The mean DSCs were improved by 0.08 and 0.03 for the LiTS17 and SIRT validation CT datasets, respectively, after training the CNN with the masked images. The CNN is able to focus on learning features inside the liver by masking the input image through a liver mask. Although the liver mask is manually delineated in the experiment, CNN liver segmentation can also be used. The liver mask does not need to follow the liver boundary precisely. As long as the CNN liver segmentation does not miss a large part of the liver, the time for correcting the CNN segmentation should be limited.

Small lesions can become invisible in the input image at a low resolution ($3 \times 3 \times 3$ mm$^3$). Therefore, a high-resolution U-net described in Section 7.3.2.2 was developed to process the input images at a higher resolution ($1 \times 1 \times 3$ mm$^3$), which can better preserve the details of small lesions. The low-resolution and high-resolution U-net models using the masked input were trained and validated with the same CT datasets to evaluate the influence of the input image resolution on lesion segmentation performance. The high-resolution U-net achieved mean DSCs 0.03 and 0.05 higher than the mean DSCs for the low-resolution U-net for the LiTS17 and SIRT validation CT datasets, respectively. Besides, the high-resolution U-net improved the mean sensitivities for the LiTS17 and SIRT validation CT datasets by 0.07 and 0.16, respectively, compared to the low-resolution U-net. This proves that input images at higher resolution help improve the performance on lesion detection and segmentation.

The high-resolution U-net using the masked input achieved optimal performance on lesion segmentation for CT images. Therefore, it was trained, validated, and tested with T1-weighted MR images (see Table 7.1). The model achieved a median DSC of 0.60 and a mean DSC of 0.48 for the test MR datasets. The high-resolution U-net had poorer performance on MR images than on CT images. In SIRT, MR image are used for HCC detection while in our center CT images are used for liver metastases. HCCs in MR images can appear diffuse and heterogeneous, which increases the intensity and texture variations. Liver metastases are usually more homogenous in CT images. Therefore, lesion segmentation on MR images is more difficult than on CT images. Besides, multiple MR images at different phases (the arterial, portal, or venous phase) were selected from the MR imaging study of the same patient for our experiment, while each CT image was from a different imaging study. Therefore, the training MR datasets contain fewer anatomical variations from different patients than the training CT datasets. The CNN trained with fewer data variations is more prone to overfitting, which might result in poor performance on unseen datasets.

The high-resolution U-net using the masked input was trained with both CT and T1-weighted MR images for model generalization on multiple modalities. The generalized high-resolution U-net achieved comparable performance on CT lesion segmentation to the high-resolution U-net trained with CT images (CT-specialized), and comparable performance on MR lesion segmentation to the high-resolution U-net trained with T1-weighted MR images (MR-specialized). Training the CNN with multiple modalities does not help improve lesion segmentation performance. However, this experiment

proves the feasibility of using a single U-net to produce lesion segmentations for both CT and MR images, which are comparable to lesion segmentations generated by the specialized models. A single generalized CNN model is easier to be implemented in clinical routine than two CNN models for different image modalities.

The CT-specialized high-resolution U-net using the masked input had a lower sensitivity for the validation CT datasets than the MR-specialized high-resolution U-net using the masked input did for the test MR datasets (see Figure 7.11 and Figure 7.13b). In SIRT, CT and MR images are used for detection of liver metastases and HCCs, respectively. The number of metastases in a liver can be much larger than the number of HCCs and liver metastases can be dispersed throughout the liver (see Figure 7.17). Small metastases usually have similar intensities with the surrounding liver tissue, which makes it difficult to detect them. The sensitivity became small when the CNN missed these small lesions. If the CNN segmented the relatively large metastases, the DSC could still remain large. For MR images, the CNN was able to segment at least a small part of HCCs despite of the heterogeneous and diffuse appearance of some HCCs. This could result in a high sensitivity but a low DSC. The tumors with very small sizes are usually not used to calculate the injected activity in SIRT or can be easily added manually. Therefore, failed CNN segmentation for these small tumors is acceptable for SIRT.

In summary, masking the input image through the liver segmentation helps the CNN focus on learning features of the liver, which increased the performance of CNN lesion segmentation for CT images. Small lesions can be barely visible in the input image at low resolution. The high-resolution U-net was developed to process the image at higher resolution, which improved the performance of CT lesion segmentation compared to the low-resolution U-net. Good lesion segmentation performance for CT images was achieved by the high-resolution U-net using the masked input, which had comparable performance on the LiTS17 CT datasets to the best-performing model during the LiTS17 challenge. The high-resolution U-net using the masked input image achieved poorer performance on lesion segmentation for T1-weighted MR images than it achieved for CT images. Lesions in MR images have more intensity and texture variations than lesions in CT images, which increases the difficulty in MR lesion segmentation. MR images from more patients are needed to improve the performance of CNN lesion segmentation for this modality. The generalized high-resolution U-net trained with both CT and T1-weighted MR images achieved comparable performance with the specialized high-resolution models. This proves the feasibility of using one single CNN to produce lesion segmentation for both CT and MR images, which is easier to be implemented in clinical routine than using two specialized models.

## 9.4 SIRT workflow

Good liver and lesion segmentations are needed for accurate absorbed dose calculation in SIRT planning and verification. Since the VOI segmentations are performed on the diagnostic images, it is critical to register these diagnostic images to the $^{99m}$Tc-MAA SPECT/CT or the $^{90}$Y PET/CT or PET/MR, so that the VOI segmentations can be directly used on the activity distribution map. Manual VOI segmentation and registration of multi-modality images are tedious and time-consuming. Therefore, it is important to develop (semi-)automatic liver and lesion segmentation and multi-modality image registration methods, so that accurate and personalized SIRT planning and verification can be achieved without increasing too much manual workflow for the doctor. These automatic methods are designed in individual modules which can be called from the MIM software by the physician (see Figure 8.1 and Figure 8.2). To deal with various situations of patients, the order of these modules to be used in the SIRT workflow can be changed according to the physician's need. This promotes the standardization of each image processing task while ensuring enough flexibility in the workflow.

CNN liver segmentation has been applied in the clinical SIRT workflow to obtain liver segmentation. As shown in Figure 8.3, around 76% of CNN liver segmentations for 37 $^{99m}$Tc-MAA CT

images were used in the SIRT workflow without any correction since 2021. Most corrections for CNN liver segmentations were minor, consuming little time from the operating physician.

By comparing the performance of the multi-modality DIR with the performance of the CNN&LM- and CNN-guided registration methods, the multi-modality DIR of MIM proves its clinical value for SIRT. It achieved comparable performance to the CNN&LM-guided registration and does not require liver and lesion segmentation in advance of registration. Therefore, the multi-modality DIR is applied for multi-modality image registration in the clinical workflows for SIRT planning and verification.

CNN lesion segmentation achieved comparable DSC for the challenge CT datasets to the highest DSC of 0.70 achieved during the LiTS17 challenge. Although the CNN's value for clinical application remains to be fully studied, it can be tested on the diagnostic CT images from the pre-treatment studies of SIRT. By applying the CNN to the SIRT workflow, we can promote the process for verifying its clinical value. For CNN lesion segmentation on MR images, the performance still needs to be improved for producing segmentations of good quality. However, the CNN achieved high sensitivity for lesion detection. It would be valuable to apply the CNN in the clinical workflow, which can assist the physician for lesion detection on MR images.

## 9.5 Limitations and future perspectives

### 9.5.1 CNN liver segmentation

The generalized U-net model (described in Section 4.3.2) achieved good liver segmentation on CT and MR images and has been applied in the clinical workflow. As is shown in Figure 8.4, there were still jagged edges on the boundary of the CNN liver segmentation. CNN liver segmentation was performed on the input image resampled to an isotropic voxel size of 3 mm. After that, the thresholded CNN output was resampled to the original resolution of the image, whose image in-plane pixel size was usually around 1 mm. This results in the jagged edges in the up-sampled CNN liver segmentation. However, these up-sampled liver segmentations are sufficient for SIRT planning and verification, which do not require very smooth liver contours. If there is higher clinical demand for the contour smoothness in the future, the high-resolution U-net mentioned in Section 7.3.2.2 can be used to produce liver segmentation on the input image at higher resolution. The original image can be resampled to an anisotropic voxel size of $1 \times 1 \times 3$ mm$^3$ and patches with smaller size can be randomly sampled from the resampled image. By training the high-resolution U-net with these image patches, it should be able to produce good liver segmentation at higher resolution. After resampling the CNN output to the original image resolution, there should be few jagged edges on the boundary.

### 9.5.2 Multi-modality image registration

Through the study in Chapter 5, it was found that guidance by segmentation of the liver and lesions can help reduce the registration errors. The current multi-modality DIR tool of MIM is only based on images, and achieved comparable performance with our CNN&LM-guided registration method. It might be helpful to incorporate the segmentations of the liver and lesions into the multi-modality DIR for further registration improvement, since the registration precision of the liver and lesions is critical for SIRT. CNN liver segmentations can be used for the affine registration step of the DIR tool, which can help the tool focus on registration of the livers instead of the whole FOV. After that, the segmentations of the liver and lesion segmentations for the reference and floating images can be used to guide the deformable registration of the multi-modality DIR. The liver segmentation used for registration can be either directly from the CNN or the corrected CNN segmentation, depending on the quality of CNN liver segmentation. The lesion segmentation can be acquired manually or automatically, depending on the quality of automatic lesion segmentation. When CNN lesion segmentation is further improved, the acquisition of lesion segmentation can be fully automated.

Our registration methods use mutual information as the similarity metric during optimization. However, mutual information, as a hand-crafted similarity metric, does not always ensure good registration of structures in multi-modal images. Therefore, segmentations of structures in the images are needed to guide registration of these structures. Another possible solution is to use deep learning to learn a similarity metric which better matches registration of multi-modal images. Some studies have been performed to obtain similarity metrics from deep learning. Cheng *et al*. proposed a deep learning network to classify the input pairs of CT and MR as aligned or not [135]. They used the network output, which is a probabilistic value, as the similarity score. Haskins *et al*. developed a CNN to predict the target registration error between ultrasound and MR images [136]. They used the predicted target registration error as the similarity measurement. To learn a similarity metric, well-registered image pairs are required. In reality, it is difficult to obtain these image pairs. Besides, it is difficult to ensure the smooth derivative of these learned similarity metrics for optimization.

Registration of multi-modal images is more difficult than registration of mono-modal images. The registration problem will be simplified if multi-modal images can be transformed into mono-modal images. Generative adversarial network (GAN) has been used for translating images to another modality in recent studies. Salehi *et al*. used GAN to transform T1-weighted MR images to T2-weighted MR images [137]. After that, a CNN was trained to register T2-weighted MR images for fetal brain. Tanner *et al*. used cycle-GAN to translate multi-modal images into mono-modal images and use a mono-modal image similarity measure to register CT and MR [138]. However, the accuracy of intensity mapping between CT and MR remains to be further studied, which is crucial for registration accuracy.

In the SIRT routine, treatment planning is always performed for each SIRT patient while treatment verification is not always performed, depending on the clinical need. The dosimetry calculation is mainly performed on the $^{99m}$Tc-MAA CT during treatment planning. Therefore, the $^{99m}$Tc-MAA CT was chosen as the reference image in our registration studies. For treatment verification, the $^{90}$Y PET/CT or PET/MR can be registered to the the $^{99m}$Tc-MAA CT. In this way, the registered $^{90}$Y PET/CT or PET/MR and the VOI segmentations mapped to the $^{99m}$Tc-MAA CT during treatment planning are in the same space. However, the registration errors are introduced into both the registered VOIs and the registered activity maps ($^{90}$Y PET). It would be better to map the VOI segmentations directly to the $^{90}$Y PET by registering the diagnostic images to the $^{90}$Y PET/CT or PET/MR, so that registration errors are only introduced to the registered VOIs. Therefore, it would be valuable to evaluate the registration performance and the dose estimation difference caused by registering the diagnostic images to the $^{90}$Y CT or MR. We would expect similar performance by using either the $^{99m}$Tc-MAA CT or the $^{90}$Y CT as reference, since both images are non-contrast-enhanced. However, the registration of the diagnostic images (CT and MR) to the $^{90}$Y MR would need more careful study.

### 9.5.3 CNN liver lesion segmentation

Due to the lack of sufficient CT images, no test datasets were separated from the total datasets. The CNN learned liver features from the training datasets, while the validation datasets were only used to evaluate the influence of different CNN structures and parameters on the unseen datasets. Therefore, the performance of the CNN on the validation datasets is expected to be very close to its performance on test datasets. MR is the main modality for HCC detection, while CT is used for liver metastases detection in case that the [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET is absent. Because HCC has become the main indication for SIRT, CNN lesion segmentation on CT images is less essential than on MR images. Besides, the collection and annotation of new test datasets are time-consuming. Due to the above two reasons, we did not continue to evaluate the CNN's performance of lesion segmentation on test CT datasets. However, it would be valuable to apply CNN lesion segmentation to new clinical CT images. In this way, the CNN's performance on lesion segmentation can be assessed by the physicians during the clinical routine and valuable feedback can be collected from the doctor to help improve its performance.

The high-resolution U-net mentioned in Section 7.3.2.2 achieved comparable mean DSC of 0.69 for the challenge validation CT datasets to the highest mean DSC achieved during the LiTS17 challenge. The mean DSC for the high-resolution U-net was 0.62 for the SIRT validation CT datasets from our hospital, which is lower than the mean DSC for the challenge datasets. The clinical CT images from the hospital usually have lower contrast, higher noise level, and more variations of lesions than the challenge CT images. Therefore, CNN lesion segmentation on clinical CT images is more difficult than on challenge CT images. The clinical CT datasets only occupied 37% of the total CT datasets used for training the CNN, which might not be sufficient to obtain a CNN model with optimal performance on clinical images. When the high-resolution U-net is applied in clinical routine for lesion segmentation, these CNN lesion segmentations are corrected by the expert. With a sufficient number of corrected lesion segmentations accumulated, these segmentations and the segmented images can be used to train a CNN that is more specialized in the clinical datasets.

The high-resolution U-net using the masked input achieved a mean DSC of 0.48 for the test T1-weighted MR datasets. However, it achieved a mean DSC of 0.83 for the training T1-weighted MR datasets, which was equivalent to the mean DSC achieved by the model on the training CT datasets. This large discrepancy of the mean DSCs between the training and test MR datasets shows poor generalization of the CNN on MR datasets. The performance of CNN lesion segmentations for T1-weighted MR images requires to be further improved for clinical use. T1-weighted MR images are mainly used to detect HCCs, which can be heterogeneous and diffuse. This increases the variations of lesions in T1-weighted MR images. The current training MR datasets are only from 90 patients, which might not be able to provide sufficient anatomical variations from different patients. T1-weighted MR images from more patients are needed to facilitate the CNN to learn more general features that apply to various lesions. This can prevent the CNN from overfitting on the training datasets that only contain a limited portion of lesion variations. An improvement of the CNN performance on lesion segmentation for T1-weighted MR images is expected when the CNN is trained with more representative MR datasets from a larger group of patients.

Except for collecting more representative MR datasets, another possible solution is to use a deep learning model with better capability of generalization. Bayesian neural networks combine Bayesian inference with the structure of neural networks. Deterministic weights are learned during training for traditional neural networks. For Bayesian neural networks, weights are sampled from probabilistic distributions and the parameters of probabilistic distributions are learned during training. Bayesian neural networks are claimed to be able to avoid over-fitting when data are limited, since weights contain uncertainty. Some recent studies applied Bayesian neural networks in the field of medical image segmentation. Hiasa *et al*. implemented a Bayesian neural network with a U-net structure for segmentation of muscles from musculoskeletal for CT images [138]. Liu et al. proposed a Bayesian neural network for segmentation of the peripheral zone and transition zone of the prostate for multi-parametric MR images [139]. Compared with traditional neural networks, Bayesian neural networks have more sophisticated structure with more weights needed for tuning. This results in a much longer training time for Bayesian neural networks. The performance and generalization capability of Bayesian neural networks compared with traditional neural networks remain to be further studied.

In the current experiments, only T1-weighted MR images are used to train a CNN for lesion segmentation. In clinical practice, multi-modality MR images from one imaging study, such as T1-weighted MR, T2-weighted MR, and diffusion-weighted MR, are used to jointly detect HCCs and the other types of lesions. Tissues full of water have high signal (intensity) in the T2-weighted MR, which can help distinguish cysts from HCCs. Tissues with high cellularity appear bright in the diffusion-weighted MR, so highly cellular tumors are evident in this modality. It would be helpful to concatenate MR sequences of different modalities from the same imaging study into one volume and train a CNN with these volumes from imaging studies of different patients to segment lesions or classify the types of lesions. Since the MR imaging protocol is less standardized than the CT imaging protocol, it is

challenging to select the correct sequence for each MR modality. And a certain type of sequence is not always present in each imaging study. Therefore, the selection of MR sequences for CNN training needs more expertise and collaboration from radiologists.

Molecular imaging can help simplify the cumbersome task of lesion segmentation. This was achieved for segmentation of liver metastases that are [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE avid. Segmentations of liver metastases were easily obtained by thresholding [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET images. However, there is no molecular imaging available for detection of HCC. Therefore, automatic segmentation of HCC for MR images is still necessary, which is very challenging. When molecular imaging is applied for HCC detection in the future, it will be very helpful to perform HCC segmentation on it.

### 9.5.4 Dosimetry estimation of the error caused by medical image analysis techniques

This dissertation mainly focuses on development of automatic methods for segmentation of the liver and lesions, and multi-modal image registration and evaluation of the errors of segmentation and registration for these methods. For multi-modal image registration, the dosimetry errors caused by different registration methods were evaluated on a relatively small dataset of 10 SIRT patients. The difficulties and prospects of performing evaluation of dosimetry errors caused by automatic segmentation and registration methods are discussed in the following paragraphs.

The generalized CNN model (see Section 4.3.2) for liver segmentation achieved good performance on CT and T1-weighted MR images for most patients. With liver segmentations, only the mono-compartment method (see Section 1.5.3.1) could be used to compute the mean absorbed dose within the liver. A CNN liver segmentation was expected to make a small difference to the mean absorbed dose, compared with a manual liver segmentation. To use more advanced dose calculation methods, including the multi-compartment method (see Section 1.5.3.2) and the voxel-level method (see Section 1.5.3.3), lesion segmentation would be required. Even with manual lesion segmentation, the influence of CNN liver segmentation error on dosimetry difference was not expected to be large, because the liver has a much larger volume than lesions and its boundary is usually not close to the region with a high activity concentration. A relatively large dosimetry error caused by a CNN liver segmentation can only be expected when it misses part of or the whole lesion with high activity concentration. The above case is uncommon for our CNN model. Therefore, a further study on dosimetry errors caused by CNN liver segmentations was not conducted in order to save time for other tasks.

Lesion segmentations from multiple observers or CNN models are expected to have a critical influence on dosimetry errors, especially for lesions with high activity concentration. However, the performance of lesion segmentation of the current CNN model (see Section 7.3.2.2) was suboptimal, especially for MR images. Its segmentation accuracy needs to be further improved, before its lesion segmentations can be compared with manual lesion segmentations for evaluation of dosimetry errors. To improve CNN lesion segmentation, experienced experts are needed to spend time on lesion delineation for obtaining more training data with good annotation. Besides, the inter-observer variability of manual lesion segmentation between different experts is expected to be considerable, which can have substantial influence on dosimetry errors. However, this study on the inter-observer variability needs at least two experts performing manual lesion segmentations, which will be very time-consuming. A study on dosimetry errors caused by automatic methods for lesion segmentation or manual lesion segmentations from different observers can be conducted in the future, which will require more input from the medical experts.

The dosimetry errors caused by our segmentation-guided registration methods (see Section 5.3.3) and the multi-modality DIR tool of MIM (see Section 6.3.2) were evaluated for 10 SIRT patients. Since selecting patients with suitable images is very difficult (see Section 5.3.1) and annotating images is

time-consuming, the number of data is restricted to 10 patients. This evaluation can be performed on a larger scale of data in the future, which will be statistically more convincing. Besides, lesions can have various appearances in different imaging modalities and their size can change over time due to growth or treatment. The large values of the error metrics (see Figure 5.6, Figure 5.7, and Figure 6.1) obtained in the registration evaluation can be caused by variability of lesion appearance in different imaging modalities and changes of tumor size, in addition to the imperfect image registration. It is very difficult to separate these factors from the registration performance, when evaluating dosimetry errors caused by registration errors.

# 10. Summary

Selective internal radiation therapy (SIRT) aims at treating surgically unresectable liver tumors, including primary heptatic tumors and metastases. In SIRT, microspheres loaded with radionuclides, like $^{90}$Y, are injected through a catheter into selected branches of the hepatic artery that supply the target tumors. These microspheres, deposited near the target tumors, emit ionizing radiation, which kill tumor cells. SIRT planning and verification highly depend on medical imaging. The VOIs of the liver and lesions are obtained from the diagnostic CT image, the diagnostic MR image, or the [$^{18}$F]FDG/[$^{68}$Ga]Ga-DOTA-TATE PET/CT. The estimated and measured activity distribution maps of $^{90}$Y are from the $^{99m}$Tc-MAA SPECT/CT and the $^{90}$Y PET/CT or PET/MR, respectively. The VOIs need to be mapped to the activity distribution maps for quantified treatment planning and verification through co-registration of multi-modality images. Manual VOI delineation is time-consuming and labour-intensive. Therefore, automatic VOI segmentation is required to apply accurate VOIs for precise and personalized SIRT planning and verification.

In this dissertation, (semi-)automatic methods for liver and lesion segmentation and co-registration of multi-modality images have been developed and evaluated to assist the nuclear medicine physician in precise SIRT planning and verification.

Convolutional neural network (CNN) was used for automatic liver and lesion segmentation. A generalized U-net model was developed, which achieved good liver segmentations for both CT and T1-weighted MR images. This model has been applied in the SIRT routine and produced clinically usable liver segmentations without much manual correction needed for clinical CT images with low contrast and low dose. A U-net model using the high-resolution input images masked by the liver segmentation was developed and achieved good performance on CT lesion segmentation. Its lesion segmentations for the CT datasets from the LiTS17 challenge are comparable to the segmentations from the best-performing CNN during the challenge. The high-resolution U-net using the masked input had poorer lesion segmentation performance on T1-weighted MR images than it achieved on CT images. Lesions in MR images have more intensity and texture variations than lesions in CT images, which increase the difficulty in MR lesion segmentation. Representative MR images from more patients are needed to provide more lesion variations for CNN training to improve the performance of CNN lesion segmentation for MR images. However, the high-resolution U-net achieved high sensitivity for lesion detection, which was clinically applicable.

A multi-modality registration method guided by CNN liver segmentations and manually delineated landmarks (mainly lesions) was developed to improve registration accuracy of the liver and landmarks from CT and MR images. This segmentation-guided registration method achieved comparable performance to the multi-modality deformable image registration (DIR) of MIM with regard to registration errors and dose estimation difference caused by registration errors. Both registration methods can be applied to the volume-level and voxel-level dosimetry for integration of multi-modality image information in SIRT.

In summary, the developed methods for liver and lesion segmentation and multi-modality image registration can be integrated into the SIRT routine, contributing to standardization and automation of the clinical workflows. These techniques facilitate personalized and precise SIRT planning and verification.

# 11. Bibliography


[1] M. Konijnenberg, K. Herrmann, C. Kobe and et al, "EANM position paper on article 56 of the Council Directive 2013/59/Euratom (basic safety standards) for nuclear medicine therapy," *Eur J Nucl Med Mol Imaging,* vol. 48, p. 67–72, 2021.

[2] K. Si-Tayeb, F. P. Lemaigre and S. A. Duncan, "Organogenesis and Development of the Liver," *Developmental Cell,* vol. 18, no. 2, pp. 175-189, 2010.

[3] S. R. Z. Abdel-Misih and M. Bloomston, "Liver Anatomy," *Surg Clin North Am,* vol. 90, p. 643–653, 2010.

[4] J. Cantlie, "On a new arrangement of the right and left lobes of the liver," *J Anat Physiol,* vol. 32, pp. 4-10, 1898.

[5] C. Couinaud, "Le foie: études anatomiques et chirurgicales," *Masson,* 1957.

[6] M. Naito, G. Hasegawa and K. Takahashi, "Development, differentiation, and maturation of Kupffer cells," *Microscopy research and technique,* vol. 39, pp. 350-364, 1997.

[7] A. Geerts, "History, heterogeneity, developmental biology, and functions of quiescent hepatic stellate cells," *Seminars in liver disease,* vol. 21, pp. 311-336, 2001.

[8] P. S. Tietz and N. F. LaRusso, "Cholangiocyte biology," *Current opinion in gastroenterology,* vol. 22, pp. 279-287, 2006.

[9] L. H. Blumgart and J. Belghiti, Surgery of the liver, biliary tract, and pancreas, Philadelphia: Saunders Elsevier, 2007.

[10] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians,* vol. 68, no. 6, pp. 394-424, 2018.

[11] World Cancer Report 2014, World Health Organization, 2014, p. Chapter 5.6.

[12] I. Ahmed and D. N. Lobo, "Malignant tumours of the liver," *Surgery (Oxford),* vol. 27, no. 1, pp. 30-37, 2009.

[13] J. S. Dooley, A. S. F. Lok, G. Garcia-Tsao and M. Pinzani, Sherlock's Diseases of the Liver and Biliary System, New York: John Wiley & Sons, 2018, p. 705–729.

[14] S. Emre and G. J. McKenna, "Liver tumors in children," *Pediatric transplantation,* vol. 8, pp. 632-638, 2004.

[15] B. Ariff, C. R. Lloyd, S. Khan, M. Shariff, A. V. Thillainayagam, D. S. Bansi, S. A. Khan, S. D. Taylor-Robinson and A. K. Lim, "Imaging of liver cancer," *World journal of gastroenterology,* vol. 15, p. 1289–1300, 2009.

[16] J. D. Brierley, H. Asamura, E. v. Eycken and B. Rous, TNM Atlas, 7th Edition, Hoboken: Wiley-Blackwell, 2021.

[17] J. Balogh, D. Victor, E. H. Asham, S. G. Burroughs, M. Boktour, A. Saharia, X. Li, R. M. Ghobrial and H. P. Monsour, "Hepatocellular carcinoma: a review," *Journal of hepatocellular carcinoma,* vol. 3, pp. 41-53, 2016.

[18] V. Mazzaferro, S. Bhoori, C. Sposito, M. Bongini, M. Langer, R. Miceli and L. Mariani, "Milan criteria in liver transplantation for hepatocellular carcinoma: an evidence-based analysis of 15 years of experience," *Liver Transplantation,* vol. 17, pp. 44-57, 2011.

[19] H. Rhim and H. K. Lim, "Radiofrequency Ablation of Hepatocellular Carcinoma: Pros and Cons," *Gut and Liver,* vol. 4, pp. 113-118, 2010.

[20] J. H. Kim, H. J. Won, Y. M. Shin, K.-A. Kim and P. N. Kim, "Radiofrequency ablation for the treatment of primary intrahepatic cholangiocarcinoma," *American journal of roentgenology,* vol. 196, pp. 205-209, 2011.

[21] R. Miraglia, G. Pietrosi, L. Maruzzelli, I. Petridis, S. Caruso, G. Marrone, G. Mamone, G. Vizzini, A. Luca and B. Gridelli, "Efficacy of transcatheter embolization/chemoembolization (TAE/TACE) for the treatment of single hepatocellular carcinoma," *World Journal of Gastroenterology,* vol. 13, p. 2952–2955, 2007.

[22] D. I. Tsilimigras, P. Brodt, P.-A. Clavien, R. J. Muschel, M. I. D'Angelica, I. Endo, R. W. Parks, M. Doyle, E. d. Santibañes and T. M. Pawlik, "Liver metastases," *Nat Rev Dis Primers,* vol. 7, p. 27, 2021.

[23] Y. Minami and M. Kudo, "Hepatic malignancies: Correlation between sonographic findings and pathological features," *World journal of radiology,* vol. 2, p. 249, 2010.

[24] D. V. Sahani, M. A. Bajwa, Y. Andrabi, S. Bajpai and J. C. Cusack, "Current status of imaging and emerging techniques to evaluate liver metastases from colorectal carcinoma," *Annals of surgery,* vol. 259, pp. 861-872, 2014.

[25] H. Bahri, L. Laurence, J. Edeline, H. Leghzali, A. Devillers, J.-L. Raoul, M. Cuggia, H. Mesbah, B. Clement, E. Boucher and E. Garin, "High prognostic value of 18F-FDG PET for metastatic gastroenteropancreatic neuroendocrine tumors: a long-term evaluation," *Journal of Nuclear Medicine,* vol. 55, pp. 1786-1790, 2014.

[26] D. Moris, D. I. Tsilimigras, J. Chakedis, E. W. Beal, E. Felekouras, S. Vernadakis, D. Schizas, J. J. Fung and T. M. Pawlik, "Liver transplantation for unresectable colorectal liver metastases: A systematic review," *Journal of surgical oncology,* vol. 116, pp. 288-297, 2017.

[27] L. Crocetti, T. d. Baére, P. L. Pereira and F. P. Tarantino, "CIRSE Standards of Practice on Thermal Ablation of Liver Tumours," *CardioVascular and Interventional Radiology,* vol. 43, p. 951–962, 2020.

[28] Mathematics and Physics of Emerging Biomedical Imaging, Washington (DC): National Academies Press (US), 1996.

[29] S. Surti, "Update on time-of-flight PET imaging," *Journal of nuclear medicine,* vol. 56, p. 98–105, 2015.

[30] S. S. Gambhir, "Molecular imaging of cancer with positron emission tomography," *Nature Reviews Cancer,* vol. 2, p. 683–693, 2002.

[31] J.-F. Chatal and C. A. Hoefnagel, "Radionuclide therapy," *The Lancet,* vol. 354, no. 9182, pp. 931-935, 1999.

[32] T. Das and S. Banerjee, “Theranostic Applications of Lutetium-177 in Radionuclide Therapy,” *Curr Radiopharm,* vol. 9, pp. 94-101, 2016.

[33] F. D. G. Liberal, J. M. O'Sullivan, S. J. McMahon and K. M. Prise, “Targeted Alpha Therapy: Current Clinical Applications,” *Cancer Biotherapy and Radiopharmaceuticals,* vol. 35, pp. 404-417, 2020.

[34] G. B. L. Sandri, G. M. Ettorre, V. Giannelli, M. Colasanti, R. Sciuto, G. Pizzi, R. Cianni, G. D’Offizi, M. Antonini, G. Vennarecci and P. Lucatelli, “Trans-arterial radio-embolization: a new chance for patients with hepatocellular cancer to access liver transplantation, a world review,” *Translational gastroenterology and hepatology,* vol. 2, p. 98, 2017.

[35] V. Vilgrain, M. Abdel-Rehim, A. Sibert, M. Ronot, R. Lebtahi, L. Castéra and G. Chatellier, “Radioembolisation with yttrium–90 microspheres versus sorafenib for treatment of advanced hepatocellular carcinoma (SARAH): study protocol for a randomised controlled trial,” *Trials,* vol. 15, p. 474, 2014.

[36] M. Gandhi, S. P. Choo, C. H. Thng, S. B. Tan, A. S. C. Low, P. C. Cheow, A. S. W. Goh, K. H. Tay, R. H. G. Lo, B. K. P. Goh, J. S. Wong, D. C. E. Ng, K. C. So and Asia-Pacific Hepatocellular Carcinoma Trials Group, “Single administration of Selective Internal Radiation Therapy versus continuous treatment with sorafeNIB in locally advanced hepatocellular carcinoma (SIRveNIB): study protocol for a phase iii randomized controlled trial,” *BMC cancer,* vol. 16, p. 856, 2016.

[37] G. A. v. Hazel, V. Heinemann, N. K. Sharma and et al, “SIRFLOX: Randomized Phase III Trial Comparing First-Line mFOLFOX6 (Plus or Minus Bevacizumab) Versus mFOLFOX6 (Plus or Minus Bevacizumab) Plus Selective Internal Radiation Therapy in Patients With Metastatic Colorectal Cancer,” *Journal of clinical oncology,* vol. 34, p. 1723–1731, 2016.

[38] J. Wolstenholme, F. Fusco, A. M. Gray, J. Moschandreas, P. S. Virdee, S. Love, G. V. Hazel, P. Gibbs, H. S. Wasan and R. A. Sharma, “Quality of life in the FOXFIRE, SIRFLOX and FOXFIRE-global randomised trials of selective internal radiotherapy for metastatic colorectal cancer,” *International Journal of Cancer,* vol. 147, no. 4, pp. 1078-1085, 2020.

[39] R. Salem, A. Gabr, A. Riaz, R. Mora, R. Ali, M. Abecassis, R. Hickey, L. Kulik, D. Ganger, S. Flamm, R. Atassi, B. Atassi, K. Sato, A. B. Benson, M. F. Mulcahy, N. Abouchaleh, A. A. Asadi, K. Desai, B. Thornburg, M. Vouche, A. Habib, J. Caicedo, F. H. Miller, V. Yaghmai, J. R. Kallini, S. Mouli and R. J. Lewandowski, “Institutional decision to adopt Y90 as primary treatment for hepatocellular carcinoma informed by a 1,000-patient 15-year experience,” *Hepatology,* vol. 68, no. 4, pp. 1429-1440, 2018.

[40] M. L. J. Smits, M. Elschot, D. Y. Sze, Y. H. Kao, J. F. W. Nijsen, A. H. Iagaru, H. W. A. M. d. Jong, M. A. A. J. v. d. Bosch and M. G. E. H. Lam, “Radioembolization Dosimetry: The Road Ahead,” *CardioVascular and Interventional Radiology,* vol. 38, p. 261–269, 2015.

[41] M. L. Smits, M. Elschot, M. A. v. d. Bosch, G. H. v. d. Maat, A. D. v. h. Schip, B. A. Zonnenberg, P. R. Seevinck, H. M. Verkooijen, C. J. Bakker, H. W. d. Jong, M. G. Lam and J. F. Nijsen, “In Vivo Dosimetry Based on SPECT and MR Imaging of 166Ho-Microspheres for Treatment of Liver Malignancies,” *Journal of Nuclear Medicine,* vol. 54, pp. 2093-2100, 2013.

[42] Y. N. Park, C. P. Yang, G. J. Fernandez, O. Cubukcu, S. N. Thung and N. D. Theise, “Neoangiogenesis and sinusoidal "capillarization" in dysplastic nodules of the liver,” *The American journal of surgical pathology,* vol. 22, p. 656–662, 1998.

[43] W. E. Bolch, L. G. Bouchet, J. S. Robertson and et al, "MIRD pamphlet No. 17: the dosimetry of nonuniform activity distributions--radionuclide S values at the voxel level. Medical Internal Radiation Dose Committee," *Journal of nuclear medicine,* vol. 40, p. 11–36, 1999.

[44] J. A. Siegel, S. R. Thomas, J. B. Stubbs and et al, "MIRD pamphlet no. 16: Techniques for quantitative radiopharmaceutical biodistribution data acquisition and analysis for use in human radiation dose estimates," *Journal of nuclear medicine,* vol. 40, p. 37–61, 1999.

[45] R. Bastiaannet, S. C. Kappadath, B. Kunnen, A. J. A. T. Braat, M. G. E. H. Lam and H. W. A. M. d. Jong, "The physics of radioembolization," *EJNMMI physics,* vol. 5, p. 22, 2018.

[46] "Biocompatibles UK Ltd. Package Insert–TheraSphere® Yttrium-90 Glass Microspheres–Rev," 2014.

[47] H. Ilhan, A. Goritschan, P. Paprottka and et al, "Predictive Value of 99mTc-MAA SPECT for 90Y-Labeled Resin Microsphere Distribution in Radioembolization of Primary and Secondary Hepatic Tumors," *Journal of nuclear medicine,* vol. 56, p. 1654–1660, 2015.

[48] J. K. Mikell, A. Mahvash, W. Siman, V. Baladandayuthapani, F. Mourtada and S. C. Kappadath, "Selective Internal Radiation Therapy With Yttrium-90 Glass Microspheres: Biases and Uncertainties in Absorbed Dose Calculations Between Clinical Dosimetry Models," *International journal of radiation oncology, biology, physics,* vol. 96, p. 888–896, 2016.

[49] J. O. Doherty, "A review of 3D image-based dosimetry, technical considerations and emerging perspectives in 90Y microsphere therapy," *Journal of diagnostic imaging in therapy,* vol. 2, pp. 1-34, 2015.

[50] W. A. Dezarn, J. T. Cessna, L. A. DeWerd and et al, "Recommendations of the American Association of Physicists in Medicine on dosimetry, imaging, and quality assurance procedures for 90Y microsphere brachytherapy in the treatment of hepatic malignancies," *Medical physics,* vol. 38, pp. 4824-4845, 2011.

[51] W. C. Scarfe and A. G. Farman, "What is Cone-Beam CT and How Does it Work?," *Dental Clinics of North America,* vol. 52, no. 4, pp. 707-730, 2008.

[52] L. Rusko, G. Bekes, G. Nemeth and M. Fidrich, "Fully automatic liver segmentation for contrast-enhanced CT images," *MICCAI Wshp. 3D Segmentation in the Clinic: A Grand Challenge,* vol. 2, 2007.

[53] R. Susomboon, D. S. Raicu and J. Furst, "A hybrid approach for liver segmentation," *Proceedings of MICCAI workshop on 3D segmentation in the clinic: a grand challenge,* pp. 151-160, 2007.

[54] R. Pohle and K. D. Toennies, "Segmentation of medical images using adaptive region growing," *Medical Imaging 2001: Image Processing,* vol. 4322, pp. 1337-1346, 2001.

[55] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding,* vol. 61, pp. 38-59, 1995.

[56] T. Heimann, I. Wolf and H. P. Meinzer, "Active shape models for a fully automated 3D segmentation of the liver--an evaluation on clinical data," *Med Image Comput Comput Assist Interv,* vol. 9, pp. 41-48, 2006.

[57] D. Kainmüller, T. Lange and H. Lamecker, "Shape constrained automatic segmentation of the liver based on a heuristic intensity model," in *Proc. MICCAI Workshop 3D Segmentation in the Clinic: A Grand Challenge*, 2007.

[58] M. Erdt, S. Steger, M. Kirschner and S. Wesarg, "Fast automatic liver segmentation combining learned shape priors with observed shape deviation," in *Proceedings of the 23rd international symposium on computer-based medical systems*, 2010.

[59] M. Moghbel, S. Mashohor, R. Mahmud and M. Saripan, "Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography," *Artificial Intelligence Review,* vol. 50, pp. 497-537, 2018.

[60] X. Zhou, T. Kitagawa, T. Hara and et al, "Constructing a probabilistic model for automated liver region segmentation using non-contrast X-ray torso CT images," *Med Image Comput Comput Assist Interv,* vol. 9, pp. 856-863, 2006.

[61] P. Slagmolen, A. Elen, D. Seghers, D. Loeckx, F. Maes and K. Haustermans, "Atlas based liver segmentation using nonrigid registration with a B-spline transformation model," in *Proceedings of MICCAI workshop on 3D segmentation in the clinic: a grand challenge*, 2007.

[62] E. van Rikxoort, Y. Arzhaeva and B. Ginneken, "Automatic segmentation of the liver in computed tomograpy scans with voxel classification and atlas matching," in *Proceedings of the MICCAI Workshop*, 2007.

[63] M. G. Linguraru, Z. Li, F. Shah and R. M. Summers, "Automated liver segmentation using a normalized," in *SPIE medical imaging*, 2009.

[64] C. Li, X. Wang, S. Eberl, M. Fulham, Y. Yin and D. Feng, "Fully automated liver segmentation for low-and high-contrast CT volumes based on probabilistic atlases," in *2010 IEEE International Conference on Image Processing*, 2010.

[65] X. Jiang, R. Zhang and S. Nie, "Image Segmentation Based on Level Set Method," *Physics Procedia,* vol. 33, pp. 840-845, 2012.

[66] S. Pan and B. M. Dawant, "Automatic 3D segmentation of the liver from abdominal CT images: a level-set approach," *Int Soc Opt Photonics Med Imaging,* vol. 4322, pp. 128-138, 2001.

[67] J. Wang, Y. Cheng, C. Guo, Y. Wang and S. Tamura, "Shape-intensity prior level set combining probabilistic atlas and probability map constrains for automatic liver segmentation from abdominal CT images," *Int J Comput Assist Radiol Surg,* vol. 11, pp. 817-826, 2016.

[68] Y. Boykov, O. Veksler and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 23, pp. 1222-1239, 2001.

[69] R. Beichel, C. Bauer, A. Bornik, E. Sorantin and H. Bischof, "Liver segmentation in CT data: A segmentation refinement approach," in *Proceedings of" 3D Segmentation in The Clinic: A Grand Challenge*, 2007.

[70] K. Wang, A. Mamidipalli, T. Retson and et al, "Automated CT and MRI liver segmentation and biometry using a generalized convolutional neural network," *Radiology: Artificial Intelligence,* vol. 1, p. 180022, 2019.

[71] K. Sharma, C. Rupprecht, A. Caroli and et al, "Automatic Segmentation of Kidneys using Deep Learning for Total Kidney Volume Quantification in Autosomal Dominant Polycystic Kidney Disease," *Scientific reports,* vol. 7, pp. 1-10, 2017.

[72] P. Bilic, P. F. Christ, E. Vorontsov and et al, “The Liver Tumor Segmentation Benchmark (LiTS),” 2019.

[73] O. Ronneberger, P. Fischer and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[74] G. Chlebus, H. Meine, S. Thoduka and et al, “Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections,” *PLOS ONE,* vol. 14, p. e0217228, 2019.

[75] S. Park, K. Seo and J. Park, “Automatic Hepatic Tumor Segmentation Using Statistical Optimal Threshold,” in *International Conference on Computational Science*, 2005.

[76] H. A. Nugroho, D. Ihtatho and H. Nugroho, “Contrast enhancement for liver tumor identification,” in *MICCAI workshop,* 2008.

[77] A. Choudhary, N. Moretto, F. P. Ferrarese and G. A. Zamboni, “An entropy based multi-thresholding method for semi-automatic segmentation of liver tumors,” in *MICCAI workshop,* 2008.

[78] J. Moltz, L. Bornemann, V. Dicken and H.-O. Peitgen, “Segmentation of Liver Metastases in CT Scans by Adaptive Thresholding and Morphological Processing,” in *MICCAI workshop,* 2008.

[79] N. H. Abdel-massieh, M. M. Hadhoud and K. M. Amin, “Fully automatic liver tumor segmentation from abdominal CT scans,” in *International Conference on Computer Engineering & Systems*, 2010.

[80] D. Wong, J. Liu, Y. Fengshou and et al, “A semi-automated method for liver tumor segmentation based on 2D region growing with knowledge-based constraints,” in *MICCAI workshop,* 2008.

[81] Y. Qi, W. Xiong, W. K. Leow and et al, “Semi-automatic Segmentation of Liver Tumors from CT Scans Using Bayesian Rule-based 3D Region Growing,” *The MIDAS Journal,* 2008.

[82] D. Smeets, D. Loeckx, B. Stijnen, B. D. Dobbelaer, D. Vandermeulen and P. Suetens, “Semi-automatic level set segmentation of liver tumors combining a spiral-scanning technique with supervised fuzzy pixel classification,” *Med Image Anal,* vol. 14, pp. 13-20, 2010.

[83] D. Jimenez-Carretero, L. Fernandez-de-Manuel, J. Pascau and et al, “Optimal multiresolution 3D level-set method for liver segmentation incorporating local curvature constraints,” in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* 2011.

[84] A. Choudhary, N. Moretto, F. P. Ferrarese and G. A. Zamboni, “An entropy based multi-thresholding method for semi-automatic segmentation of liver tumors,” in *MICCAI workshop,* 2008.

[85] L. Bi, J. Kim, A. Kumar and D. Feng, “Automatic liver lesion detection using cascaded deep residual networks,” 2017.

[86] X.-F. Xi, L. Wang, V. S. Sheng, Z. Cui, B. Fu and F. Hu, “Cascade u-resnets for simultaneous liver and lesion segmentation,” *IEEE Access,* vol. 8, pp. 68944-68952, 2020.

[87] G. Chlebus, A. Schenk, J. H. Moltz, B. v. Ginneken, H. K. Hahn and H. Meine, "Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing," *Scientific reports,* vol. 8, pp. 1-7, 2018.

[88] M. J. Jansen, H. J. Kuijf, M. Niekel and et al, "Liver segmentation and metastases detection in MR images using convolutional neural networks," *Journal of Medical Imaging,* vol. 6, p. 044003, 2019.

[89] A. Hatamizadeh, A. Hoogi, D. Sengupta and et al, "Deep active lesion segmentation," in *International Workshop on Machine Learning in Medical Imaging*, 2019.

[90] W.-C. C. Lee, M. E. Tublin and B. E. Chapman, "Registration of MR and CT images of the liver: comparison of voxel similarity and surface based registration algorithms," *Computer Methods and Programs in Biomedicine,* vol. 78, pp. 101-114, 2005.

[91] A. Foruzan and H. Motlagh, "Multimodality liver registration of Open-MR and CT scans," *Int J CARS,* vol. 10, p. 1253–1267, 2015.

[92] G. Cazoulat, D. Elganainy, B. M. Anderson and et al, "Vasculature-Driven Biomechanical Deformable Image Registration of Longitudinal Liver Cholangiocarcinoma Computed Tomographic Scans," *Advances in Radiation Oncology,* vol. 5, pp. 269-278, 2020.

[93] A. Charnoz, V. Agnus, G. Malandain, C. Forest, M. Tajine and L. Soler, "Liver Registration for the Follow-Up of Hepatic Tumors," in *Medical Image Computing and Computer-Assisted Intervention,* 2005.

[94] G. Gunay, M. H. Luu, A. Moelker, T. v. Walsum and S. Klein, "Semiautomated registration of pre- and intraoperative CT for image-guided percutaneous liver tumor ablation interventions," *Medical Physics,* vol. 44, pp. 3718-3725, 2017.

[95] X. Huang, B. Wang, R. Liu, X. Wang and Z. Wu, "CT-MR image registration in liver treatment by maximization of mutual information," in *IEEE International Symposium on IT in Medicine and Education,* 2008.

[96] K. D. Moor, J. Nuyts, L. Plessers, S. Stroobants, F. Maes and P. Dupont, "Non-rigid registration with position dependent rigidity for whole body PET follow-up studies," in *IEEE Nuclear Science Symposium Conference Record,* 2006.

[97] N. Spahr, S. Thoduka, N. Abolmaali and et al, "Multimodal image registration for liver radioembolization planning and patient assessment," *Int J CARS,* vol. 14, p. 215–225, 2019.

[98] C. Rieder, S. Wirtz, J. Strehlow and et al, "Automatic alignment of pre- and post-interventional liver CT images for assessment of radiofrequency ablation," in *Proc. SPIE 8316,* 2012.

[99] K. A. Hasenstab, G. M. Cunha, A. Higaki and et al, "Fully automated convolutional neural network-based affine algorithm improves liver registration and lesion co-localization on hepatobiliary phase T1-weighted MR images," *Eur Radiol Exp,* vol. 3, p. 43, 2019.

[100] C. Breedis and G. Young, "The blood supply of neoplasms in the liver," *The American journal of pathology,* vol. 30, pp. 969-977, 1954.

[101] B. N. Gray, M. A. Burton, D. Kelleher, P. Klemp and L. Matz, "Tolerance of the liver to the effects of Yttrium-90 radiation," *Int J Radiat Oncol Biol Phys,* vol. 18, pp. 619-623, 1990.

[102] M. Cremonesi, C. Chiesa, L. Strigari and et al, “Radioembolization of hepatic lesions from a radiobiology and dosimetric perspective,” *Front Oncol,* vol. 4, p. 210, 2014.

[103] A. S. Kennedy, C. Nutting, D. Coldwell, J. Gaiser and C. Drachenberg, “Pathologic response and microdosimetry of (90)Y microspheres in man: review of four explanted whole livers,” *International journal of radiation oncology,* vol. 60, pp. 1552-1563, 2004.

[104] R. De Gersem, G. Maleux, H. Vanbilloen and et al, “Influence of time delay on the estimated lung shunt fraction on 99mTc-labeled MAA scintigraphy for 90Y microsphere treatment planning,” *Clinical nuclear medicine,* vol. 38, pp. 940-942, 2013.

[105] N. M. Maughan, M. Eldib, D. Faul and et al, “Multi institutional quantitative phantom study of yttrium-90 PET in PET/MRI: the MR-QUEST study,” *EJNMMI Phys,* vol. 5, p. 7, 2018.

[106] C. L. Wright, K. Binzel, J. Zhang, E. J. Wuthrick and M. V. Knopp, “Clinical feasibility of 90Y digital PET/CT for imaging microsphere biodistribution following radioembolization,” *European journal of nuclear medicine and molecular imaging,* vol. 44, pp. 1194-1197, 2017.

[107] E. Garin, L. Lenoir, Y. Rolland and et al, “Dosimetry based on 99mTc-macroaggregated albumin SPECT/CT accurately predicts tumor response and survival in hepatocellular carcinoma patients treated with 90Y-loaded glass microspheres: preliminary results,” *Journal of nuclear medicine,* vol. 53, p. 255–263, 2012.

[108] S. Ho, W. Y. Lau, T. W. Leung, M. Chan, Y. K. Ngar, P. J. Johnson and A. K. Li, “Partition model for estimating radiation doses from yttrium-90 microspheres in treating hepatic tumours,” *European journal of nuclear medicine,* vol. 23, p. 947–952, 1996.

[109] E. J. Rangraz, W. Coudyzer, G. Maleux, K. Baete, C. M. Deroose and J. Nuyts, “Multi-modal image analysis for semi-automatic segmentation of the total liver and liver arterial perfusion territories for radioembolization,” *EJNMMI research,* vol. 9, p. 19, 2019.

[110] K. Kamnitsasa, C. Ledig, V. F. Newcombe and et al, “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” *Medical image analysis,* vol. 36, pp. 61-78, 2017.

[111] J. Long, E. Shelhamer and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[112] K. H. Zou, S. K. Warfield and A. Bharatha, “Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports,” *Academic radiology,* vol. 11, pp. 178-189, 2004.

[113] T. Heimann, B. v. Ginneken, M. A. Styner and et al, “Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets,” *IEEE Transactions on Medical Imaging,* vol. 28, pp. 1251-1265, 2009.

[114] X. Tang, E. Jafargholi Rangraz, R. Heeren and et al, “Segmentation-guided multi-modal registration of liver images for dose estimation in SIRT,” *EJNMMI Physics,* vol. 9, p. 3, 2022.

[115] G. Chartrand, T. Cresson, R. Chav, A. Gotra, A. Tang and J. A. De Guise, “Liver Segmentation on CT and MR Using Laplacian Mesh Optimization,” *IEEE transactions on bio-medical engineering,* vol. 64, p. 2110–2121, 2017.

[116] S. Mulay, G. Deepika, S. Jeevakala, K. Ram and M. Sivaprakasam, “Liver segmentation from multimodal images using HED-mask R-CNN,” in *International Workshop on Multiscale Multimodal Medical Imaging*, 2019.

[117] P. F. Christ, F. Ettlinger, F. Grün and et al, “Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks,” arXiv preprint, 2017.

[118] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint,* 2014.

[119] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality (complete samples),” *Biometrika,* vol. 52, pp. 591-611, 1965.

[120] A. Imam, M. Usman and M. A. Chiawa, “On Consistency and Limitation of paired t-test, Sign and Wilcoxon Sign Rank Test,” *IOSR Journal of Mathematics,* vol. 10, pp. 1-6, 2014.

[121] S. A. Padia, R. J. Lewandowski, G. E. Johnson and et al, “Radioembolization of Hepatic Malignancies: Background, Quality Improvement Guidelines, and Future Directions,” *Journal of Vascular and Interventional Radiology,* vol. 28, no. 1, pp. 1-15, 2017.

[122] R. Lhommel, L. van Elmbt, P. Goffette and et al, “Feasibility of 90Y TOF PET-based dosimetry in liver metastasis therapy using SIR-Spheres,” *Eur J Nucl Med Mol Imaging,* vol. 37, p. 1654–1662, 2010.

[123] H. M. Luu, C. Klink, W. Niessen, A. Moelker and T. van Walsum, “Non-Rigid Registration of Liver CT Images for CT-Guided Ablation of Liver Tumors,” *PLoS ONE,* vol. 11, p. e0161600, 2016.

[124] A. A. Alsultan, C. van Roekel and M. W. Barentsz, “The Efficacy of Coil Embolization to Obtain Intrahepatic Redistribution in Radioembolization: Qualitative and Quantitative Analyses,” *Cardiovasc Intervent Radiol,* vol. 43, p. 391–401, 2020.

[125] G. Nodari, R. Popoff, J. M. Riedinger and et al, “Impact of contouring methods on pre-treatment and post-treatment dosimetry for the prediction of tumor control and survival in HCC patients treated with selective internal radiation therapy,” *EJNMMI Res,* vol. 11, p. 24, 2021.

[126] E. Jafargholi Rangraz, X. Tang, C. Van Laeken and et al, “Quantitative comparison of pre-treatment predictive and post-treatment measured dosimetry for selective internal radiation therapy using cone-beam CT for tumor and liver perfusion territory definition,” *EJNMMI Res,* vol. 10, p. 94, 2020.

[127] X. Tang, E. Jafargholi Rangraz, W. Coudyzer and et al, “Whole liver segmentation based on deep learning and manual adjustment for clinical use in SIRT,” *Eur J Nucl Med Mol Imaging,* vol. 47, p. 2742–2752, 2020.

[128] H. Levillain, O. Bagni, C. M. Deroose and et al, “International recommendations for personalised selective internal radiation therapy of primary and metastatic liver diseases with yttrium-90 resin microspheres,” *Eur J Nucl Med Mol Imaging,* vol. 48, p. 1570–1584, 2021.

[129] S. Calusi, G. Labanca, M. Zani and et al, “A multiparametric method to assess the MIM deformable image registration algorithm,” *Journal of applied clinical medical physics,* vol. 20, pp. 75-82, 2019.

[130] N. Fukumitsu, K. Nitta, T. Terunuma and et al, “Registration error of the liver CT using deformable image registration of MIM Maestro and Velocity AI,” *BMC Med Imaging,* vol. 17, p. 30, 2017.

[131] X. Gu, B. Dong, J. Wang and et al, “A contour-guided deformable image registration algorithm for adaptive radiotherapy,” *Physics in Medicine & Biology,* vol. 58, no. 6, p. 1889, 2013.

[132] V. Kearney, S. Chen, X. Gu and et al, “Automated landmark-guided deformable image registration,” *Physics in Medicine and Biology,* vol. 60, p. 101, 2014.

[133] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, 2016.

[134] W. H. Kruskal and W. W. Wallis, “Use of Ranks in One-Criterion Variance Analysis,” *Journal of the American Statistical Association,* vol. 47, no. 260, pp. 583-621, 1952.

[135] X. Cheng, L. Zhang and Y. Zheng, “Deep similarity learning for multimodal medical images,” *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization,* vol. 6, pp. 248-252, 2018.

[136] G. Haskins, J. Kruecker, U. Kruger and et al, “Learning deep similarity metric for 3D MR–TRUS image registration,” *International journal of computer assisted radiology and surgery,* vol. 14, pp. 417-425, 2019.

[137] S. S. M. Salehi, S. Khan, D. Erdogmus and A. Gholipour, “Real-Time Deep Pose Estimation With Geodesic Loss for Image-to-Template Rigid Registration,” *IEEE Transactions on Medical Imaging,* vol. 38, pp. 470-481, 2019.

[138] Y. Hiasa, Y. Otake, M. Takao, T. Ogawa, N. Sugano and Y. Sato, “Automated Muscle Segmentation from Clinical CT Using Bayesian U-Net for Personalized Musculoskeletal Modeling,” *IEEE Transactions on Medical Imaging,* vol. 39, pp. 1030-1040, 2020.

[139] Y. Liu, G. Yang, M. Hosseiny and et al, “Exploring Uncertainty Measures in Bayesian Deep Attentive Neural Networks for Prostate Zonal Segmentation,” *IEEE Access,* vol. 8, pp. 151817-151828, 2020.

[140] C. Tanner, F. Ozdemir, R. Profanter, V. Vishnevsky, E. Konukoglu and O. Goksel, “Generative adversarial networks for MR-CT deformable image registration,” *arXiv preprint,* 2018.

# 12. Acknowledgements, personal contribution and conflict of interest statements

## 12.1 Acknowledgements

I would like to thank my supervisor (Johan Nuyts), co-supervisors (Christophe Deroose and Kristof Baete), and Mark Gooding for conceptualizing the project, contributing their valuable ideas, mentoring me as a young researcher, and providing scientific / clinical support.

We thank the members of the examining committee (Geert Maleux, Patrick Dupont, Arthur J.A.T. Braat, and Mark Konijnenberg) and the supervisory committee (Steven Dymarkowski, Lieven Dupont, Inge Depoortere, Tatiana Kouznetsova) for their guidance and support on the research process.

I would like to thank Georg Schramm and Ahmadreza Rezaei for providing valuable ideas about developing and improving the algorithms. I would like to acknowledge Esmaeel Jafargholi Rangraz for the dosimetry analysis tools and support on liver segmentation and image registration. Jeroen Bertels and David Robben provided the framework in python for building our CNN models for liver and lesion segmentation, which is much appreciated.

I would like to thank Walter Coudyzer and Richard ‘s Heeren for providing manual liver and lesion segmentation. We wish to thank Vincent Vandecaveye and Frederik De Keyzer for providing important knowledge about MR imaging and selecting MR images for CNN lesion segmentation. I would also like to acknowledge Wies Deckers for designing clinical workflows and providing instructions on MIM.

## 12.2 Funding

This project was funded by the European Horizon2020 ITN project (HYBRID, MSCA 764458) and the Research Foundation Flanders (FWO, grant G082418N).

## 12.3 Personal contribution

My personal contributions to this dissertation include (but are not limited to): algorithm development, experiment design, data acquisition, data delineation, data analysis, clinical software development, and thesis writing.

The authors’ contributions to the published chapters of thesis are listed below.

**Chapter 3**. We declare that all authors have contributed to the performed research and the manuscript. Conception and design: Xikai Tang, Esmaeel Jafargholi Rangraz, Jeroen Bertels, David Robben, Georg Schramm, Mark J. Gooding, Kristof Baete, Christophe M. Deroose, Johan Nuyts; Data acquisition: Xikai Tang, Esmaeel Jafargholi Rangraz, Geert Maleux, Chris Verslype, Walter Coudyzer, Christophe M. Deroose; Data delineation: Xikai Tang, Walter Coudyzer, Christophe M. Deroose; Data analysis: Xikai Tang, Georg Schramm, Mark J. Gooding, Christophe M. Deroose, Johan Nuyts; Writing manuscript and contributing to revising and approving the manuscript: all.

**Chapter 5**. We declare that all authors have contributed to the performed research and the manuscript. Conception and design: Xikai Tang, Esmaeel Jafargholi Rangraz, Mark J. Gooding, Kristof Baete, Christophe M. Deroose, Johan Nuyts; Data acquisition: Xikai Tang, Esmaeel Jafargholi Rangraz, Geert Maleux, Chris Verslype, Walter Coudyzer, Christophe M. Deroose; Data delineation: Xikai Tang, Richard ‘s Heeren, Walter Coudyzer, Christophe M. Deroose; Data analysis: Xikai Tang, Mark J.

Gooding, Christophe M. Deroose, Johan Nuyts; Writing manuscript and contributing to revising and approving the manuscript: all.

## 12.4 Conflict of interest statement

Esmaeel Jafargholi Rangraz is employed by Quirem Medical B.V., Deventer, Netherlands. Christophe M. Deroose is a Senior Clinical Investigator at the Research Foundation Flanders (FWO). Mark Gooding is employed by Mirada Medical Ltd, Oxford, UK, a medical software company. The department of nuclear medicine at KU Leuven receives support from GE for image reconstruction research. No other potential conflicts of interest relevant to this dissertation exist.

# Xikai Tang

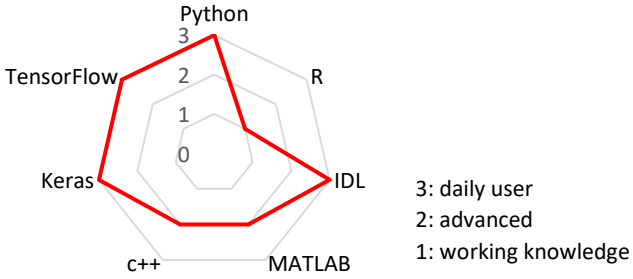
**Medical imaging scientist**

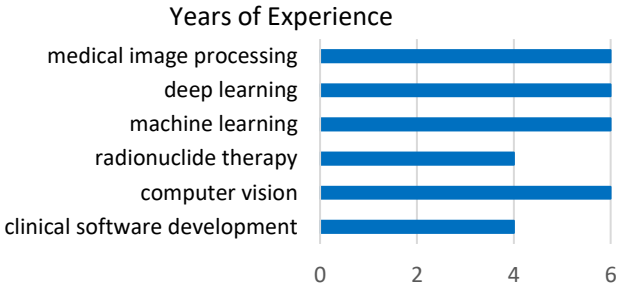
Phone: +32 483 603770 email: xikaitang@outlook.com

Address: 3001 Leuven, Belgium https://www.linkedin.com/in/xikai-tang/


## Profile

- Professional medical imaging scientist with 6 years of experience in medical image processing, machine learning, deep learning.
- Proven track record of developing high-quality medical image analysis algorithms used in the hospital.
- Successful design and deployment of clinical AI platform applied to patients.
- Successful standardization and automation of the clinical workflow of selective internal radionuclide therapy using AI.




## Work experience

Aug 2018 – current

### Research associate, medical image analysis, KU Leuven

Leuven, Belgium

AI-based medical image analysis platform development for selective internal radionuclide therapy for liver cancer

**Programming**: python, c++, TensorFlow, Keras, IDL, Linux

- AI system for automatic liver and tumor contouring for CT and MR, applied to over 100 patients.
- AI-guided image registration algorithm, comparable performance with commercial tools.
- Assembled AI platform integrated into the commercial software, used in the clinical routines of the hospital for 3 years.
- Successful automation and standardization of the clinical workflow using the AI platform.
- Substantial reduction of the clinical manual workload, like manual liver delineation.
- Substantial performance boosting of AI models for real-world clinical images.
- Collection, annotation, and statistical analysis of clinical images and dosimetry analysis.

Sep 2017 – Jun 2018

### Research engineer, medical image analysis, Philips

Best, Netherlands

Machine learning system for segmentation and classification of materials for CBCT with dual-layer detector

**Programming**: MATLAB

- Edge extraction for CBCTs, using edge detector, interest point detection, and unsupervised learning (clustering).
- Material separation using edges and dual-energy information of CBCTs.

May – Jul 2017

### Student researcher, machine learning, TU/e

Eindhoven, Netherlands

Improving 3D Ultrasound Needle Localization in noisy detection using Conditional Random Fields (CRF)

- 3D CRF programming in c++ based on 2D CRF.
- Needle localization improvement from noisy detection in 3D ultrasound volumes using 3D CRF.

## Education and relevant courses

2018 – 2022

### Ph.D., biomedical sciences, KU Leuven

- Funded by EU's H2020 programme

2016 – 2018

### MSc, Electrical engineering, Eindhoven University of Technology

- Scholarships from the university and the Dutch government
- GPA: 8.52/10

2013 – 2017

### BSc, Electrical engineering, Zhejiang University

- GPA: 89/100

## Professional skills

- Medical image analysis.
- Machine learning and deep learning.
- Image registration.
- Clinical workflow development and deployment.
- Clinical data analysis and dosimetry analysis.
- Computer vision.
- Python, c++, TensorFlow, Keras, Matlab, Linux.

## Selected publications

- Whole liver segmentation based on deep learning and manual adjustment for clinical use in SIRT. Eur J Nucl Med Mol Imaging 47, 2742–2752 (2020).
- Segmentation-guided multi-modal registration of liver images for dose estimation in SIRT. EJNMMI Physics 9, 3 (2022).
- Quantitative comparison of pre-treatment predictive and post-treatment measured dosimetry for selective internal radiation therapy using cone-beam CT for tumor and liver perfusion territory definition. EJNMMI Research 10, 94 (2020).

## Conferences

- Preliminary results of whole liver segmentation with deep learning for SIRT. 32nd Annual Congress of the European-Association-of-Nuclear-Medicine (EANM). Barcelona. 12-16 October 2019.
- Liver registration for CT images guided by CNN liver segmentation. IEEE International Symposium on Biomedical Imaging. Nice. 13-16 April 2021.
- Segmentation-guided liver registration for CT and MR images. ESMRMB 2021. Barcelona. 7-9 October 2021.

## Language

- Chinese: native
- English: fluent
- Dutch: A2 level

## Hobbies

Badminton lover, enthusiastic beginner for tennis, fitness addicted, traveling with friends, cooking lover