SCANNERLESS MRI GENERATION USING GENERATIVE ADVERSARIAL NETWORKS WITH MULTIPLE SURROGATE SIGNALS

K.J.W. (Koen) Damme

MSC ASSIGNMENT

Committee:
dr. N. Strisciuglio
A. Cordon, MSc
dr. ir. M. Abayazid
dr. ing. G. Englebienne

October, 2024

070RaM2024
Robotics and Mechatronics
EEMCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Scannerless MRI Generation Using Generative Adversarial Networks With Multiple Surrogate Signals

Koen Jacobus Wilhelmus Damme (s2832690)
Robotics and Mechatronics Lab
University of Twente
Enschede, The Netherlands
k.j.w.damme@student.utwente.nl

***Abstract*—Respiratory-induced motion (RIM) presents a challenge in targeting liver tumors during medical procedures, as it causes the tumor to shift position within the body. Motion models can track the position of a liver tumor based on a surrogate signal, compensating for RIM to enable more accurate ablation and biopsy procedures. However, interpreting tumor position as an XYZ-coordinate would be challenging for clinicians. This study presents a conditional progressively growing generative adversarial network (cProGAN) that can generate scannerless MR-images using one or multiple surrogate signals for guidance during liver interventions. We compared three signals: a heat camera measuring the airflow, and an ultrasound transducer and external markers, to capture the internal and external abdominal motion, respectively. This study is validated in seven human subject experiments, where MR-images and the three surrogate signals are simultaneously collected while each subject is following a specific breathing protocol. The quality of the scannerless images is assessed by the structural similarity index measure (SSIM) and by extracting the superior-inferior movement of the liver border in the real and scannerless images and comparing the resulting waveform using the mean absolute error (MAE, in millimeters and as a percentage of the average liver movement) and the coefficient of determination metrics. The model trained on external markers generated images with the most accurate liver positions during breathing (MAE of 5.02 ± 3.74 mm) and breath holds (MAE of 9.14 ± 1.31 mm). The highest SSIM was for the combined model during breathing (51.42%) and for the external marker model during breath holds (36.47%). Models using the other surrogate signals resulted in a significantly higher MAE and lower SSIM. These results suggest that external marker tracking provides the most accurate respiratory motion modeling for scannerless MRI generation, though further research is needed to improve image quality. The proposed solution can potentially be expanded by adding more sources of motion and generating entire 3D volumes and we believe that this could greatly improve the precision of percutaneous procedures in the liver and make them easier to perform.**

***Index Terms*—Respiratory-induced motion, surrogate signals, generative adversarial networks**

I. Introduction

Liver cancer was the cause of death for 830,000 people in 2020 and it was among the top three causes of cancer deaths in 46 countries around the world [1]. Rumgay et al. [1] expect that the number of deaths and people diagnosed will increase by more than 55% between 2020 and 2040. Primary liver cancer, or hepatocellular carcinoma (HCC) is one of the most difficult type of cancers to treat and has a high recurrence chance of 70% in five years [2].

Biopsies and ablation are percutaneous procedures that are commonly performed in the liver. For tumors with atypical imaging characteristics, a biopsy should be done for a definitive diagnosis [3]. Obtaining a reliable piece of tissue is essential to find the correct diagnosis and prevent a false-negative. Gonzales et al. [3] found that the false-negative rate for tumors smaller than 2 cm was 30%. Ablation is the most common treatment for early stage HCC (single tumor $\leq$ 5 cm or up to three tumors $\leq$ 3 cm) [4]. During ablation treatment, focal tumors are destroyed by applying chemicals (such as ethanol and acetic acid) or thermal/non-thermal energy, delivered via needle-like applicators [5].

A. Image-guided Interventions

During these interventions, medical imaging modalities like computed tomography (CT), ultrasound (US), magnetic resonance imaging (MRI), and others are commonly applied [6]. Unfortunately, neither have both a high contrast and a high temporal resolution, which are both important for real-time guidance. For example, MRI, which is the preferred imaging modality for liver interventions [6], offers a high contrast but with a low temporal resolution. The low temporal resolution is problematic because respiratory induced motion (RIM) can cause the tumor location to shift [7] by up to 35 millimeters [8]. This movement may result in missing the target, potentially causing misdiagnosis and false-negatives during a biopsy and exposing healthy liver tissue to ablative doses, potentially causing radiation induced liver disease [7].

B. Respiratory Induced Motion Compensation

Keall et al. [9] reviewed different techniques to manage RIM, that enable higher doses to be delivered to the tumor while sparing healthy tissue or to obtain a more reliable piece of tissue during biopsy. These methods mostly consist of specific breathing techniques, like deep or shallow breath holding to reduce RIM. Additionally, respiratory monitoring can be used to detect whether the lung inflation is at the same reproducible state. The review also discussed respiratory gating, where the treatment is only applied in a specific part of the patient's breathing cycle. This part is called the "gate" and the target should have a similar position throughout these gates. A problem with these methods is that the procedure can take a considerable amount of extra time because the patient needs to be coached or treatment has to be halted until the same gate is reached. Deep breath holding can also be uncomfortable and some patients might be unable to hold their breath for a sufficient amount of time.

Keall et al. [9] also suggest to track the tumor in real-time by inferring the tumor position from a surrogate signal. A surrogate signal is an additional signal that is simultaneously collected with imaging data to find a relationship between this signal and the tumor location. According to McClelland et al. [10], a surrogate signal should be relatively easy to acquire, have a high temporal resolution and be highly correlated with the internal motion. The relationship between this signal and the tumor location is described as a correspondence model that is fitted with a learning algorithm. Several surrogate signals have been proposed in literature. Some examples are optical markers [6], a depth camera [11], a reference needle [12], and an ultrasound transducer [13]–[16].

Using a motion model to accurately estimate the tumor position could be especially advantageous for guiding a surgical robot, enabling precise targeting during treatment. However, interpreting the tumor position as an XYZ-coordinate can be challenging, which limits its usability when a clinician is actively performing the procedure. Another solution is therefore to instead use the surrogate signal to generate scannerless images. In MRI guided procedures, the proposed solutions could provide two main advantages [13]. First, temporal resolution can be increased when the patient is inside the scanner ("in-bore") by image interpolation between acquired MR-images. Secondly, when the patient is moved out of the scanner ("out-of-bore"), scannerless MR-images can continue to be generated, while only relying on the surrogate signals. The application in the "in-bore" and "out-of-bore" situations are visualized in Fig. 1.

C. Research Questions

This research aims to compare and combine different surrogate signals to generate scannerless MR-images using a Generative Adversarial Network (GAN). The following research question was defined: "How can a generative adversarial network be utilized to generate real-time scannerless MR-images, using a number of surrogate signals?" To answer this research question, it is divided in the following two sub-questions:

1) How can information about inter- and intra-variable respiratory patterns be extracted from the surrogate signals and combined as input for the GAN?
2) How much do the scannerless images resemble the real images in different respiratory patterns, and what is each surrogate signal's influence on this result?

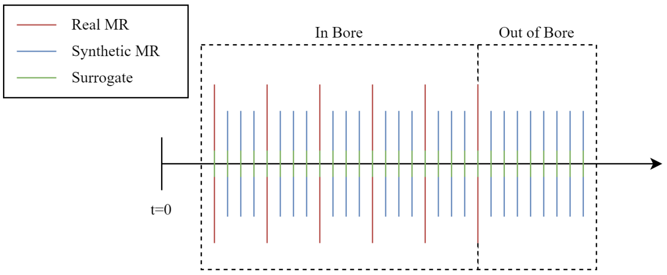


Fig. 1: Visualization of the applications of a motion model that can generate scannerless MR-images in the "in-bore" and "out-of-bore" situations. In the "in-bore" situation, the motion model can interpolate between real MR-images, greatly increasing its temporal resolution. In the "out-of-bore" situation, scannerless MR-images can be generated while only relying on the surrogate signals.

It is hypothesized that using a combination of multiple surrogate signals should improve the overall accuracy of the scannerless images.

D. Contributions

The contribution of this research is firstly to employ the conditional Progressive GAN architecture for real-time scannerless MR-image generation of the liver. A GAN was similarly employed in [17] to generate MR-images of the lungs, but they used the Pix2pix architecture from [18] that requires an image as input. The conditional Progressive GAN was introduced by Karras et al. [19] and later conditioned in [20]. When training this network, it starts generating images in a low resolution that is gradually increased. This improves training stability and allows shorter training times. Additionally, the conditioning does not necessarily have to be in image format, providing a more flexible way to add conditioning and making it simpler to combine multiple surrogate signals.

The second contribution is the combination of multiple surrogate signals for scannerless MR-image generation of the liver and comparing the influence of each surrogate signal on the quality of the results. The following three signals were compared: a thermal camera measuring airflow, an ultrasound transducer capturing internal abdominal motion, and external markers tracking external abdominal motion. This comparison is conducted through an ablation study, evaluating the quality of scannerless images generated by models trained with different combinations of surrogate signals. To our knowledge, this has not been done before. We additionally believe that the construction of a synchronized data set of the mentioned surrogate signals and MR-images of the liver is also a great contribution for future research.

E. Structure

This report is structured as follows: Section II describes the required background knowledge. Section III introduces related research papers. Section IV describes our proposed approach in detail. Section V describes the experiments that were conducted to acquire the results. Section VI introduces the results, that are discussed in Section VII. Section VIII provides the limitations of the proposed solution. Section IX proposes directions for future research and Section X concludes this research.

II. Background

This section introduces the background knowledge where this work is based upon. Firstly, the general information surrounding motion models and its components are described and lastly, the GAN architecture is explained.

A. Motion Model

The review made by McClelland et al. [10] describes a motion model as a model that estimates respiratory motion by finding the relationship between this motion M and some surrogate data s. This relationship is captured in the correspondence model ϕ and can be described as

$$M = \phi(s). \tag{1}$$

The surrogate signals should be able to easily acquire data that is highly correlated with the true respiratory motion with a high temporal resolution. Additionally, it should capture both intra-cycle and inter-cycle respiratory variability, which refer to the variability within a single breathing cycle and the variability between different breathing cycles, respectively. Some examples of surrogate signals that have previously been used are depth cameras [11], [21], ultrasound transducer [13], [14], [16], [17], [22], optical markers [6], and a reference needle [12].

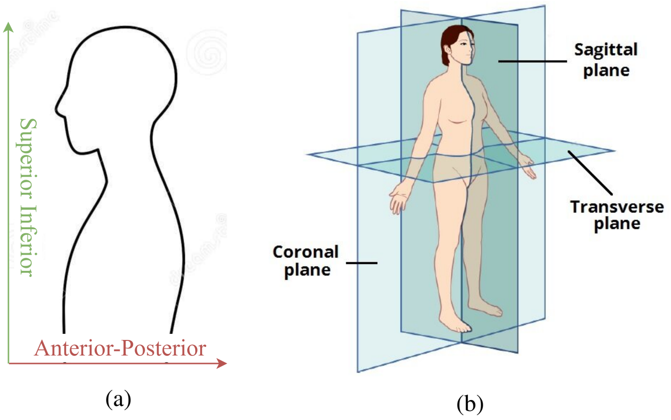
The true respiratory motion, which is the ground truth of the motion model can be extracted from some imaging modality [10]. For example, Fahmi et al. [6] segmented the liver border using thresholding and morphological operations of each MR-image, and used its location as the true motion. Cordon and Abayazid [11] similarly tracked the liver wall but from ultrasound as imaging modality.

According to Keall et al. [9], the primary organ motion is in the superior-inferior (SI) direction, while the displacement in the anterior-posterior (AP) and lateral direction is typically less than 2 millimeters. This needs to be considered when choosing the plane for the imaging modality, since the transverse plane does not capture the main breathing motion in the SI direction. The SI and AP breathing directions and the imaging planes are visualized in Fig. 2.

The correspondence model ϕ can be any regression model, like a linear regressor, polynomial regressor, or a more complex deep learning model like a neural network.

B. Generative Models

The Generative Adversarial Network, introduced by Goodfellow et al. [23], is a generative model, where a generator G tries to generate data that fools the discriminator D, which attempts to classify given samples as real or fake. To train this model, a minimax game is played, where D maximizes the probability that it correctly classifies a given sample, and G minimizes the probability that its output is classified as fake. This can be formulated in an objective function as



(a) (b)

Fig. 2: (a) The Anterior-Posterior (AP) and Superior Inferior (SI) respiratory motion directions with respect to the side view of the human body and (b) a visualization of the coronal, sagittal, and transverse plane with respect to the human body.

$$\min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_r}[\log D(x)] + \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[\log(1-D(\tilde{x}))], \tag{2}$$

where $\mathbb{P}_r$ is the distribution of the real data, $\mathbb{P}_g$ is the distribution of the data generated by the generator and $\tilde{x} = G(z)$ with z being random noise.

After the introduction of GAN, many improvements were introduced to broaden its use cases. One that was introduced shortly after the GAN by Mirza and Osindero [24], is the conditional GAN (cGAN). The cGAN can be conditioned by an extra piece of information y by simply adding y as an additional input in G and D.

Arjovsky et al. [25] recognized the instability when training a GAN. This instability can cause mode collapse, where the generator would only generate samples that represent a small subset of the training data. They attempted to improve this by introducing the Wasserstein GAN (WGAN) [25], that uses the Earth Movers distance to measure the difference between the distributions $\mathbb{P}_r$ and $\mathbb{P}_g$. The objective function can then be formulated using the Kantorovich-Rubinstein duality [26] as

$$\min_G \max_{D\in\mathcal{D}} \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})], \tag{3}$$

where $\mathcal{D}$ are all the Lipschitz functions and $\tilde{x} = G(z)$. To satisfy the 1-Lipschitz constraint, the authors clip the weights to $[-c, c]$. According to Gulrjani et al. [27], this weight clipping can lead to exploding or vanishing gradients and limited model capacity. The authors therefore introduce a gradient penalty to enforce the Lipschitz constraint. The equation for the gradient penalty is written as

$$GP = \lambda \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2], \tag{4}$$

where $\hat{x}$ is uniformly sampled from a straight line between pairs sampled from $\mathbb{P}_g$ and $\mathbb{P}_r$, and λ is the gradient penalty coefficient. The objective for the WGAN with gradient penalty (WGAN-GP) then becomes

$$\min_{G} \max_{D \in \mathcal{D}} \mathbb{E}_{x \sim \mathbb{P}_r}[D(x)] - \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g}[D(\tilde{x})] + GP. \tag{5}$$

Karras et al. [19] attempted to further improve training stability by proposing the Progressively Growing GAN (ProGAN) that attempts to learn features of the training data in a coarse to fine manner. This is done by gradually increasing the resolution of the generated samples by adding new layers to G and D. A new layer is smoothed in by first considering it as a residual block, weighted by α, that linearly increases from 0 to 1. It uses the Wasserstein Loss function with gradient penalty [27] for optimization.

III. Related Work

Preiswerk et al. [13] proposed a system that uses an ultrasound transducer as surrogate signal to greatly enhance the temporal resolution when the subject is "in-bore", or to create scannerless MR-images when the subject is "out-of-bore". The first step is to create a subject specific history of paired MRI and ultrasound data. After the history is constructed, a similarity measure is computed between the current ultrasound signal, and all the ultrasound signals in the history. A scannerless image is finally constructed by taking the average of all the scans in the history, weighted by the similarity measures. This work was attempted to be improved by Shokry [15] by separately applying the methodology from Preiswerk et al. [13] to all the entries of the k-space. Both solutions faced the problem that the correlation between ultrasound and MRI data was lost when the ultrasound transducer was displaced. It also faced efficiency problems when a large history was formed, since the computational complexity is dependent on the size of the history. Veenstra [16] compared the methodology from Preiswerk et al. [13] and Shokry [15] after applying several pre- and postprocessing techniques, aimed at enhancing performance and efficiency.

Preiswerk et al. [14] attempted to improve their previous work from [13] by combining a Convolutional Neural Network (CNN) and a long-short term memory (LSTM) with ultrasound data as input to generate scannerless MR-images. The use of a deep learning solution was motivated by having a constant complexity at inference stage that is thus independent of the data set size. The authors achieved a ten times faster reconstruction time with only a slight increase in error between real and scannerless images (Pixel-wise sum of squared error of 39.0 ± 12 pixels for [14] vs. 33.9 ± 7 pixels for [13]).

Giger et al. [17] proposed a solution where a 2D ultrasound transducer is used as a surrogate signal to generate scannerless MR-images. They used the Pix2pix model, introduced by [18], which is a GAN architecture specifically made for image-to-image translation. Navigator-based 4D MRI and 2D ultrasound was simultaneously collected and the GAN was trained to predict the navigator deformation field that can be used to reconstruct a scannerless 3D MR-volume.

IV. Methodology

This work proposes a deep learning approach that combines multiple surrogate signals to generate out-of-bore scannerless MR-images for real-time guidance during liver biopsy and ablation. Fig. 3 shows that the system is divided into a training phase and a testing phase. In the training phase, a dataset is created by simultaneously collecting surrogate data and MR-images. From each surrogate signal, the breathing waveform is extracted and synchronized with the MR-images. The final step of the training phase is to train the generative model by using the breathing waveforms as input, and the MR-images as ground truth data. In the testing phase, the subject can be placed out-of-bore and the trained generator can be used to create scannerless MR-images by inputting the real-time breathing waveforms, extracted from the surrogate signals.

A. Data Collection

The data is collected in human subject experiments, where seven subjects perform specific breathing protocols, while MR-images and surrogate data is simultaneously collected. The surrogate signals capture data on the internal breathing motion, external breathing motion, and the breathing airflow from the subject's mouth. Each signal captures a different element of the breathing information, which potentially improves the capture of intra-cycle variability. All the surrogate data is gathered on the same machine and the timestamp is recorded for each signal. For each subject, three sessions are performed to test repeatability. The schematic view of the data collection setup is introduced in Fig. 4.

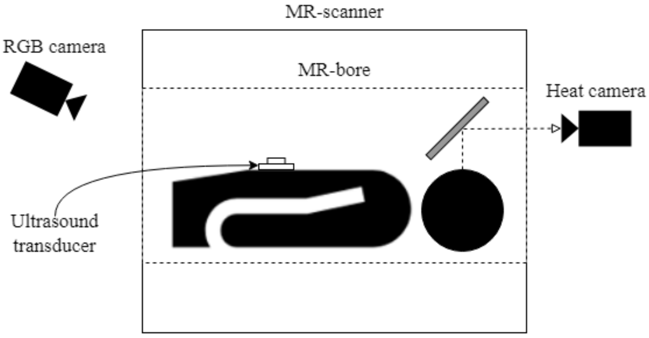


Fig. 4: Schematic side view of the data collection setup, where the subject is laying inside the MR scanner. The ultrasound transducer is attached to the subject's abdomen and an RGB-camera is pointed at the scanner's receiver coil to track its movement. A mirror is placed above the subject's face, to make the heat that results from breathing, visible from the heat camera.

1) MRI: MRI has been chosen as ground truth data because, according to Panych and Tokuda [28], it is considered to be ideal for image guided interventions. Unlike CT, X-ray, and PET, it is free of ionizing radiations, making it noninvasive. Additionally, it provides better detail on soft tissue than CT and US [16]. It was chosen to acquire images in the sagittal

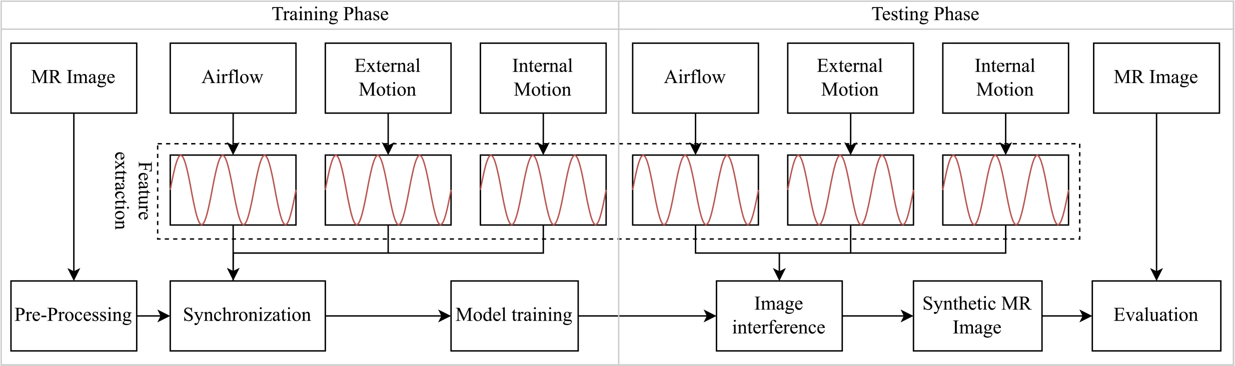


Fig. 3: Flow diagram of the proposed approach: In the training phase, the MR-images and surrogate signals (airflow, external motion, and internal motion) are simultaneously collected. The MR-images are pre-processed and the breathing waveforms are extrected from the surrogate signals in the feature extraction step. Next, the MR-images and surrogate signals are synchronized and used to train the deep learning model. The trained model is used in the testing phase, where it uses the breathing waveforms from the surrogate signals to generate scannerless MR-images. The scannerless and real MR-images are finally compared in the evaluation step.

TABLE I: The most relevant settings that were used in the MRI-scanner.

Setting	Value
Field strength	1.5 T
Orientation	Sagittal
Slice thickness	10 mm
Spatial resolution	1.9 mm by 1.9 mm
Base resolution	192×192 pixels^2
Repetition time	355.60 ms
Echo Time	1.17 ms
Trajectory	Cartesian

plane, so both the SI and AP motion of the liver border are visible.

The MRI machine that was used is the SIEMENS MAGNETOM AERA 1.5 T (Siemens Healthineers, Erlangen, Germany), located in the TechMed centre at the University of Twente. The most relevant settings that are used can be found in Table I.

2) Internal Breathing Motion: An ultrasound transducer was used to capture internal breathing motion. It has previously been employed in [13], [14], [16] as a surrogate signal to create scannerless MR-images. Madore et al. [22] also extracted a single dimensional breathing waveform from the ultrasound data. An additional benefit is that the transducer is attached to the body and can thus follow the patient through different locations [22].

TABLE II: The settings that were applied in the ultrasound transducer.

Setting	Value
Pulse voltage	240 V (level 10)
Pulse width	2.8 μs
Sampling frequency	33.3 MHz
Analog filter	2-6 MHz
Gain	24 (pre-amplifier) + 15 (constant) dB
Delay	10 μs
(Measurement) window	80 μs
Trigger	timer (PRF)
PRF	50 Hz

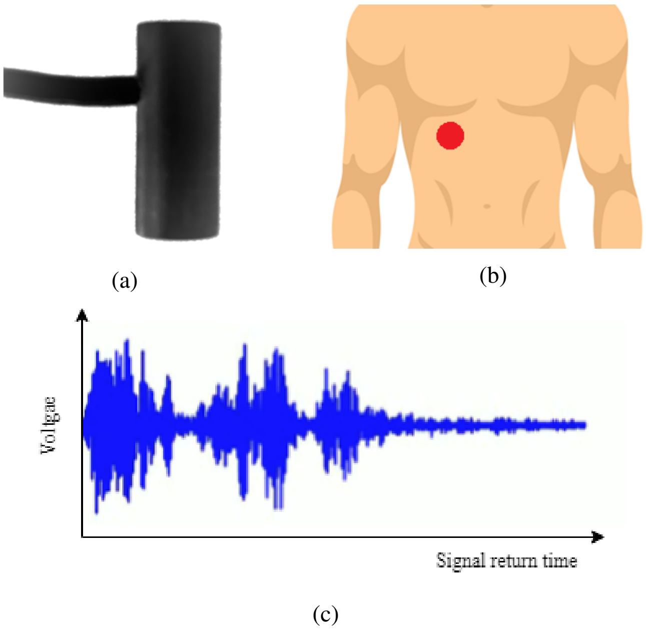
This work uses the MR-compatible Optel Opbox 2.1 ultrasound transducer. This same device was successfully used before by Veenstra [16] and therefore, the same settings were applied. These settings can be found in Table II.

The ultrasound probe is placed inside a 3D printed housing and attached with double sided tape to a plaster on the right side of the subject's abdomen, just below the ribs. A picture of the transducer, the location where it is placed on the abdomen and what a single firing of this sensor looks like can be found in Fig. 5.

3) External Breathing Motion: The external breathing motion is captured by visually tracking interest points on the subject that move during breathing, using the Intel RealSense L515 LiDAR depth camera (Santa Clara, California, United States). Fig. 6 shows the receiver coil of the MR scanner that was chosen as interest point. The coil was easy to track and clearly moved during breathing, making it highly correlated to the subject's RIM. A depth camera was used because the original plan was to use the subject's volume as a surrogate signal, but this could not be reliably extracted due to the range and positioning of the camera. The camera was capturing RGB-images at 30 frames per second and a resolution of 640×480 pixels^2.

4) Breathing Airflow: The FLIR A615 heat camera was used to capture information on airflow that results from exhalation and inhalation. The subject had to wear a face mask that absorbs the heat from the breathing airflow and a mirror was placed above the subject's face to make the heat signature visible from outside the magnetic field. The face mask in the mirror and the view from the heat camera can be found in Fig. 7. The camera captures heat images at 50 frames per second (fps) and a resolution of 640×480 pixels^2.

5) Breathing Protocol: The subjects were asked to perform the following breathing protocol when placed inside the scan-



(a) (b)

(c)

Fig. 5: The Optel Opbox ultrasound transducer (a) is placed on the right side of the subject's abdomen, just below the ribs, as shown as the red circle in (b). The graph in (c) shows a single firing of the ultrasound transducer with the signal return time on the horizontal axis, and the voltage of the received signal on the vertical axis.

ner:

- Fully inhaled breath hold: 15 seconds
- Shallow breathing: 3 minutes
- Half exhaled breath hold: 15 seconds
- Regular breathing: 3 minutes
- Fully exhaled breath hold: 15 seconds
- Deep breathing: 3 minutes

The protocol was created to evaluate whether the method is generalizable across different subjects. During shallow breathing, the subject is asked to breath faster but with a lower amplitude, regular breathing is how the subject would normally breath, and deep breathing means slower breathing with a higher amplitude. The breath holds are performed in different positions, so it is possible to evaluate whether the inconsistency between breath holds is accurately modeled. Additionally, the subject was asked to breath through the stomach, to enhance the movement of the coil, and to breath through the mouth, to improve the heat signature on the face mask.

6) Subjects: Seven healthy subjects (A-G) participated in this study, including five males and two females. All subjects signed a consent form and the experiments are approved by the Natural Sciences and Engineering Sciences ethics committee from the University of Twente. All subjects were asked to perform the previously described breathing protocol three times. In some instances, a subject was briefly removed from the scanner between sessions, or sessions were performed on different days. The first session from subjects C and D are excluded from the final evaluation, because the subject did not follow protocol and there was not enough time to do a rerun.

B. *Feature Extraction*

From each surrogate signal, the breathing waveform is extracted. A surrogate signal captures the secondary motion caused by respiration, where the breathing waveform resembles a sinusoidal wave due to the periodic nature of breathing. This subsection describes how the waveforms are extracted from each surrogate signal.

1) Internal Breathing Motion: The breathing waveform is extracted from the internal breathing motion following the method outlined by Madore et al. [22]. A single firing of the ultrasound transducer generates a one-dimensional signal, capturing the amplitude of the transducer voltage as a function of the travel time of the ultrasound pulse. All the signals in the collected dataset can be seen as a two dimensional image

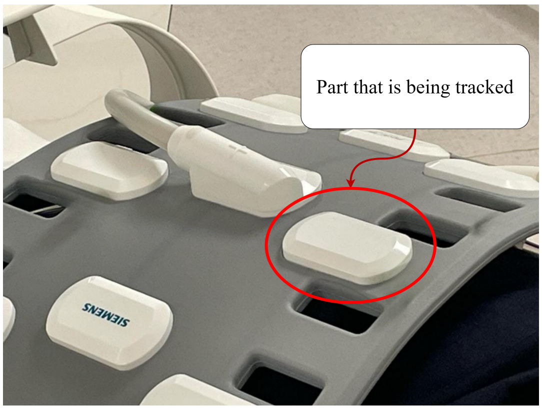
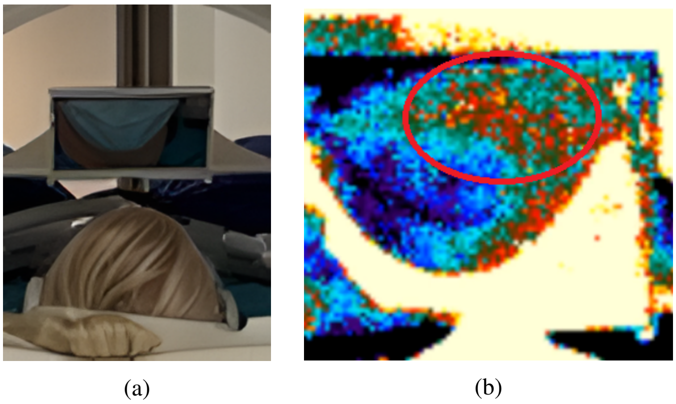


Fig. 6: The MRI receiver coil of which the movement is being tracked to capture the subject's external motion that is caused by breathing.



(a) (b)

Fig. 7: The face mask visible in the mirror placed above the subject's face (a) and a single heat image captured by the camera with the region of interest where the average temperature is extracted from (b).

with axes travel time t and timestamps T. These signals were converted into velocity measurements using

$$V = (\lambda/2\pi) * (\Delta\theta/\Delta T), \tag{6}$$

where θ is the phase and λ is the wavelength. The velocity measurements were transformed back into a single dimensional signal $v(T)$, by applying the median operator along the t axis. Finally, the breathing waveform is extracted through the integral

$$z(T) = \int_0^T v(T')dT'. \tag{7}$$

Furthermore, Madore et al. [22] removed a linear trend from $z(T)$ but it was found that the trend in the waveforms from this work's data were different for each breathing pattern. The trends were therefore separately removed for each pattern. The trend was determined by finding pairs of the most similar signals in the magnitude and computing the slope between these pairs. Finally, these slopes are averaged and a linear line with this average as slope is removed from the signal.

2) External Breathing Motion: The breathing waveform is extracted by tracking the vertical movement of the MR-scanner's receiver coil in the RGB-images. The receiver coil is placed over the subject's abdomen and moves up and down as a result of breathing. This movement can be tracked by tracking the outline of one of the clear white areas on the surface of the coil that can be seen in Fig. 6. The RGB-images are converted to grayscale images and simple thresholding is sufficient to reliably filter out the white areas. A vertical line is manually placed over one of the white areas, and the top intersection between the line and the masked area is recorded for each image. This intersection over all the images results in the breathing waveform.

3) Breathing Airflow: The breathing waveform can easily be extracted from the heat images by taking the average temperature from the subject's face mark.

C. MRI Preprocessing

The preprocessing steps that are applied to the MR-images are visualized in Fig. 8. The first step is to enhance the image's contrast and to crop it to the liver and lungs region. The region under the liver is removed, because it includes a lot of artifacts that are caused by the ultrasound transducer. The resulting image is finally scaled to a range of $[-1, 1]$, to match the pixel value range of the generator's output, and used as ground truth images during training.

The next step is to extract the breathing waveform from the MR-images, that is used for synchronization and evaluation. The respiratory-induced motion can be extracted by thresholding the preprocessed image and applying morphological operators. Finally, a vertical line is manually placed at the highest point of the liver, and the top intersection between the mask and the line is tracked. Taking the y-position of the point in each frame results in the breathing waveform.

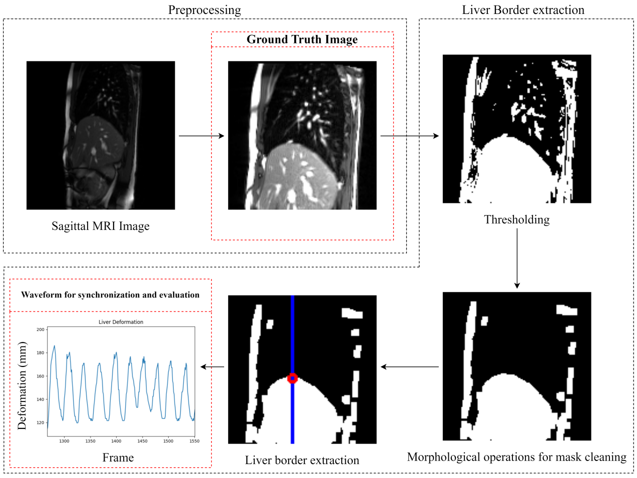


Fig. 8: Preprocessing steps of an MR-image from subject A. The image's contrast is enhanced before it is cropped to cut of the artifacts. Next, the breathing motion is extracted by thresholding and morphological operations to find the liver border.

D. Synchronization

As stated before, the data of all surrogate signals is gathered on the same machine. This means that the surrogate signals can be synchronized based on their timestamps. Additionally, the breathing waveform of external breathing motion is interpolated to match the temporal resolution of the ultrasound transducer and the heat camera, which is 50 Hz for both.

To synchronize the surrogate signals with the ground truth MR-images, an interest point in the ultrasound waveform is matched to the same interest point in the breathing waveform that is extracted from the MR-images. The interest points that were easily located in most instances, was the first fully inhaled position during deep breathing. The MR breathing

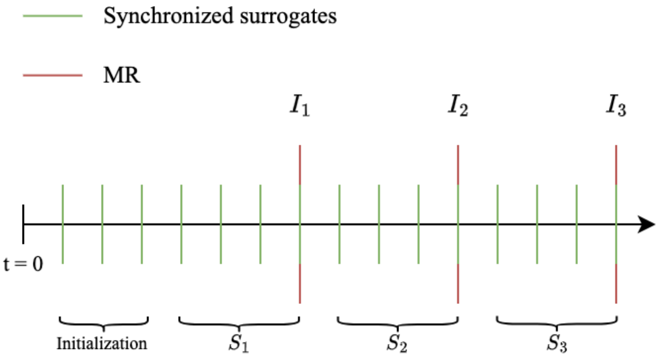


Fig. 9: A timeline with the MR-images I_t and the synchronized surrogate signals S_t. Since the temporal resolution of the surrogate signals is higher than that of the MR-images, all the surrogates collected between I_t and I_{t-1} were coupled to I_t.

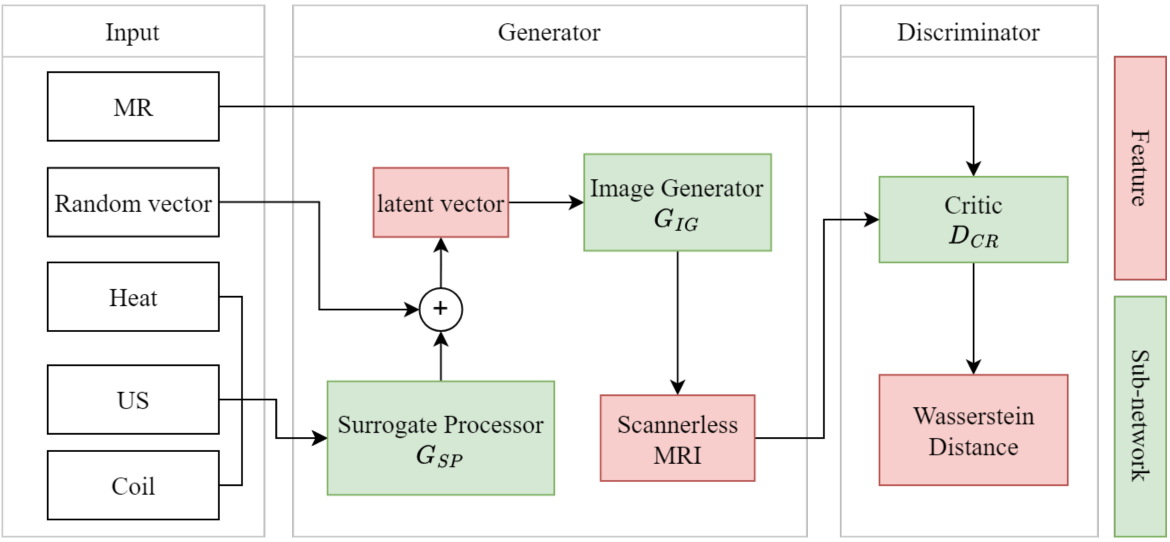


Fig. 10: Flow diagram of the Conditioned ProGAN architecture, inspired from [20]. The network consists of a Generator, that takes a random vector and the surrogate signals to generate a scannerless MR-Image. The Discriminator takes scannerless and real MR-images, and computes the Wasserstein Distance between their distributions, used for optimization.

waveform is extracted by tracking the SI displacement of the liver border. It was chosen to perform the synchronization based on the ultrasound data because Madore et al. [22] showed great correlation between this and the movement of the liver border.

Fig. 9 illustrates how input data is paired to the ground truth data. Each MR-image I_t is paired to surrogate data S_t, which is a combination of the ultrasound data U_t, tracked coil data C_t, and heat data H_t, collected between I_t and I_{t-1}.

E. Model Architecture

The deep learning architecture used in this work is strongly inspired by Arshad and Beksi [20], who made a conditional version of the Progressive Growing Generative Adversarial Network (ProGAN), that was first introduced by Karras et al. [19].

The main components of the conditional ProGAN (cProGAN) architecture are the Discriminator D and Generator G. The networks are optimized using the Wasserstein loss function, first introduced by Arjovsky et al. [25], with gradient clipping, which was later introduced by Gulrajani et al. [27]. Fig. 10 introduces an overview of the entire architecture. The structure and details of the subnetworks can be found in Appendix A.

1) Generator: The generator G uses a random vector $\mathbf{z}_{\text{in}} \sim \mathcal{N}(0,1)$, where $\mathbf{z}_{\text{in}} \in \mathbb{R}^{32}$, and surrogate signals $S_t = \{U_t, C_t, H_t\}$ to generate a scannerless MR-Image $\hat{I}_t$. The generator consists of two subnetworks: the surrogate processor G_{SP} and the image generator G_{IG}.

G_{SP} is a multi-layer perceptron (MLP) with a single hidden layer and its purpose is to process S_t and extract its important features. The surrogate signals are first concatenated into an input vector $\mathbf{x}_{\text{t}} = (U_t, C_t, H_t)$ and fed through G_{SP} to construct the feature vector $\hat{\mathbf{x}}_{\text{t}} \in \mathbb{R}^{32}$. This feature vector $\hat{\mathbf{x}}_{\text{t}}$ and $\mathbf{z}_{\text{in}}$ are concatenated to latent vector $\mathbf{z} = (\hat{\mathbf{x}}_{\text{t}}, \mathbf{z}_{\text{in}})$, which is the input for G_{IG}.

The subnetwork G_{IG} is the progressively growing generator, similar as in the original ProGAN architecture [19]. The main difference is that G_{IG} is not trained directly from the lowest resolution possible (4×4), but starts training at a resolution of 32×32. In a total of three steps, it is doubled twice using nearest neighbor interpolation and reaches a resolution of 128×128. It was found that starting from the lowest possible resolution resulted in images with a lower sharpness that include artifacts. The Tanh activation function is applied on the output layer, making the pixel value range $[-1, 1]$.

2) Discriminator: The Discriminator D consists of a single network, the critic D_{CR}. Unlike in [20], the critic is not conditioned by the surrogate signals because it was found that conditioning reduced the smoothness of the motion in the scannerless images. The critic progressively grows during training as outlined in the original ProGAN [19] and its architecture is an exact mirror of G_{IG}. It takes either a real or scannerless MR-image as input, and outputs a single scalar value with a linear activation function, used to calculate the Wasserstein loss function. The resolution of the real MR-images is reduced using average pooling and its pixel values are scaled to a range of $[-1, 1]$, to ensure that the real and scannerless images match pixel value range and resolution at each step.

F. Performance Metrics

The performance of the trained generator is evaluated using two metrics. The first metric is the Structural Similarity Index Measure (SSIM), which assesses the similarity of two given images. The other metric that we use is comparing the extracted motion from the real and scannerless images, to assess how well the RIM is captured by the model.

Structural Similarity Index Measure: The SSIM was introduced by Wang et al. [29], and compares two nonnegative images, x and y, based on their luminance $l(\cdot)$, contrast $c(\cdot)$ and structure $s(\cdot)$, using the following equation:

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \tag{8}$$

When the parameters α, β, and γ are set to 1, the equation can be written as the following:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{9}$$

The equations use the means μ and standard deviations σ of the images, and small constants C_1 and C_2 that avoid zero divisions. In practice, SSIM is applied locally, by sliding a window over the image and averaging the result.

Liver Deformation: The liver deformation in the real and scannerless images are compared to evaluate how well the model captures the respiratory motion. The liver deformation waveform can be extracted from the real and scannerless images as explained in Section IV-C and compared using the Mean Absolute Error (MAE) that is computed by the following equation:

$$MAE(y,\hat{y}) = \frac{1}{N}\sum_{i=0}^{N} |y_i - \hat{y}_i|, \tag{10}$$

where N is the number of data points, y is the ground truth liver motion, and $\hat{y}$ is the estimated liver motion. The MAE can be computed in millimeters, making it easily interpretable. It can also be computed as a percentage of the average peak-to-trough (PTT) distance of the ground truth data to make the results comparable in different breathing amplitudes. The PTT distance is computed as the average distance between the peaks and the troughs in the breathing waveform that is extracted from the ground truth MR-images. We refer to this as the MAE percentage (MAE %).

Another metric to evaluate the motion in the scannerless images is the coefficient of determination, often denoted as R^2. It is computed with the following equation [30]:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}, \tag{11}$$

where $TSS = \sum(y_i - \bar{y})^2$ is the total sum of squares, and $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares. In this equation, $\bar{y}$ is the mean value of y.

V. Experiments

All experiments are run offline after the data sets for the different subjects are collected. This section describes how the performance of the trained generator and quality of the scannerless MR-images are evaluated.

A. Data Preparation

Each breathing pattern in a single session's data is split into training data (80%), validation data (10%), and testing data (10%) and a single model is trained using a combination of the training subsets. Validation and testing is performed separately for each breathing pattern. Excluded from the training data are the breath holds, since these had very few samples. For each breath hold, 50% of the data was used in evaluation, and 50% in testing.

TABLE III: Peak to trough distance of the liver motion in millimeters ($\mu \pm \sigma$), extracted from the ground truth images.

Subject	**Shallow**	**Regular**	**Deep**
A	19.13 ± 5.12	35.18 ± 10.41	78.22 ± 21.81
	21.57 ± 6.30	36.86 ± 8.71	97.26 ± 16.54
	34.76 ± 16.92	51.37 ± 14.59	99.73 ± 14.09
B	22.97 ± 3.18	21.59 ± 6.75	53.12 ± 17.55
	28.77 ± 5.45	26.52 ± 8.98	81.40 ± 12.60
	20.89 ± 3.16	17.44 ± 5.28	74.73 ± 10.09
C	81.43 ± 13.08	86.43 ± 14.20	112.09 ± 23.82
	42.28 ± 5.36	67.25 ± 11.29	115.38 ± 4.84
	44.52 ± 4.78	101.34± 18.68	119.85 ± 4.68
D	33.13 ± 6.11	27.80 ± 5.24	21.15 ± 6.70
	16.74 ± 4.12	27.48 ± 5.45	53.28 ± 9.09
	21.84 ± 5.65	25.48 ± 6.76	65.94 ± 7.16
E	45.00 ± 6.26	52.86 ± 7.10	114.98 ± 16.71
	33.23 ± 6.32	31.66 ± 5.66	66.29 ± 18.39
	20.40 ± 4.27	23.15 ± 3.83	39.59 ± 11.63
F	34.08 ± 5.49	57.93 ± 11.50	112.85 ± 10.70
	35.40 ± 6.46	56.72 ± 5.78	108.44 ± 7.74
	31.73 ± 36.56	46.86 ± 9.01	112.18 ± 19.05
G	65.52 ± 35.91	67.20 ± 8.32	137.95 ± 6.13
	92.26 ± 37.27	77.20 ± 9.98	127.74 ± 32.89
	52.95 ± 14.67	65.08 ± 13.33	124.34 ± 5.67
Mean	38.03 ± 11.07	47.78 ± 9.09	91.26 ± 13.23

The data from different sessions was kept separated and training and evaluation is performed on a single session. The reason for this is that some sessions were performed right after each other without moving the subject and other times, there was a break between sessions, or they were even performed on different days. Therefore, MR-images from different sessions may look slightly different, which deteriorates the comparative metrics.

B. Training Procedure and Model Selection

For each session, the combined training data was used to train the conditional ProGAN model for sixty epochs, taking 765 seconds on average when trained on a GPU (NVIDIA RTX 4070). The generator and the discriminator are both trained using the Adam optimizer [31], with parameters $\beta_1 = 0$, $\beta_2 = 0.99$, and $\varepsilon = 10^{-8}$. A linear schedular was used for the learning rate α, that started as 0.001 and was linearly reduced to 0.00001 after sixty epochs. During training, the SSIM is computed on the validation data for each epoch and model checkpoints are saved. The model checkpoint from the epoch that resulted in the highest average SSIM score over all breathing patterns is selected for final evaluation.

C. Evaluation Criteria

To evaluate the importance and quality of each surrogate signal, an ablation study is performed by comparing the accuracy of the model when it is trained using different combinations of surrogate signals. Evaluating all possible combinations would be infeasible, so it was decided to only train on all surrogate signals combined, and all of them separately (a total of four models).

TABLE IV: Table with the results during different breathing patterns. MAE is in millimeters and MAE % indicates the MAE as a percentage of the average PTT distance during corresponding breathing pattern. The SSIM and R^2 are percentages.

Model	Metric	Deep Breathing	Shallow Breathing	Regular Breathing	Mean
Combined	MAE ↓	7.42 ± 4.95	4.75 ± 3.23	4.85 ± 3.55	5.67 ± 3.91
	MAE % ↓	8.36	14.05	11.73	11.38
	R^2 ↑	55.09	-17.29	22.90	20.24
	SSIM ↑	46.49 ± 7.16	54.85 ± 6.13	52.93 ± 6.65	51.42 ± 6.65
External	MAE ↓	7.11 ± 5.27	3.65 ± 2.72	4.30 ± 3.23	5.02 ± 3.74
	MAE % ↓	8.19	11.76	10.29	10.08
	R^2 ↑	54.44	14.26	42.74	37.15
	SSIM ↑	44.52 ± 7.16	55.01 ± 5.02	52.71 ± 6.25	50.75 ± 6.14
Internal	MAE ↓	14.61 ± 9.49	6.78 ± 4.10	9.26 ± 5.72	10.22 ± 6.44
	MAE % ↓	15.74	19.04	21.11	18.63
	R^2 ↑	-29.78	-105.81	-81.42	-72.34
	SSIM ↑	37.13 ± 6.84	46.93 ± 5.11	41.94 ± 6.05	42.00 ± 6.00
Airflow	MAE ↓	17.67 ± 12.53	8.44 ± 5.40	8.42 ± 5.76	11.51 ± 7.90
	MAE % ↓	18.49	23.14	18.13	19.92
	R^2 ↑	-73.48	-203.04	-34.40	-103.64
	SSIM ↑	35.75 ± 8.84	44.12 ± 5.84	44.11 ± 6.00	41.32 ± 6.89

TABLE V: Table with the results during different breath holding positions. MAE is in millimeters and the SSIM is in percentages.

Model	Metric	Half Exhaled	Fully Exhaled	Fully Inhaled	Mean
Combined	MAE ↓	7.25 ± 2.38	11.08 ± 1.48	17.36 ± 1.97	11.90 ± 1.95
	SSIM ↑	39.09 ± 2.09	40.02 ± 1.52	25.08 ± 1.68	34.73 ± 1.76
External	MAE ↓	7.34 ± 1.57	8.05 ± 1.02	12.04 ± 1.32	9.14 ± 1.31
	SSIM ↑	38.88 ± 1.22	43.81 ± 1.15	26.71 ± 1.46	36.47 ± 1.28
Internal	MAE ↓	10.86 ± 1.89	16.54 ± 1.80	28.25 ± 1.80	18.55 ± 1.83
	SSIM ↑	33.81 ± 1.28	30.26 ± 1.11	18.80 ± 1.19	27.63 ± 1.20
Airflow	MAE ↓	12.15 ± 2.87	14.23 ± 3.08	28.37 ± 2.96	18.25 ± 2.97
	SSIM ↑	32.95 ± 2.21	35.27 ± 1.71	19.62 ± 1.88	29.28 ± 1.93

Each model was evaluated by assessing the quality of the scannerless images that were generated using the testing data. The first component of this evaluation is to assess how well the model captures the respiratory motion by comparing the extracted liver deformation in the real and scannerless images using the MAE and R^2 metrics. The other component is to assess how well the scannerless images represent the real images. This is done by computing the SSIM with a window size of 11 × 11 pixels^2 between the real and corresponding scannerless images. A high SSIM indicates a high similarity between two images. Section IV-F described how these metrics are computed.

VI. Results

This section presents the results that are obtained by evaluating the accuracy of the scannerless MR-images, created by the generator. Table III shows the RIM, extracted from the ground truth images, that is represented by the average peak-to-trough (PTT) distance over the separate breathing patterns. It is shown that the average RIM over all the subjects is 38.03 ±11.07 mm during shallow breathing, 47.78 ± 9.09 mm during regular breathing, and 91.26 ± 13.23 mm during deep breathing.

The results of the evaluation of the scannerless images during breathing are presented in Table IV. The external model achieved the lowest MAE over the extracted motion, with an average value of 5.02 ± 3.74 mm. The external motion model also performed the best in terms of the MAE as a percentage of the average PTT distance and the R^2 value, scoring 10.08% and 37.15%, respectively. However, the SSIM is slightly higher for the the combined model, being 51.42%, compared to 50.75% for the external motion model. The combined and external motion model performed best during deep breathing according to the MAE as percentage of the average PTT and the R^2 metrics. Clearly, the models trained exclusively on the internal motion and airflow surrogates perform significantly worse across all metrics with respect to the external and combined models, with the R^2 even being entirely negative.

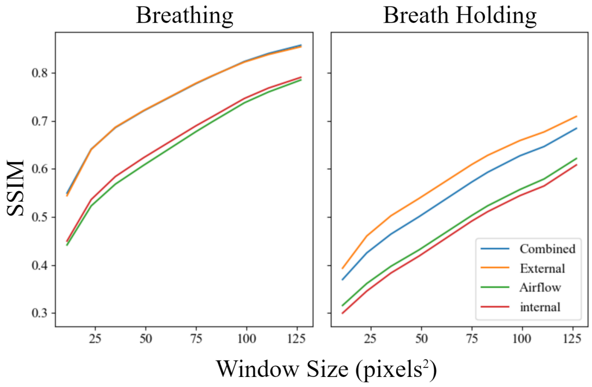


Fig. 11: SSIM between real and scannerless images, as a function of the window size. The left figure represents the average SSIM over all the sessions from all the subjects during breathing, and the right figure during breath holding.

The MAE and SSIM during the different breath holding

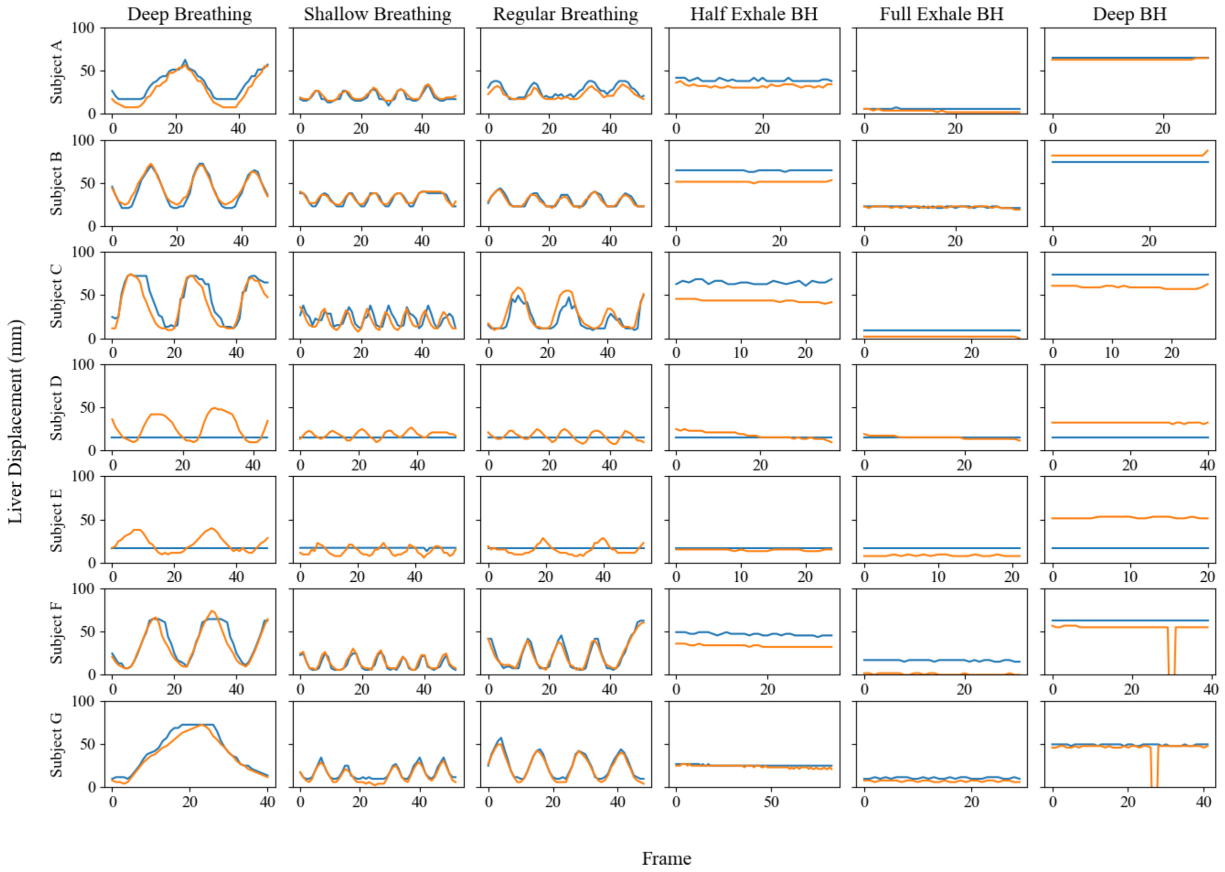


Fig. 12: Breathing motion extracted from the real (orange line) and scannerless (blue line) images using only the external breathing motion as a surrogate signal for a single session of all the subjects. The vertical axis represents the liver displacement in millimeters, and the horizontal axis the frame number.

positions are presented in Table V. The performance of the model trained on the external surrogate signal exclusively has the best MAE and SSIM, being 9.14 ± 1.31 mm and 36.47%, respectively. This model was the most accurate during the half exhaled breath hold, with an MAE and SSIM of 7.34 ± 1.57 mm and 38.88%, respectively. Again, the internal motion and airflow models perform significantly worse than the combined and external models.

Examples of the comparison between the motion in the real and scannerless images that are generated using the external motion model are presented in Fig. 12. The motion extracted from images that are generated using the combined model are presented in Fig. 13. The figures include the motion in the different breathing patterns for a single session of all the subjects. Subjects F and G are missing an MR frame, resulting in the zero value during the deep breath hold. This frame is excluded from the final evaluation.

The average interference time per image, when generating 10,000 images, is 3.6 ms when using a GPU (NVIDIA RTX 4070) and 7.9 ms when using a CPU (AMD Ryzen 7 5800X). The difference between the different models was negligible.

As previously stated, the SSIM values are all computed with a window size of 11 × 11 pixels^2. Fig. 11 presents the SSIM values as a function of the window size. It can be seen that the SSIM gradually increases as the window becomes larger. During breathing, the combined model and the external model result in similar SSIM values across all window sizes. During breath holding, the external model results in a higher SSIM value than the combined model. During both breathing and breath holding, the airflow and internal motion models perform significantly worse with respect to the combined and external motion models.

Three comparisons of real and scannerless images are presented in Fig. 14. The examples include the image with the highest, lowest, and median SSIM over all three sessions of subject A.

VII. Discussion

This research proposed a deep learning solution to generate scannerless MR-images for assistance during image guided interventions. Different surrogate signals were compared that capture the internal breathing motion, the external breathing motion, and the breathing airflow. A combination of all surrogate signals was also investigated. Data was collected

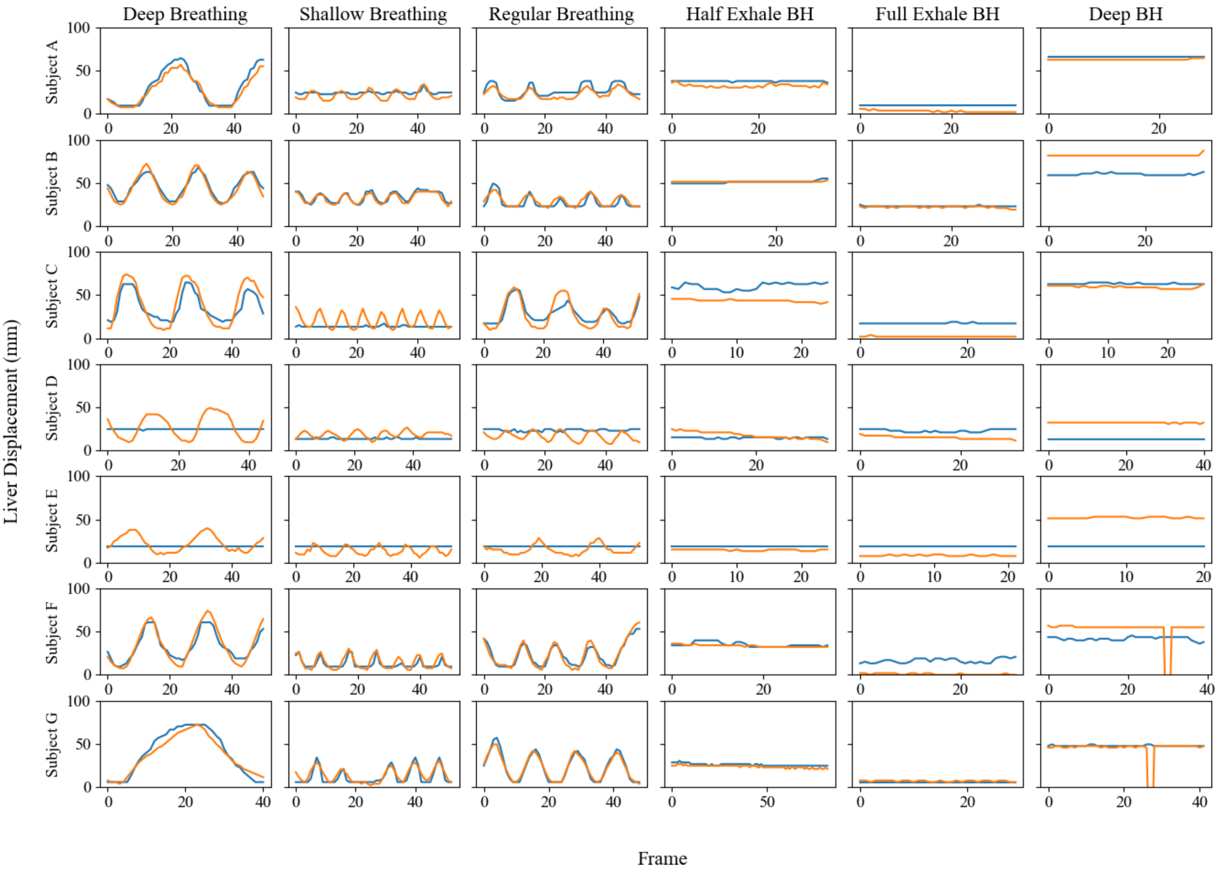


Fig. 13: Breathing motion extracted from the real (orange line) and scannerless (blue line) images using all the surrogates for for a single session of all the subjects. The vertical axis represents the liver displacement in millimeters, and the horizontal axis the frame number.

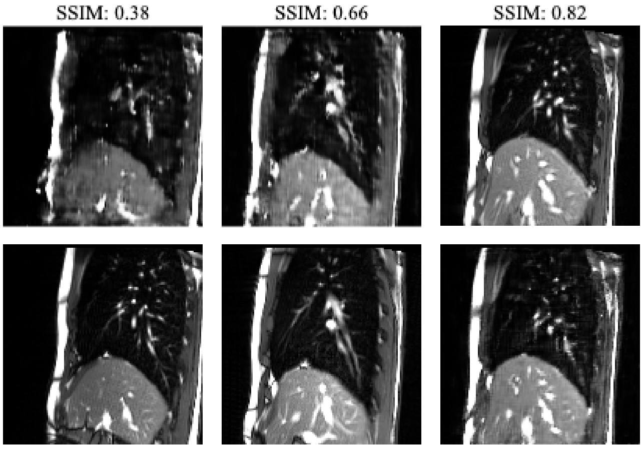


Fig. 14: Example images, where each column represents the image with the worst, median, and best SSIM from subject A, generated using the combined model. The top row are the scannerless images, and the bottom row are the corresponding real images. The SSIM is computed with a window size of 11.

in human subject experiments, where different breathing patterns were induced and separately evaluated. Evaluation was performed using the SSIM and by comparing the extracted motion from the real and scannerless images using the MAE and R^2 metrics.

It was hypothesized that the use of a combination of all surrogates would improve the model's capability to differentiate between the different breathing patterns and that its overall performance would increase. However, it was found that the model, trained on a combination of all surrogates, performed slightly worse than the model trained on just the external breathing motion surrogate. This is due to the poor quality of the internal and airflow surrogate signals. Especially, the airflow was very noisy and in some instances would only result in a reliable waveform during deep breathing. This is mainly due to the heat camera being positioned outside the magnetic field of the scanner, making the range from the subject's face mask fairly high. There was also trending present in the airflow surrogate, because the mirror and surrounding scanner would slowly heat up during image acquisition. The internal motion surrogate did result in high quality breathing waveforms, but the linear detrending was not sufficient in instances where the signal included a nonlinear trend. A better detrending method must therefore be investigated to improve the internal breathing motion waveform, or the feature extraction could be performed by a deep learning method, similar to [14].

The internal breathing and airflow surrogates being of low quality is confirmed by the bad performance of the models that are exclusively trained on one of these surrogates. The high amount of noise in the airflow surrogate caused multiple temperature observations to correspond with the same liver position, and thus roughly the same image. This caused the model to collapse in most instances and always generate the same low quality image. The trending in the internal breathing surrogate signal caused the model to be entirely unfamiliar with the observations in the testing data, causing inaccurate movement in generated images, or mode collapse. The low quality of these surrogate signals also slightly deteriorates the results of the combined model. This is especially clear during shallow breathing, as can be confirmed by comparing the liver border position for subjects A and C in Fig. 12 and Fig. 13. The model trained on all surrogate signals generates images with minimal movement due to the noisy airflow surrogate, while the model trained exclusively on the external motion surrogate signal produces fairly accurate movement.

Another challenge was the manual synchronization between the surrogate signals and the real MR-images. This was mainly the case when the peaks of the MR and US breathing waveforms were rounded instead of sharp, because this made it challenging to find two perfectly matching positions. This problem can be clearly seen in the extracted breathing motion from subject C in Fig. 12. The breathing motion from the scannerless images seems to be slightly delayed. This is also the cause of the higher MAE as a percentage of the PTT during shallow breathing, because the error caused by the delay is proportionally larger when the PTT is low. We believe that bad synchronization also caused mode collapse in subjects D and E. The waveforms from the internal breathing surrogate from these subjects were wide and flat. This made it difficult to pinpoint the exact peak of the breathing cycle, leading to synchronization errors of up to half a cycle. This means that bad synchronization can be a result of low quality waveforms and that improving the quality of the waveforms should make synchronization more accurate. It is however still recommended to investigate a system that can detect pulses from the MR-scanner to achieve perfect synchronization that is independent of the waveform's quality and not prone to human error. This has previously been done in [13], [14].

It is also clear that the performance during any of the breath holds is significantly lower than during breathing. This could have several reasons. Firstly, the fully inhaled and fully exhaled breath holds are generally at extreme liver positions that are not reached during any of the breathing patterns, meaning that there is no training data for the liver position during these breath holds. Another possible reason is that a window of surrogate data is used as input for the generator, meaning that the gradient is never entirely zero in the training data, that only consists of breathing data. When a window of surrogate data during a breath hold is used as input, its gradient is zero everywhere, which the model has not seen before. A possible improvement for these problems is to include breath holds in the training data or to decrease the window size of the input vectors, so that an entirely flat input window is more likely reached in the extremes of the breathing data.

Finally, the SSIM is relatively low, being only 51.42% and 50.75% on average for the combined and external motion model during breathing, respectively. This is substantially lower than the results from Veenstra [16]. Fig. 11 showed that the SSIM quickly increases as the window size is increased, indicating that the overall structure of the images are accurately modeled but are lacking in details. This is confirmed by visually inspecting the results. We generally saw that boundaries like the liver border and abdominal wall are slightly blurred and not well defined. We believe that adding complexity to the model, by increasing its number of parameters could improve the quality and sharpness of the scannerless images, improving the SSIM. Also, the vessels in the lungs were mostly not well defined. The heartbeat causes slight changes in the brightness and shape of these vessels, whereas they remain constant in the scannerless images. This is because the generator has no information about the heartbeat. Adding a surrogate signal that is highly correlated to the heartbeat could improve the modelling of the blood vessels, which should increase the SSIM.

The average interference time for a single image of 3.6 ms when using a GPU and 7.9 ms when using a CPU suggests that the model can be used for real-time image generation. However, further testing is needed to confirm that the overhead of the waveform extraction is minimal enough to sustain real-time performance.

We believe that the short training times allow for a clinical scenario where the data collection and training of the model is done right before the actual procedure, where the scannerless MR-images can be used to give clinicians extra information on the current liver position while the patient is out-of-bore.

VIII. Limitations

The proposed solution is able to generate scannerless MR-images while taking into account the RIM. However, a limitation to the proposed solution is that RIM is not the only source of motion and deformation in the liver. Other sources of deformation might include the clinicians touching and moving tissue, insertion of a needle during ablation or a biopsy, and the heartbeat. These sources are not modeled and will therefore not be reflected in the scannerless images.

Another limitation is the location of the RGB-camera in regard to the position that is being tracked. When the subject is moved out-of-bore, after training the model, the position of the tracked point in the images changes, potentially causing the correlation between the surrogate and the MR-images to be entirely lost. Regularization techniques that make the input data independent on the position of the camera with respect to the point that it is tracking could be investigated to resolve this issue. Another solution could be to mount the camera on the bed where the subject is laying on, ensuring a fixed position of the camera, relative to the point that it is tracking.

IX. Future Work

The presented work serves as a strong foundation for further developments aimed at enhancing its clinical usability. To improve the practical utility of this approach, several key extensions are suggested.

The first suggestion is to expand the single plane image generation of the presented work, to the generation of the entire 3D volume. This would give the clinician significantly more information about the tumor's surrounding tissue and could help deciding the best needle insertion angle, potentially increasing the accuracy and safety of procedures.

Another suggestion is the addition of other sources of motion in the model by using surrogate signals that are highly correlated to the target motion and adding them as conditioning of the generator. An example could be to incorporate tissue and needle properties so that the tool-tissue interaction can be accounted for in the scannerless images. This could also make it possible to show the current location of the needle in the scannerless images. Another example is to incorporate heartbeat monitoring data in the model, to account for slight movements caused by the heartbeat. These enhancements would not only improve the realism of the scannerless images but also provide clinicians with critical information that could aid in decision-making during surgery.

X. Conclusion

This research proposed the use of a progressively growing GAN for scannerless MR-image generation using multiple surrogate signals. The following main research question was addressed: How can a generative adversarial network be utilized to generate real-time scannerless MR-images using a number of surrogate signals? Data was collected in human subject experiments where surrogate data and MR-images were simultaneously collected. The following surrogate signals were used: an ultrasound transducer to capture the internal breathing motion, optical tracking of visual markers to capture the external breathing motion, and a heat camera to capture the breathing airflow. This work compared the use of a combination of these surrogate signals with using each of them separately.

In response to the first sub-question—How can information about inter- and intra-variable respiratory patterns be extracted from the surrogate signals and combined as input for the GAN?—it was found that extracting the external breathing information was the most reliable. The internal breathing information included a linear trend that was hard to remove from the data and the airflow data was very noisy, making it challenging to extract reliable breathing information.

The findings from the first sub-questions are confirmed in the second sub-question: "How much do the scannerless images resemble the real images in different respiratory patterns, and what is each surrogate signal's influence on this result?" It was found that the model trained exclusively on the external breathing surrogate resulted in scannerless images with the most accurate breathing motion. However, the scannerless images lack in detail but the overall structure is captured fairly accurately. The results did show that it is challenging to generate scannerless MR-images during breath holds when there are not included in the training dataset. The combined model showed slightly worse results and the internal breathing model and airflow model were significantly worse across all metrics and breathing patterns. This is due to the lower quality breathing waveforms that resulted from these surrogate signals.

It was found that tracking the external breathing motion was a reliable surrogate signal that was additionally cost efficient and easy to setup. Its limitation is that the correlation between MR-images and the tracked markers are lost when the subject is moved, meaning that regularization techniques have to be investigated to make it usable in a real world scenario. We were unable to get accurate results using the ultrasound transducer, but it could be greatly improved when using a more adequate detrending method, or by using a deep learning feature extractor. Its advantage is that it is attached to the subject, meaning that the correlation is preserved when moving the subject. The heat camera seems to be an unsuitable surrogate signal when using MR as an imaging modality, since it has to be positioned far away from the subject.

For further research, it is recommended to reconsider the chosen GAN architecture, or add complexity to the model used in the proposed approach to improve the quality and details of the scannerless images. Another suggestion is to generate the entire 3D volume of the abdominal area instead of a single plane, to provide more information on the surrounding tissue of a tumor. Additionally, it should be investigated how other sources of motion could be incorporated in the model. These sources could include the heartbeat, needle insertion, and the clinician touching tissue. We believe that this research demonstrated the potential of using a GAN for scannerless MRI generation and that together with the proposed directions of future work, percutaneous procedures in the liver can become more precise and easier to perform.

Acknowledgements

I would like to express my gratitude to my supervisors, Momen Abayazid and Ana Cordon, for their invaluable guidance, insightful feedback, and encouragement throughout this research. Next, I would like to thank everyone who participated in the experiments for their willingness to help and sacrificing their time to this study. Lastly, I want to extend my appreciation to my family, friends, and especially to my girlfriend, Faye Loos, for their support, encouragement, and patience throughout my master's program.

Appendix A
Generator and Discriminator architecture

Table VI introduces the structure of the surrogate processor network. This network takes either one of the surrogate signals, or a concatenation of the surrogate signals. Therefore, the shape of the input is denoted as N. The network consists of

TABLE VI: Architecture of the surrogate processor

Surrogate Processor	Act.	Output shape	Params
Input	-	$1\times N$	-
Fully connected	ReLU	1×64	$N\cdot 64+64$
Fully connected	ReLU	1×32	2080
Fully connected	ReLU	1×32	1056
Total			64N + 3200

TABLE VII: Architecture of the generator

Generator	Act.	Output shape	Params
Latent vector	-	256×1×1	-
Conv 4×4	ReLU	256×4×4	1,048,832
Conv 3×3	ReLU	256×4×4	590,080
Upsample	-	256×8×8	-
Conv 3×3	ReLU	256×8×8	590,080
Conv 3×3	ReLU	128×8×8	295,040
Upsample	-	128×16×16	-
Conv 3×3	ReLU	128×16×16	147,584
Conv 3×3	ReLU	64×16×16	73,792
Upsample	-	64×32×32	-
Conv 3×3	ReLU	64×32×32	36,928
Conv 3×3	ReLU	32×32×32	18,464
Upsample	-	32×64×64	-
Conv 3×3	ReLU	32×64×64	9,248
Conv 3×3	ReLU	16×64×164	4,624
Upsample	-	16×128×128	-
Conv 3×3	ReLU	16×128×128	2,320
Conv 3×3	ReLU	8×128×128	1,160
Conv 1×1	Tanh	1×128×128	9
Total			2,818,161

an input layer, one hidden layer, and an output layer with the ReLU activation function after each layer.

Table VII introduces the structure of the generator. The generator consists of six blocks that each returns a scannerless image in a different resolution. Each block first upsamples the output of the previous block, using nearest neighbor interpolation, and then performs two convolutions with a kernel size of 3×3. The ReLU activation function is applied after each convolution. The final convolution with a kernel size of 1×1 is to transfer the eight channel output of the previous convolution to a single channel grayscale image. The final Tanh activation function is performed so that the pixel range of the output image is $[-1,1]$.

Table VIII introduces the structure of the discriminator. It is an exact mirror of the generator, again with six blocks where two convolutions are performed with a kernel size of 3×3. Each block takes an image with a different resolution and it downsamples the image at the end using average pooling. The LeakyReLU activation function with a slope of 0.2 is applied after each convolution. In the final block, minibatch standard deviation was applied to increase the variation, which is specified in [19]. The first convolution with the 1×1 kernel size is to transform the single channel input image to the amount of channels that the is required by the block. The final part is the fully connected layer with a linear activation function that outputs a single value that is used to compute the Wasserstein distance.

TABLE VIII: Architecture of the discriminator

Discriminator	Act.	Output shape	Params
Input image	-	1×128×128	-
Conv 1×1	LReLU	8×128×128	16
Conv 3×3	LReLU	8×128×128	584
Conv 3×3	LReLU	16×128×128	1,168
Downsample	-	16×64×64	-
Conv 3×3	LReLU	16×64×64	2,320
Conv 3×3	LReLU	32×64×64	4,640
Downsample	-	32×32×32	-
Conv 3×3	LReLU	32×32×32	9,248
Conv 3×3	LReLU	64×32×32	18,496
Downsample	-	64×16×16	-
Conv 3×3	LReLU	64×16×16	36,928
Conv 3×3	LReLU	128×16×16	73,856
Downsample	-	128×8×8	-
Conv 3×3	LReLU	128×8×8	147,584
Conv 3×3	LReLU	256×8×8	295,168
Downsample	-	256×4×4	-
Minibatch std		257×4×4	-
Conv 3×3	LReLU	256×4×4	592,384
Conv 4×4	LReLU	256×1×1	1,048,832
Fully connected	Linear	1×1×1	257
Total			2,231,481

References

[1] H. Rumgay, M. Arnold, J. Ferlay, O. Lesi, C. J. Cabasag, J. Vignat, M. Laversanne, K. A. McGlynn, and I. Soerjomataram, "Global burden of primary liver cancer in 2020 and predictions to 2040," *Journal of Hepatology*, vol. 77, no. 6, pp. 1598–1606, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168827822030227

[2] C.-Y. Liu, K.-F. Chen, and P.-J. Chen, "Treatment of liver cancer," *Cold Spring Harb Perspect Med*, vol. 5, no. 9, p. a021535, Jul. 2015.

[3] S. A. Gonzalez and E. B. Keeffe, "Diagnosis of hepatocellular carcinoma: Role of tumor markers and liver biopsy," *Clinics in Liver Disease*, vol. 15, no. 2, pp. 297–306, 2011, diagnosis and Therapy of Hepatocellular Carcinoma: Status Quo and a Glimpse at the Future. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1089326111000134

[4] J. M. Llovet, J. Fuster, and J. Bruix, "The barcelona approach: Diagnosis, staging, and treatment of hepatocellular carcinoma," *Liver Transplantation*, vol. 10, no. S2, pp. S115–S120, 2004. [Online]. Available: https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/lt.20034

[5] S. A. Wells, J. L. Hinshaw, M. G. Lubner, T. J. Ziemlewicz, C. L. Brace, and F. T. Lee, Jr, "Liver ablation: Best practice," *Radiol Clin North Am*, vol. 53, no. 5, pp. 933–971, Sep. 2015.

[6] S. Fahmi, "Respiratory motion estimation of the liver with abdominal motion as a surrogate : a supervised learning approach," August 2017. [Online]. Available: http://essay.utwente.nl/73643/

[7] M. Gargett, C. Haddad, A. Kneebone, J. T. Booth, and N. Hardcastle, "Clinical impact of removing respiratory motion during liver SABR," *Radiat Oncol*, vol. 14, no. 1, p. 93, Jun. 2019.

[8] C. Ozhasoglu and M. J. Murphy, "Issues in respiratory motion compensation during external-beam radiotherapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 52, no. 5, pp. 1389–1399, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360301601027894

[9] P. J. Keall, G. S. Mageras, J. M. Balter, R. S. Emery, K. M. Forster, S. B. Jiang, J. M. Kapatoes, D. A. Low, M. J. Murphy, B. R. Murray *et al.*, "The management of respiratory motion in radiation oncology report of aapm task group 76 a," *Medical physics*, vol. 33, no. 10, pp. 3874–3900, 2006.

[10] J. McClelland, D. Hawkes, T. Schaeffter, and A. King, "Respiratory motion models: A review," *Medical Image Analysis*, vol. 17, no. 1, pp. 19–42, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136184151200134X

[11] A. Avila and M. Abayazid, "Liver respiratory-induced motion estimation using abdominal surface displacement as a surrogate: robotic phantom and clinical validation with varied correspondence models," *International journal of computer assisted radiology and surgery*, 05 2024.

[12] M. Abayazid, T. Kato, S. Silverman, and N. Hata, "Using needle orientation sensing as surrogate signal for respiratory motion estimation in percutaneous interventions," *International journal of computer assisted radiology and surgery*, vol. 13, no. 1, pp. 125–133, Jan. 2018, springer deal.

[13] F. Preiswerk, M. Toews, C. Cheng, J. Chiou, C.-S. Mei, L. Schaefer, W. Hoge, B. Schwartz, L. Panych, and B. Madore, "Hybrid mri-ultrasound acquisitions, and scannerless real-time imaging," *Magnetic Resonance in Medicine*, vol. 78, 10 2016.

[14] F. Preiswerk, C.-C. Cheng, J. Luo, and B. Madore, *Synthesizing Dynamic MRI Using Long-Term Recurrent Convolutional Networks*. Springer International Publishing, 09 2018, pp. 89–97.

[15] K. Shokry, "Generating high frame rate mri images using a surrogate signal a supervised learning approach," September 2018. [Online]. Available: http://essay.utwente.nl/76692/

[16] G. Veenstra, "Generating high frame rate mr images using surrogate signals," August 2019.

[17] A. Giger, R. Sandkühler, C. Jud, G. Bauman, O. Bieri, R. Salomir, and P. C. Cattin, "Respiratory motion modelling using cgans," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*. Springer, 2018, pp. 81–88.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2018.

[19] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2018.

[20] M. S. Arshad and W. Beksi, "A progressive conditional generative adversarial network for generating dense and colored 3d point clouds," pp. 712–722, 11 2020.

[21] Z. Zhou, S. Jiang, Z. Yang, N. Zhou, S. Ma, and Y. Li, "A high-dimensional respiratory motion modeling method based on machine learning," *Expert Systems with Applications*, vol. 242, p. 122757, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423032591

[22] B. Madore, F. Preiswerk, J. S. Bredfeldt, S. Zong, and C.-C. Cheng, "Ultrasound-based sensors to monitor physiological motion," *Medical physics*, vol. 48, no. 7, pp. 3614–3622, 2021.

[23] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[24] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[25] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html

[26] C. Villani, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. [Online]. Available: https://books.google.nl/books?id=hV8o5R7_5tkC

[27] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," 2017.

[28] L. P. Panych and J. Tokuda, *Real-Time and Interactive MRI*. New York, NY: Springer New York, 2014, pp. 193–209. [Online]. Available: https://doi.org/10.1007/978-1-4614-7657-3_13

[29] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[30] F. Sohil, M. Sohail, and J. Shabbir, "An introduction to statistical learning with applications in r: by gareth james, daniela witten, trevor hastie, and robert tibshirani, new york, springer science and business media, 2013, $41.98, eisbn: 978-1-4614-7137-7," *Statistical Theory and Related Fields*, vol. 6, pp. 1–1, 09 2021.

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.