Physics in Medicine & Biology

Deep learning-based image reconstruction and motion estimation from undersampled radial k-space for real-time MRI-guided radiotherapy

To cite this article before publication: Maarten Lennart Terpstra *et al* 2020 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/ab9358

Deep learning-based image reconstruction and motion estimation from undersampled radial k-space for real-time MRI-guided radiotherapy

Maarten L. Terpstra[1,2], Matteo Maspero[1,2], Federico d'Agata[1,2,3], Bjorn Stemkens[1,2], Martijn P.W. Intven[1], Jan J.W. Lagendijk[1], Cornelis A.T. van den Berg[1,2], and Rob H.N. Tijssen[1,4]

[1] Department of Radiotherapy, University Medical Center Utrecht, The Netherlands
[2] Computational Imaging Group for MR Diagnostics & Therapy, Center for Image Sciences, University Medical Center Utrecht, the Netherlands
[3] Department of Neurosciences, University of Turin, Turin, Italy
[4] Department of Radiation Oncology, Catharina Hospital, Eindhoven, the Netherlands

E-mail: m.l.terpstra-5@umcutrecht.nl

Abstract.

Purpose: To enable magnetic resonance imaging (MRI)-guided radiotherapy with real-time adaptation, motion must be quickly estimated with low latency. The motion estimate is used to adapt the radiation beam to the current anatomy, yielding a more conformal dose distribution. As the MR acquisition is the largest component of latency, deep learning (DL) may reduce the total latency by enabling much higher undersampling factors compared to conventional reconstruction and motion estimation methods. The benefit of DL on image reconstruction and motion estimation was investigated for obtaining accurate deformation vector fields (DVFs) with high temporal resolution and minimal latency.

Methods: 2D cine MRI acquired at 1.5T from 135 abdominal cancer patients were retrospectively included in this study. Undersampled radial golden angle acquisitions were retrospectively simulated. DVFs were computed using different combinations of conventional- and DL-based methods for image reconstruction and motion estimation, allowing a comparison of four approaches to achieve real-time motion estimation. The four approaches were evaluated based on the end-point-error and root-mean-square error compared to a ground-truth optical flow estimate on fully-sampled images, the structural similarity (SSIM) after registration and time necessary to acquire k-space, reconstruct an image and estimate motion.

Results: The lowest DVF error and highest SSIM were obtained using conventional methods up to $\mathcal{R} \leq 10$. For undersampling factors $\mathcal{R} > 10$, the lowest DVF error and highest SSIM were obtained using conventional image reconstruction and DL-based motion estimation. We have found that, with this combination, accurate DVFs can be obtained up to $\mathcal{R} = 25$ with an average root-mean-square error up to 1 millimeter and an SSIM greater than 0.8 after registration, taking 60 milliseconds.

Conclusion: High-quality 2D DVFs from highly undersampled k-space can be

obtained with a high temporal resolution with conventional image reconstruction and a deep learning-based motion estimation approach for real-time adaptive MRI-guided radiotherapy.

Submitted to: *Phys. Med. Biol.*
Keywords: Deep Learning, MRI, Reconstruction, Undersampling, Motion Estimation, MR-Linac, Radiotherapy, Real-time

1. Introduction

Magnetic resonance imaging-guided radiotherapy (MRIgRT) is increasingly adopted in clinical practice. Hybrid MRI scanners with an integrated linear accelerator (MR-linac) have shown to be very efficient in dealing with inter-fraction anatomical changes by employing online re-planning prior to each treatment session (Mutic and Dempsey, 2014; Winkel et al., 2019).

The future promise of hybrid MR-linac systems is to not only account for inter-fraction motion but also adapt the radiation delivery in real-time during treatment to accommodate for respiration or cardiac-induced intra-fraction motion (Keall et al., 2019; Glitzner et al., 2015; Fast et al., 2017; Kontaxis et al., 2017; Dietz et al., 2019), peristaltic motion and tissue deformation, e.g. due to bladder filling or passing air bubbles.

Real-time adaptive radiotherapy requires imaging with extremely high temporal resolution as well as a very low total latency (i.e., the time between an event and response) of the MR-linac feedback chain (Keall et al., 2006). The most significant source of latency in the MR-linac feedback chain is MR image acquisition (Borman et al., 2018). If acquisitions could be significantly undersampled, motion could be estimated with minimal latency. Although dense array radio-lucent receiver coils improve the acquisition speed of MR-linac systems by use of parallel imaging (Zijlema et al., 2019), most motion quantification techniques are image-based and rely on high-quality images, which limits the maximum acceleration factors achievable with parallel imaging (Wiesinger et al., 2004). Regularized reconstruction methods like compressed sensing (Lustig et al., 2007) may achieve even higher acceleration factors, but the iterative nature of compressed sensing reconstruction algorithms make it unsuitable for real-time applications.

Recently, deep learning (DL) has become a popular technique in many scientific fields due to its high-quality results and speed. The use of neural networks to generate a hierarchical representation of the input data to achieve high task-specific performance without the need of hand-engineered features has proven extremely powerful for imaging applications (Litjens et al., 2017; Meyer et al., 2018; Sahiner et al., 2019). In computer vision, various DL methods have been developed that outperform traditional motion

estimation algorithms (Ranjan and Black, 2017; Dosovitskiy et al., 2015; Ilg et al., 2017), while for MRI several DL methods have been proposed to replace the computationally expensive compressed sensing reconstructions (Schlemper et al., 2018; Hammernik et al., 2018; Lønning et al., 2019).

In this paper, we investigate the performance of DL for image reconstruction and motion quantification on highly undersampled golden-angle (GA) radial acquisitions for real-time MRIgRT with the goal of providing accurate motion quantification with minimal latency. We hypothesize that the benign undersampling artifacts in GA radial MRI in combination with DL image reconstruction provides high acceleration factors with image quality on par with CS reconstruction but at a fraction of the computation time. The addition of a DL-based motion quantification approach is believed to relax the requirements for high-quality images, potentially allowing even greater image acceleration factors.

In this work, we investigate a two-step process in which retrospectively undersampled dynamic GA radial data are reconstructed by classical methods or using DL models. With this approach, we assess the individual and the combined performance of DL-based image reconstruction and processing on computation time and motion estimation accuracy for acceleration factors of up to 50.

2. Materials & Methods

The study design is illustrated in Figure 1. Image reconstruction from undersampled dynamic GA radial k-space was performed with either a classical non-uniform fast Fourier transform (NUFFT) (Fessler and Sutton, 2003) or with dAUTOMAP (Schlemper et al., 2019), a convolutional neural network designed for image reconstruction. Subsequently, motion is estimated on the reconstructed images via a classical optical flow (OF) based motion estimation algorithm, or a modified version of SPyNET, a multi-resolution layered deep neural network that computes deformation vector fields (DVFs) at multiple resolutions, similar to OF (Ranjan and Black, 2017). This allowed us to compare four approaches using varying degrees of DL to estimate motion from undersampled dynamic GA radial k-space.

2.1. Patient data collection

Patients diagnosed with cancer in the abdomen undergoing radiotherapy simulation at our department between June 2015 and December 2019 were included in this study when sagittal cine MRI were acquired. In total, 135 patients were included, of whom 83 were male and 52 were female and were diagnosed with tumors to the abdomen (7), liver (40), kidneys (62) and pancreas (26). The patients were between 37 and 89 years old with a mean age of 67 ± 11 years old. Two-dimensional (2D) Cartesian balanced steady-state free precession (bSSFP) cine MRIs were acquired on a 1.5T MRI scanner (Ingenia MR-RT, Philips, Best, the Netherlands). Table 1 lists the acquisition parameters. The total

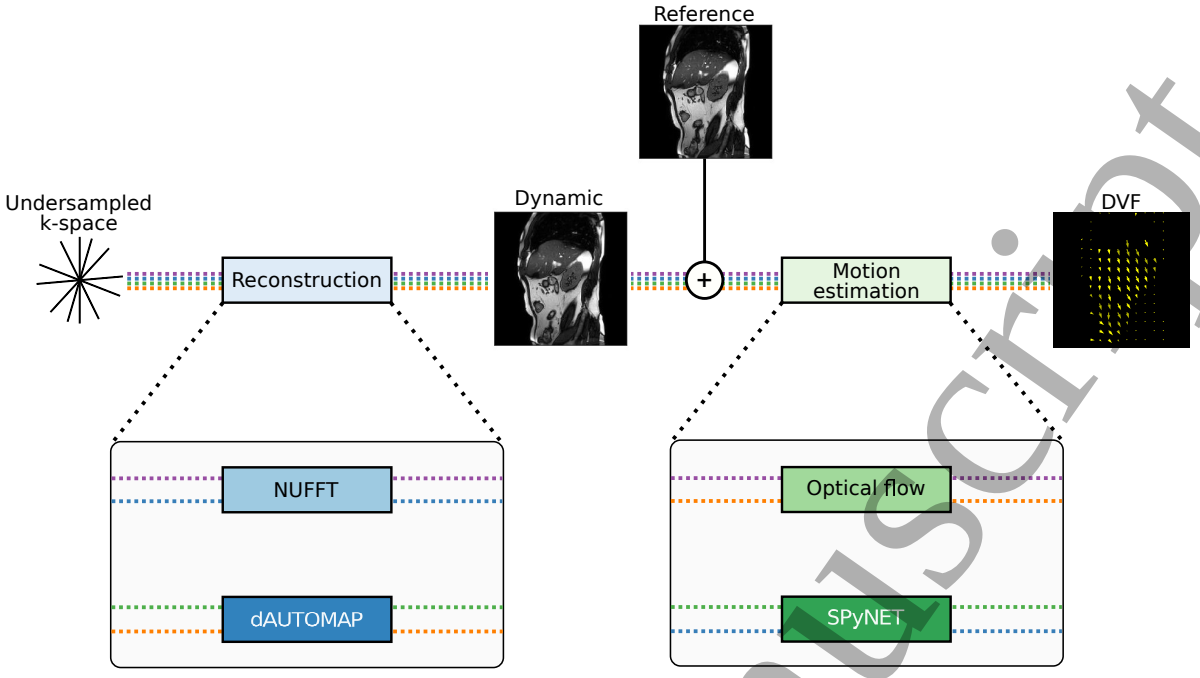

Figure 1: Schematic overview of the study design. Generation of undersampled k-space (top) and the DVFs from fully-sampled k-space (ground-truth) and undersampled k-space (bottom). Reconstruction can happen with a NUFFT or a DL-based image reconstruction model. Motion estimation with the reference image happens with optical flow or a DL-based motion estimation approach.

acquisition time was between 25 s and 2.5 min, according to the number of dynamics acquired per scan, which varied between 50 and 300. Patients were scanned on a flat tabletop in the supine position using a 16-channel anterior and a 12-channel posterior phased-array coil. Two in-house built coil bridges supported the anterior coil to avoid skin contour deformation and not to affect natural motion. In total, 31750 magnitude-only dynamics were collected from 200 cine MRIs, as for some patients the cine data were acquired multiple times. Of these 200 cine MRIs, 126 were scanned after contrast agent injection.

For 30 of the 135 patients, 42 coronal cine MRIs were also acquired. Coronal cine MRIs were used for model validation. The scan parameters of these cine MRIs are also detailed in Table 1.

2.2. Data preparation

The signal intensity over all dynamics was linearly rescaled to an output range of [0, 1], clipping to the 99$^{\text{th}}$ percentile of intensity values of the dynamics in a cine MRI. Complex k-space was obtained by adding simulated phase to the magnitude-only images and computing the non-uniform Fourier transform (NUFFT) (Fessler and Sutton, 2003) using PyNUFFT version 2019.1.1 (Lin, 2018) with an undersampled GA radial readout trajectory. The simulated phase was generated per dynamic, as suggested by Zhu et al.

Table 1: Scan parameters for sagittal and coronal 2D Cartesian balanced steady-state free precession MRI used in this work.

Parameter	**Sagittal**	**Coronal**
TE (ms)	1.4	1.4
TR (ms)	2.8	2.7
Flip angle	50°	50°
Resolution (mm^2)	1.4×1.4	2.0×2.0
FOV (mm^2)	320×320	450×450
Reconstruction resolution (px^2)	224×224	224×224
Slice thickness (mm)	7	7
Readout direction	FH	FH
Bandwidth (Hz/px)	724 - 2034	1431 - 2034
Temporal resolution (ms)	500-570	500-570
Number of dynamics	50-300	100-300

(2018), i.e. by generating two two-dimensional sinusoids with a randomly-chosen spatial frequency between 0.05 Hz and 0.25 Hz and rotating these sinusoids separately with a random angle around the origin. These sinusoids were added together and the amplitude normalized to $[-\pi, \pi]$ such that the intensity represents phase values. K-space was density-compensated with a Ram-Lak filter and gridded to a Cartesian grid.

To ensure that representative noise was present in the retrospectively undersampled k-space, additional Gaussian noise $X \sim \mathcal{N}(0, \epsilon \cdot |k_0|)$ was added separately to the real and imaginary channels, where ϵ was randomly chosen between $[3 \cdot 10^{-3}, 5 \cdot 10^{-3})$. The range for ϵ was determined from separate noise scans as the magnitude of the noise divided by the magnitude of the DC component.

The undersampling factor $\mathcal{R}$ was determined by dividing the number of spokes required for a Nyquist-sampled radial acquisition at the reconstruction resolution by the undersampling factor, i.e. $\lceil 224 \cdot \pi/2 \rceil \cdot \mathcal{R}^{-1} = 352 \cdot \mathcal{R}^{-1}$. The induced latency of this acquisition scheme is half of the acquisition time, i.e. $352 \cdot \mathcal{R}^{-1} \cdot \mathrm{TR}/2$ (Borman et al., 2018). Data was prepared for the undersampling factors $\mathcal{R} = 1$, 5, 10, 16, 20, 25, 30, 40, and 50.

2.3. Image reconstruction

The generated k-space of each dynamic was reconstructed with a conventional method and a DL-based approach.

2.3.1. Conventional Non-Cartesian k-space was reconstructed with a NUFFT adjoint reconstruction, obtaining a fast reconstruction at the cost of undersampling artifacts compared to an iterative reconstruction algorithm.

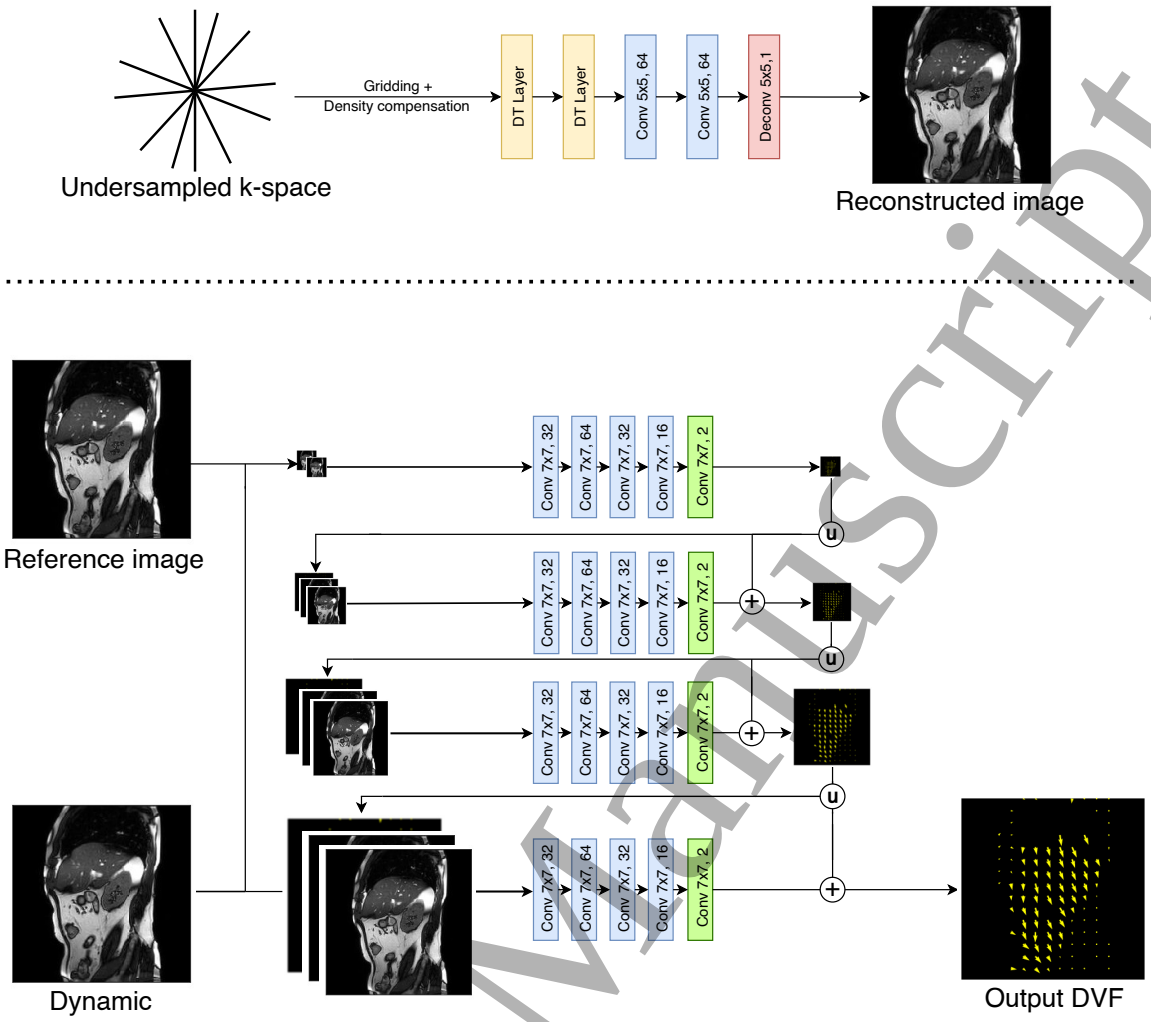

Figure 2: Schematic of the image reconstruction and motion estimation models. The dAUTOMAP model (top) reconstructs the re-gridded and density-compensated undersampled k-space to an image. SPyNET (bottom) is a multi-resolution approach that estimates a DVF between a reference image and dynamic using multiple CNNs. Blue and green layers are two-dimensional convolution layers with and without non-linear activation, respectively.

2.3.2. Deep learning For image reconstruction from undersampled k-space dAUTOMAP‡ (Schlemper et al., 2019) was trained on a GPU (Tesla P100, NVIDIA, Santa Clara, CA, USA). dAUTOMAP is a model that performs non-iterative reconstruction with low parameter count, which makes it suitable for real-time image reconstruction. As dAUTOMAP assumes that the k-space points lie on a Cartesian grid, the k-space was re-gridded and density-compensated, as illustrated in Figure 2 (top). The model was implemented in PyTorch 1.0.1 and had 913473 trainable parameters. dAUTOMAP was initialized using Xavier initialization (Glorot and Bengio, 2010) and trained using the

‡ Reference implementation as found on `https://github.com/js3611/dAUTOMAP`

Adam optimizer (Kingma and Ba, 2015) using $\beta_1 = 0.9, \beta_2 = 0.999$ and a learning rate of 10^{-3} with a batch size of 64 on an undersampled k-space with $\mathcal{R} = 10$ to minimize the mean-square-error (MSE) between reconstruction and target. After 50 epochs, the high-frequency error norm (HFEN) (Ravishankar and Bresler, 2011) was added to the loss function as it was found to improve performance. dAUTOMAP was trained until validation loss converged. $\mathcal{R} = 10$ was chosen as the undersampling factor for training as a balance between a fast acquisition and image quality, as training with higher undersampling factors became unstable. The learning rate was halved if the validation error plateaued, i.e. if the validation error has not improved with at least 10^{-8} in the last ten epochs. dAUTOMAP was trained on 119 cine MRIs from 81 patients comprising 60% of all sagittal dynamics. The hyper-parameters were validated on 38 cine MRIs from 26 patients comprising 20% of all sagittal dynamics. The final model was tested on 43 cine MRIs from 28 patients comprising 20% of all sagittal dynamics.

2.4. Motion estimation

For every sagittal cine MRI, a reference image was chosen by randomly selecting a dynamic after ensuring that the dynamic was acquired in the steady-state. This was ensured by excluding the first 30 dynamics of the cine MRI from the selection of reference images. Then, DVFs were computed between every dynamic and the reference image.

Five reference images were randomly selected per cine MRI as data augmentation strategy and to ensure that the reference images were not only on an "extreme" point of the respiratory phase, e.g. inspiration or expiration.

This yielded a total of 130475 DVFs for training and validation and 28275 DVFs for testing.

2.4.1. Conventional DVFs were computed using optical flow (Horn and Schunck, 1981; Zachiu et al., 2015a,b).

Optical flow is a registration algorithm that assumes the DVF to be smooth and the brightness of the images is preserved over time. Optical flow estimates DVFs by minimizing the energy function given in Equation 1:

$$E = \iint_{\Omega} |I_x u + I_y v + I_t| + \beta^2 \left(||\nabla u||_2^2 + ||\nabla v||_2^2 \right) \mathrm{d}x\mathrm{d}y \tag{1}$$

where $\Omega \subseteq \mathbb{R}^2$ is the image domain, u and v are components of the DVF, I_x, I_y, I_t are the spatial and temporal partial derivatives of the images, respectively, and β is the regularization parameter enforcing smoothness.

Optical flow refines the motion estimate through iteration and estimating motion at multiple resolution levels in a pyramid approach in order to resolve large displacements.

In a preliminary study that is presented in Appendix A, we compared an implementation of optical flow and Elastix (Klein et al., 2010) to assess the registration performance on our dataset.

As a result of this preliminary study, we opted to use optical flow as implemented with RealTITracker (Zachiu et al., 2015a,b) in this work. In particular, ground-truth DVFs were computed on the fully-sampled dynamics by computing optical flow between every dynamic/reference image pair with RealTITracker with $\beta = 0.6$.

2.4.2. Deep learning For motion estimation, the convolutional neural network called SPyNET (Ranjan and Black, 2017) was trained on a GPU (Tesla P100, NVIDIA, Santa Clara, CA, USA). SPyNET is a multi-resolution pyramid approach. At every resolution level in the pyramid, a small CNN of 233778 parameters is employed to estimate motion from the input images together with an upsampled motion estimate from the previous pyramid level. The motion estimation approach is illustrated in Figure 2 (bottom). The model was implemented in PyTorch 1.0.1 and was serially trained with four pyramid levels, for a total of 935112 trainable parameters. The image pyramid had an image size of 224×224 pixels at the highest resolution level down to 28×28 pixels at the lowest resolution level. SPyNET was trained separately on pairs of images reconstructed with either a NUFFT or dAUTOMAP reconstruction with $\mathcal{R} = 10$ to learn the ground-truth optical flow DVFs by minimizing the end-point-error $\left(\text{EPE} = \sqrt{(u_{est} - u_{gt})^2 + (v_{est} - v_{gt})^2}\right)$. The model weights of all networks were initialized using Kaiming uniform initialization (He et al., 2015).

The effect of the warping operator as defined in the original implementation of SPyNET, which registers the images at lower resolution levels to resolve larger displacements, was evaluated and was found to be detrimental to the motion estimation quality and therefore omitted.

Data augmentation was performed on the images consistent with the ground-truth DVF by random horizontal and vertical flips and contrast jitter to prevent overfitting (Bloice et al., 2019). The EPE was minimized using the Adam optimizer $\beta_1 = 0.9, \beta_2 = 0.999$ with a learning rate of $5 \cdot 10^{-4}$ until convergence of the validation loss. The batch size was limited by the available GPU memory and was 1024 for the lowest resolution level, and 32 for the highest resolution level.

Every SPyNET level was trained, tested, and validated on the same data partition as dAUTOMAP. That is, 119 cine MRIs from 81 patients comprising 60% of all sagittal dynamics were used for training. The hyper-parameters were validated on 38 cine MRIs from 26 patients comprising 20% of all sagittal dynamics. The final model was tested on 43 cine MRIs from 28 patients comprising 20% of all sagittal dynamics.

2.5. Experiment setup

As image reconstruction and motion estimation can be computed with conventional or DL-based methods, we investigated four different combinations to obtain DVFs from k-space:

- Using NUFFT reconstruction and optical flow motion estimation (NUFFT/OF);
- Using NUFFT reconstruction and SPyNET motion estimation (NUFFT/SPyNET);

- Using dAUTOMAP reconstruction and optical flow motion estimation (dAUTOMAP/OF);
- Using dAUTOMAP reconstruction and SPyNET motion estimation (dAUTOMAP/SPyNET).

As the goal of these methods is to estimate motion from undersampled k-space, quality is defined solely by the correctness of the DVF. The four approaches were evaluated using the following criteria:

Registration performance The image similarity after registration of fully-sampled dynamics using a DVF estimated on undersampled images was evaluated over the whole image. This was quantified by the structural similarity (SSIM) (Wang et al., 2004) over the whole image. In particular, the mean ($\pm$ std) of the SSIM after registration was computed for 100 dynamic/reference image pairs of each cine MRI for every approach. In total, a sample of 2975 dynamic/reference pairs were considered.

DVF quality The quality of the DVF was measured by the mean absolute displacement error, as well as the root-mean-square error (RMSE) compared to the ground truth in a region of interest (ROI) that was manually generated to include relevant structures, e.g. liver veins, kidney structures or tumors. The ROIs of all patients in the test set are presented in Appendix B. The root-mean-square error of displacement within the ROIs was considered as well. Bland-Altman plots (Altman and Bland, 1983) of the mean absolute displacement error were calculated to compare the average DVF magnitude within an ROI to the ground-truth optical flow. These plots reveal the bias of a model for undersampled motion estimation in the generated DVFs, computing statistical error bounds. The statistical significance was estimated using the Wilcoxon signed-rank test.

Time The time necessary to estimate motion, including MR acquisition, was reported. For a fair comparison of the different approaches, only GPU timings were considered. Given that RealTITracker, the optical flow implementation that we adopted, is available only for CPUs, we obtained the timing of conventional motion estimation using a CUDA implementation of optical flow that is part of the OpenCV library §. Note that such implementation uses a different algorithm (Farnebäck, 2003) than the optical flow implementation used to generate ground-truth data.

All the metrics were computed on the test set, consisting of 28275 sagittal image pairs as well as 27900 coronal image pairs, for undersampling factors $\mathcal{R}$ = 1, 5, 10, 16, 20, 25, 30, 40, and 50 without retraining of the DL models, which were trained on $\mathcal{R} = 10$.

§ `https://github.com/NeerajGulia/python-opencv-cuda`

3. Results

dAUTOMAP was trained on $\mathcal{R} = 10$ for 300 epochs in approximately six hours. After training, inference of the model to reconstruct a dynamic from gridded k-space was performed in 5 ms, making it as fast as NUFFT adjoint reconstruction. Examples of NUFFT and dAUTOMAP reconstructions are shown in Figure 3d and Figure 3g, respectively. It can be observed that NUFFT reconstructions at $\mathcal{R} = 20$ suffer from considerable streaking artifacts and dAUTOMAP reconstructions are overly smoothed with intensity patches, as highlighted by the red arrows. Every SPyNET level was trained $\mathcal{R} = 10$ for 12 hours until the validation error converged which took between 200 and 1000 epochs, depending on the resolution level. After training, inference of the four-level pyramid including resizing the input images and upsampling the intermediate DVFs was performed in 15 ms, which is slower than a GPU optical flow implementation that estimates motion in 5 ms. Example DVFs estimated by SPyNET are shown in Figure 3f and Figure 3i, on NUFFT and dAUTOMAP reconstructions, respectively. Example DVFs estimated by optical flow are shown in Figure 3e and Figure 3h, on NUFFT and dAUTOMAP reconstructions, respectively. In the supplementary material, an animation of Figure 3 is reported. It can be observed that optical flow DVFs in the liver are comparable to the ground-truth, but in this case SPyNET is able to improve the motion estimate in the spine, which seems more physiologically plausible than for optical flow.

3.1. Registration performance

Using the DVFs as generated by the four proposed methods to register the fully-sampled dynamics, the SSIM quantifies the registration performance across the entire image. Figure 4 shows the SSIM as a function of the undersampling factor. DVFs generated by SPyNET lead to a significantly higher SSIM after registration compared to optical flow for $\mathcal{R} > 10$ (Wilcoxon, $p < 0.001$), even though the models were trained at $\mathcal{R} = 10$. At $\mathcal{R} = 30$ an average SSIM of 0.8 is achieved using NUFFT/SPyNET, whereas using NUFFT/optical flow results in an average SSIM of 0.72. Interestingly, Using SPyNET with NUFFT reconstruction shows a similar performance when evaluated on coronal acquisitions even though SPyNET was trained on sagittal dynamics, as presented in Figure 4. Using dAUTOMAP for image reconstruction results in a 5-25% drop in performance when registering coronal images compared to sagittal images depending on the undersampling factor.

3.2. DVF quality

The root-mean-square displacement error of the DVF generated with conventional methods compared to the ground-truth within an ROI on sagittal images significantly increases for acceleration factors $\mathcal{R} \geq 20$, as presented in Figure 5. For the NUFFT/SPyNET approach, the RMSE shows a slower rise as the undersampling factor

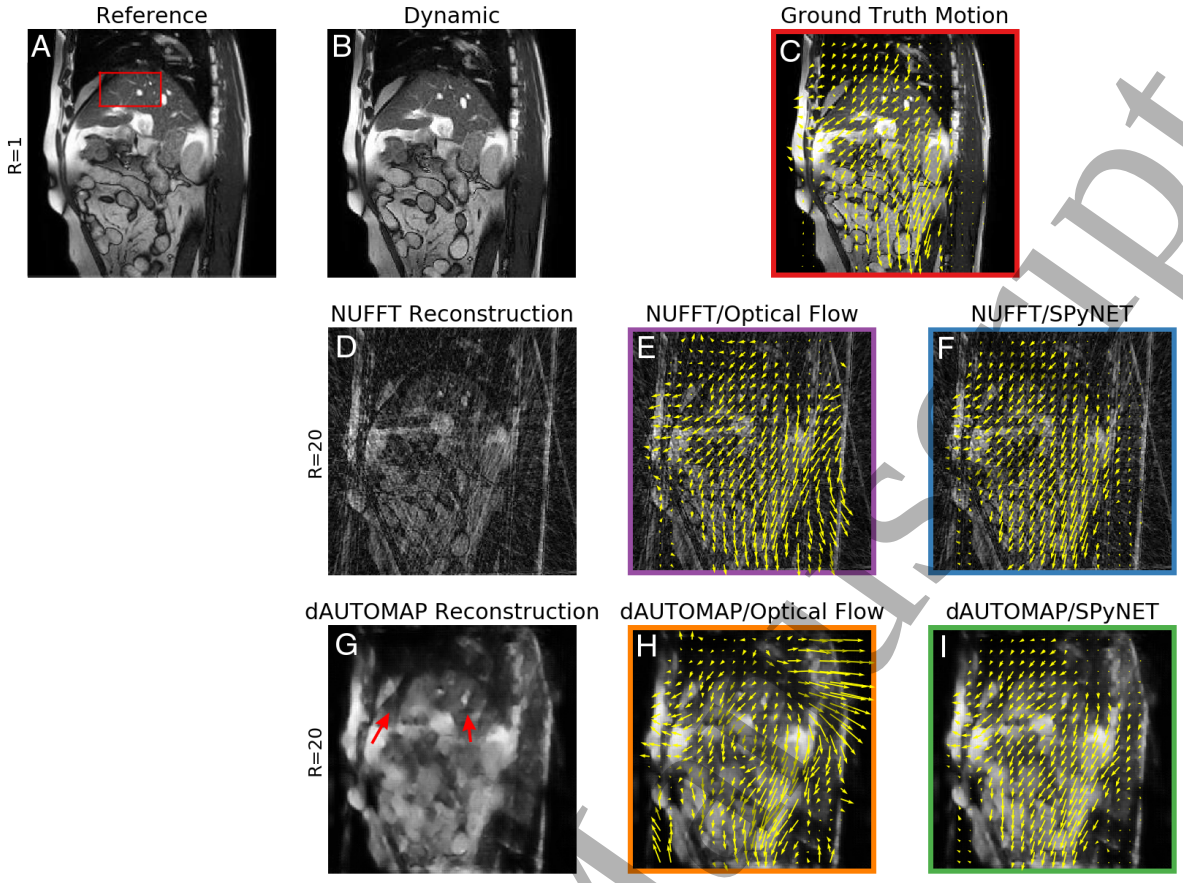

Figure 3: Example of a dynamic with image reconstruction and motion estimation. Figure 3a and Figure 3b show the fully-sampled sagittal reference image with region-of-interest in the red box and dynamic, respectively. The corresponding ground-truth DVF is shown in Figure 3c. Figure 3d shows the NUFFT adjoint reconstruction of the 20-fold retrospectively undersampled dynamic. The DVFs computed with the optical flow or SPyNET with adjoint reconstructions are shown in Figure 3e and f, respectively. Figure 3g, h, and i show the same as Figure 3d, e, and f, respectively but using dAUTOMAP for image reconstruction instead of a NUFFT adjoint. The arrows in Figure 3g indicate pseudo-random intensity patches introduced by dAUTOMAP.

increases, indicating robustness to undersampling artifacts. For NUFFT/SPyNET the root-mean-square displacement is lowest among all approaches at high undersampling factors ($\mathcal{R} \geq 20$) and remains within 1 mm with a narrower standard deviation, even for $\mathcal{R} = 30$.

Figure 6 reports Bland-Altman plots of the mean absolute displacement error within an ROI compared to the ground-truth on sagittal images. At $\mathcal{R} = 10$, there is no clear improvement of using DL rather than conventional methods. The mean difference is zero for the fully conventional method and has standard deviations within 0.95 mm, compared to a bias of -0.28 mm and a standard deviation up to 1.6 mm for dAUTOMAP/SPyNET. However, at $\mathcal{R} = 25$ the smallest error is obtained when using

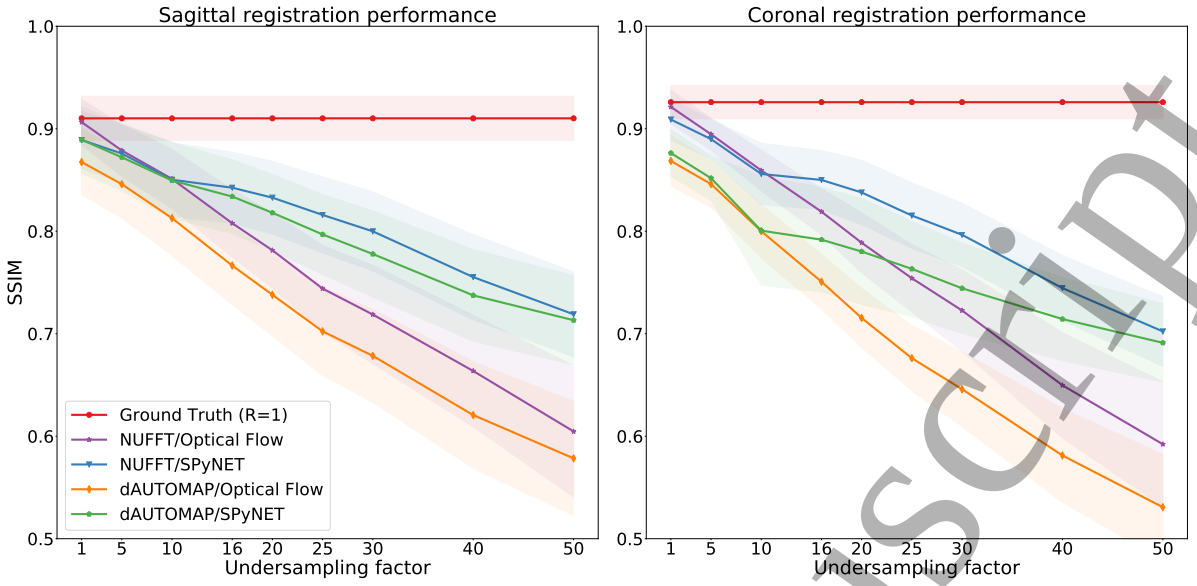

Figure 4: Comparison of the SSIM after registration over the whole image for sagittal images (left) and coronal images (right). Shaded regions indicate standard deviation.

NUFFT reconstruction with SPyNET motion estimation as the bias is reduced to -0.1 mm and the standard deviation of the absolute error remains within 2 mm, compared to a standard deviation up to 3.5 mm for NUFFT in combination with optical flow.

3.3. Time

At $\mathcal{R} = 25$, approximately 40 ms would be spent acquiring k-space of a single dynamic with TR=2.8 ms. Combined with a NUFFT adjoint reconstruction which takes 5 ms and a SPyNET forward evaluation of 15 ms, DVFs can be computed with high quality in 60 ms, which is more than adequate for real-time MRIgRT of respiratory induced moving targets.

Table 2 summarizes all quantitative results in the sagittal plane. It can be observed that almost 94% of all vectors have a root-mean-square error of less than 2 mm when computed with a NUFFT adjoint reconstruction and SPyNET for motion estimation.

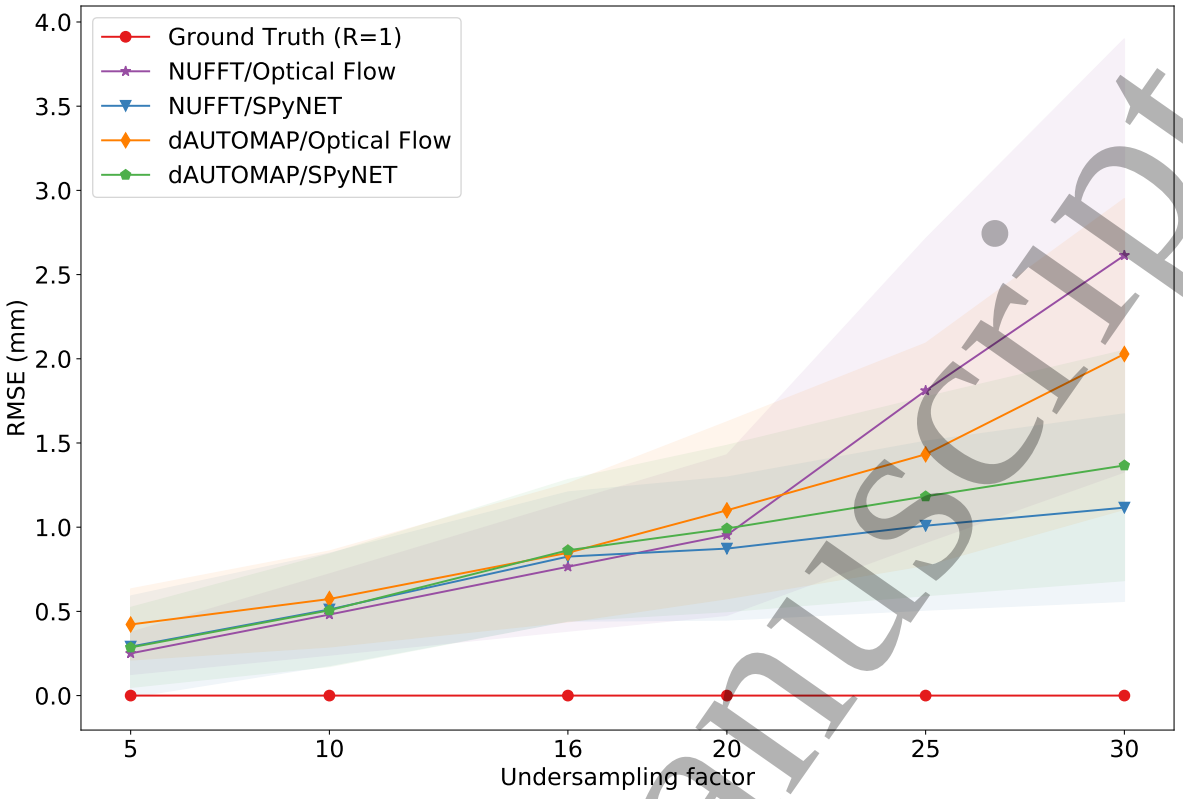

Figure 5: Root-mean-square displacement error within an ROI on sagittal images. Shaded regions indicate standard deviation.

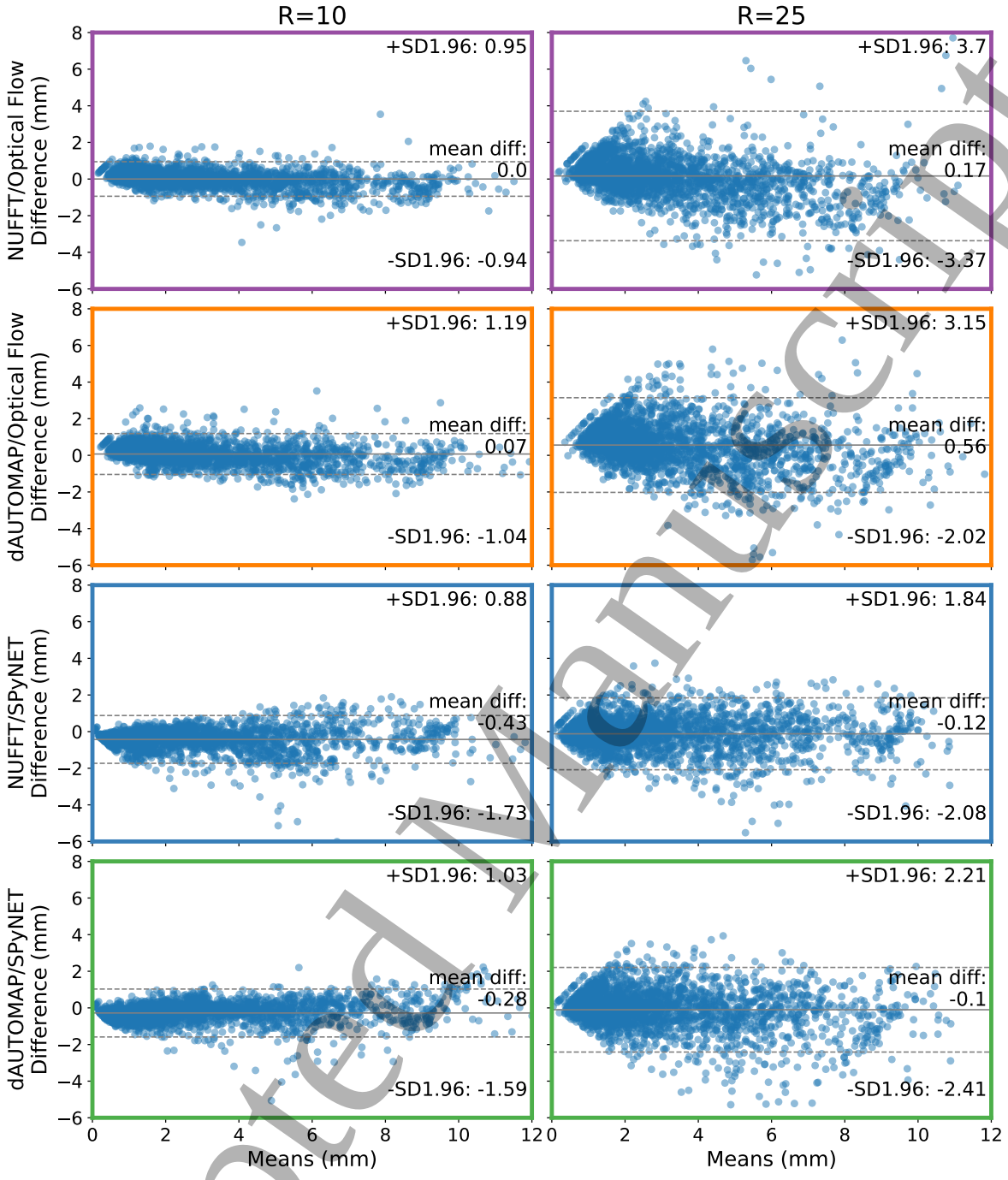

Figure 6: Bland-Altman plots of the average vector magnitude within an ROI on sagittal images as generated by the various model configurations at $\mathcal{R} = 10$ and $\mathcal{R} = 25$ compared to the ground-truth. A positive value indicates an overestimation compared to the ground-truth.

Table 2: Quantitative results for the four approaches in the sagittal plane, displaying the structural similarity index (SSIM) after registration for various undersampling factors, the root-mean-square error (RMSE) of the motion magnitude within an ROI (ROIs displayed in Appendix B), and the time it takes for MRI acquisition, image reconstruction, and motion estimation. Best results per metric per undersampling factor are marked in boldface, excluding ground-truth.

	Ground Truth (R=1)	**NUFFT Optical Flow**	**NUFFT SPyNET**	**dAUTOMAP Optical Flow**	**dAUTOMAP SPyNET**
			SSIM after registration		
$\mathcal{R}=1$	0.91 ± 0.04	$\mathbf{0.91 \pm 0.04}$	0.89 ± 0.06	0.87 ± 0.06	0.89 ± 0.06
$\mathcal{R}=5$	0.91 ± 0.04	$\mathbf{0.88 \pm 0.05}$	0.88 ± 0.06	0.85 ± 0.07	0.87 ± 0.07
$\mathcal{R}=10$	0.91 ± 0.04	$\mathbf{0.85 \pm 0.06}$	0.85 ± 0.07	0.81 ± 0.07	0.85 ± 0.07
$\mathcal{R}=16$	0.91 ± 0.04	0.81 ± 0.07	$\mathbf{0.84 \pm 0.07}$	0.77 ± 0.08	0.83 ± 0.07
$\mathcal{R}=20$	0.91 ± 0.04	0.78 ± 0.07	$\mathbf{0.83 \pm 0.07}$	0.74 ± 0.08	0.82 ± 0.07
$\mathcal{R}=25$	0.91 ± 0.04	0.74 ± 0.08	$\mathbf{0.82 \pm 0.07}$	0.70 ± 0.09	0.80 ± 0.08
$\mathcal{R}=30$	0.91 ± 0.04	0.72 ± 0.10	$\mathbf{0.80 \pm 0.08}$	0.68 ± 0.09	0.78 ± 0.08
$\mathcal{R}=40$	0.91 ± 0.04	0.66 ± 0.11	$\mathbf{0.76 \pm 0.08}$	0.62 ± 0.10	0.74 ± 0.09
$\mathcal{R}=50$	0.91 ± 0.04	0.60 ± 0.13	$\mathbf{0.72 \pm 0.08}$	0.58 ± 0.11	0.71 ± 0.09
			RMSE ≤ 1 mm (within ROI)		
$\mathcal{R}=5$	100%	**99.5%**	95.5%	98.0%	95.5%
$\mathcal{R}=10$	100%	**95.3%**	92.3%	92.8%	92.9%
$\mathcal{R}=16$	100%	86.1%	86.1%	81.4%	**88.5%**
$\mathcal{R}=20$	100%	78.1%	**83.4%**	70.1%	80.6%
$\mathcal{R}=25$	100%	69.9%	**76.6%**	56.0%	71.1%
$\mathcal{R}=30$	100%	61.8%	**70.3%**	48.1%	64.6%
			RMSE ≤ 2 mm (within ROI)		
$\mathcal{R}=5$	100%	**100.0%**	99.2%	100.0%	99.1%
$\mathcal{R}=10$	100%	**99.7%**	98.8%	99.2%	98.8%
$\mathcal{R}=16$	100%	97.8%	97.7%	96.8%	**98.1%**
$\mathcal{R}=20$	100%	95.6%	**96.8%**	93.9%	96.0%
$\mathcal{R}=25$	100%	91.7%	**94.8%**	86.7%	91.8%
$\mathcal{R}=30$	100%	85.8%	**93.9%**	80.1%	90.1%
			Time (acquisition/reconstruction/motion (ms))		
$\mathcal{R}=10$	500/1/5	**100/5/5**	100/5/15	**100/5/5**	100/5/15
$\mathcal{R}=20$	500/1/5	**50/5/5**	50/5/15	**50/5/5**	50/5/15
$\mathcal{R}=25$	500/1/5	**40/5/5**	40/5/15	**40/5/5**	40/5/15

4. Discussion

In this work, we have investigated the impact of conventional and DL-based approaches to estimate 2D DVFs from highly undersampled k-space for real-time MRIgRT applications. In particular, we have quantified how much specific deep learning models can accelerate MRI acquisition and processing over conventional techniques and in which step deep learning is beneficial to obtaining high-quality motion estimates. We have shown that motion can be estimated from heavily undersampled k-space with

high temporal resolution and low error compared to the ground-truth when images are reconstructed with a conventional NUFFT and motion is estimated with deep learning. For example, the mean absolute displacement error remained within 2 mm and the RMSE remained within 1 mm at $\mathcal{R} = 25$ while the SSIM after registration remained above 0.8 when motion is estimated with NUFFT adjoint image reconstruction and SPyNET is used. Our method can compute DVFs with these errors within 60 ms and induces a total latency of 40 ms of which 20 ms comes from MRI acquisition (Borman et al., 2018) and 20 ms comes from processing, but extra overhead may present itself in a prospective setting. This demonstrated that reconstruction of DVFs is feasible at very high undersampling factors despite severe artifacts in the reconstructed images, indicating that accurate motion estimation is more resilient to undersampling than high-quality image reconstruction.

Results show that using SPyNET for motion estimation rather than optical flow significantly improves DVF quality at undersampling factors $\mathcal{R} \geq 10$. Also, we observe that the best DL-based approach can achieve the same SSIM after registration as the fully conventional approach with approximately two times more undersampling.

Interestingly, applying SPyNET to NUFFT-reconstructed images also outperforms applying SPyNET to dAUTOMAP-reconstructed images. This indicates that general-purpose trained DL-based image reconstruction obtained with dAUTOMAP does not have added value for motion estimation. We observed that dAUTOMAP favored overly smoothed reconstructions at high undersampling factors. We hypothesize that this may be detrimental to recover motion information.

We believe we have designed a robust approach to motion estimation. Augmenting the input images with flips and rotations makes dAUTOMAP and SPyNET robust against slight angulations. Moreover, the NUFFT/SPyNET approach shows near-equivalent registration performance on coronal images compared to registration of sagittal images without retraining. When dAUTOMAP is used for image reconstruction, the performance is significantly lower on coronal images than on sagittal images as it fails to reconstruct high-quality coronal images when trained on sagittal images. Even though the networks were trained at $\mathcal{R} = 10$, evaluation at higher undersampling factors seems to have a low impact on the registration quality.

NUFFT/SPyNET is thus able to resolve incoherent streaking artifacts introduced by radial sampling. An interesting exploration would be to investigate whether other sampling strategies (e.g., variable-density spirals) achieve similar results, but this was considered out of the scope of this paper. This robustness of NUFFT/SPyNET could suggest that the model is well generalizable and might transfer to other body sites and contrasts without retraining, which is currently under investigation.

This method of a radial readout with NUFFT image reconstruction and SPyNET motion estimation could find its application in real-time MRI-guided radiotherapy applications. Keall et al. (2006) suggest that acquisition, motion estimation and dose delivery needs to happen within 200 milliseconds to maintain accuracy. By using NUFFT/SPyNET, accurate DVFs can be obtained at $\mathcal{R} = 25$ in 60 ms with a latency of

40 ms, including MR acquisition. This leaves ample time for adaptation of the radiation beam to counteract the motion. This could enable real-time tumor tracking to account for intra-fraction motion.

One of the limitations of our approach is that it requires a ground-truth motion estimate to learn. While computing a ground-truth is feasible for retrospectively undersampled data, obtaining a high-quality ground-truth motion estimate for prospectively undersampled in-vivo MR data is challenging. Prospective data will also be acquired with multiple receiver coils while this work is focused on single-coil images. Considering multi-coil images might be beneficial for motion estimation quality but also introduces new challenges. It requires more data needs to be evaluated, which might result in more parameters to train and higher inference times. Future work may investigate unsupervised approaches to learning motion or find another way to obtain motion estimates from k-space acquired with multiple receiver coils.

Another limitation is that our networks were only trained at $\mathcal{R} = 10$. Performance might be improved at high undersampling factors if they are retrained at $\mathcal{R} > 10$.

When compared to other works, our method is significantly faster while achieving similar accuracy at $\mathcal{R} = 25$, even when compared to other deep learning-based methods (Seegoolam et al., 2019; Stemkens et al., 2016; Haskell et al., 2019). Seegoolam et al. (2019) investigated motion estimation on 2D cardiac cine MRI for $\mathcal{R} = 9$ and $\mathcal{R} = 50$ achieving an average SSIM after registration of 0.93 at $\mathcal{R} = 9$ versus 0.86 in this work and an SSIM of 0.776 at $\mathcal{R} = 51.2$ versus 0.72 in this work. Also, they indicate that the motion estimation network shows better generalization than the reconstruction network for various undersampling factors, which is in accordance with what we observed. However, their reconstruction method takes approximately 1.8 seconds per frame, excluding MR acquisition which is a significant performance penalty.

Stemkens et al. (2016) obtained a 3D motion estimation with an RMSE of 1 mm using a 360 ms 2D acquisition and a few seconds of motion calculation. This error is comparable with we observed, even though their work estimates motion in three dimensions. This is, however, not a "full" 3D method but uses multi-2D cine scans in conjunction with a 4D MRI to obtain 3D motion estimates, limiting the accuracy of the method.

The approach by Haskell et al. (2019) significantly reduces motion artifacts in image space by combining a CNN with a physics-based model. This approach of combining DL to remove artifacts with conventional SENSE reconstruction (Pruessmann et al., 1999) produces the best results, which is in line with what we found. However, their approach requires fully-sampled data, and the full motion correction model requires several minutes to evaluate, making it unsuitable for real-time applications.

In this work, we showed that acquisition, reconstruction and motion estimation can be performed in approximately 60 ms for $\mathcal{R} = 25$ achieving a root-mean-square displacement error of less than 1 millimeter compared to a ground-truth motion estimate. This is of particular interest for applications with crucial time constraints, such as MRIgRT (Lagendijk et al., 2014). We believe that deep learning models play an

important role in facilitating real-time motion management on MR-Linacs, but should be carefully assessed, taking into account the entire feedback chain. Replacing an individual "classic" step in the processing pipeline by a DL alternative does not necessarily result in improved performance. We did show that using a DL-based motion estimation network in conjunction with a NUFFT yields a robust and generic method for motion estimation. The combination of highly undersampled k-space with DL-based methods yields high-quality motion estimation for a real-time MRIgRT with low latency, which makes it a worthwhile area of ongoing research.

In a future study, we will attempt to extend this method to a "full" three-dimensional real-time motion estimation method. We believe this will have a higher accuracy and performance than a multi-2D approach. Motion has been successfully estimated from fully-sampled 3D MR cardiac images (Morales et al., 2019), but the method has not been demonstrated for real-time applications. We will investigate whether the use of multi-channel MRI may further improve the current performances.

5. Conclusions

The performance of DL-based image reconstruction and motion estimation was assessed on retrospectively undersampled GA radial MRI to allow real-time motion estimation with minimal latency. It was found that DL-based motion estimation (SPyNET) allowed far greater acceleration factors than traditional optical flow based motion estimation. DL-based image reconstruction of undersampled radial data, however, did not result in better performance compared to standard NUFFT reconstructions in combination with SPyNET motion estimation. The NUFFT/SPyNET approach produced an acceptable performance for 25-fold accelerated data, thereby achieving an imaging frame rate of 25 Hz while the root-mean-square error remained within 1 millimeter.

6. Acknowledgement

This work is part of the research programme HTSM with project number 15354, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO) and Philips Healthcare. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro RTX 5000 GPU used for prototyping this research.

References

Altman, D. G. and Bland, J. M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. *The Statistician*, 32(3):307–317.

Bloice, M. D., Roth, P. M., and Holzinger, A. (2019). Biomedical image augmentation using Augmentor. *Bioinformatics*, 35(21):4522–4524.

Borman, P. T. S., Tijssen, R. H. N., Bos, C., Moonen, C. T. W., Raaymakers, B. W., and

Glitzner, M. (2018). Characterization of imaging latency for real-time MRI-guided radiotherapy. *Physics in Medicine & Biology*, 63(15):155023.

Dietz, B., Yun, J., Yip, E., Gabos, Z., Fallone, B. G., and Wachowicz, K. (2019). Single patient convolutional neural networks for real-time MR reconstruction: a proof of concept application in lung tumor segmentation for adaptive radiotherapy. *Physics in Medicine & Biology*, 64(19):195002.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., and Brox, T. (2015). FlowNet: Learning Optical Flow with Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766. IEEE.

Farnebäck, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. *Lecture Notes in Computer Science*, 2749:363–370.

Fast, M. F., Eiben, B., Menten, M. J., Wetscherek, A., Hawkes, D. J., McClelland, J. R., and Oelfke, U. (2017). Tumour auto-contouring on 2d cine MRI for locally advanced lung cancer: A comparative study. *Radiotherapy and Oncology*, 125(3):485–491.

Fessler, J. and Sutton, B. (2003). Nonuniform fast fourier transforms using min-max interpolation. *IEEE Transactions on Signal Processing*, 51(2):560–574.

Glitzner, M., de Senneville, B. D., Lagendijk, J. J. W., Raaymakers, B. W., and Crijns, S. P. M. (2015). On-line 3 D motion estimation using low resolution MRI. *Physics in Medicine and Biology*, 60(16):N301–N310.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Hammernik, K., Klatzer, T., Kobler, E., Recht, M. P., Sodickson, D. K., Pock, T., and Knoll, F. (2018). Learning a variational network for reconstruction of accelerated MRI data. *Magnetic Resonance in Medicine*, 79(6):3055–3071.

Haskell, M. W., Cauley, S. F., Bilgic, B., Hossbach, J., Splitthoff, D. N., Pfeuffer, J., Setsompop, K., and Wald, L. L. (2019). Network Accelerated Motion Estimation and Reduction (NAMER): Convolutional neural network guided retrospective motion correction using a separable motion model. *Magnetic Resonance in Medicine*, 82(4):1452–1461.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034. IEEE.

Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *2017 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 1647–1655. IEEE.

Keall, P., Poulsen, P., and Booth, J. T. (2019). See, Think, and Act: Real-Time Adaptive Radiotherapy. *Seminars in Radiation Oncology*, 29(3):228–235.

Keall, P. J., Mageras, G. S., Balter, J. M., Emery, R. S., Forster, K. M., Jiang, S. B., Kapatoes, J. M., Low, D. A., Murphy, M. J., Murray, B. R., Ramsey, C. R., Van Herk, M. B., Vedam, S. S., Wong, J. W., and Yorke, E. (2006). The management of respiratory motion in radiation oncology report of AAPM Task Group 76a). *Medical Physics*, 33(10):3874–3900.

Kim, T. H. and Haldar, J. P. (2018). The Fourier radial error spectrum plot: A more nuanced quantitative evaluation of image reconstruction quality. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 61–64. IEEE.

Kingma, D. P. and Ba, J. (2015). Adam: {A} Method for Stochastic Optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, {ICLR} 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klein, S., Staring, M., Murphy, K., Viergever, M., and Pluim, J. (2010). elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205.

Kontaxis, C., Bol, G. H., Stemkens, B., Glitzner, M., Prins, F. M., Kerkmeijer, L. G. W., Lagendijk, J. J. W., and Raaymakers, B. W. (2017). Towards fast online intrafraction replanning for free-breathing stereotactic body radiation therapy with the MR-linac. *Physics in Medicine & Biology*, 62(18):7233–7248.

Lagendijk, J. J. W., Raaymakers, B. W., Van den Berg, C. A. T., Moerland, M. A., Philippens, M. E., and van Vulpen, M. (2014). MR guidance in radiotherapy. *Physics in Medicine and Biology*, 59(21):R349–R369.

Lin, J.-M. (2018). Python Non-Uniform Fast Fourier Transform (PyNUFFT): An Accelerated Non-Cartesian MRI Package on a Heterogeneous Platform (CPU/GPU). *Journal of Imaging*, 4(3):51.

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

Lønning, K., Putzky, P., Sonke, J.-J., Reneman, L., Caan, M. W., and Welling, M. (2019). Recurrent inference machines for reconstructing heterogeneous MRI data. *Medical Image Analysis*, 53:64–78.

Lustig, M., Donoho, D., and Pauly, J. M. (2007). Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195.

Meyer, P., Noblet, V., Mazzara, C., and Lallement, A. (2018). Survey on deep learning for radiotherapy. *Computers in Biology and Medicine*, 98:126–146.

Morales, M. A., Izquierdo-Garcia, D., Aganj, I., Kalpathy-Cramer, J., Rosen, B. R., and Catana, C. (2019). Implementation and Validation of a Three-dimensional Cardiac Motion Estimation Network. *Radiology: Artificial Intelligence*, 1(4):e180080.

Mutic, S. and Dempsey, J. F. (2014). The ViewRay System: Magnetic Resonance–Guided and Controlled Radiotherapy. *Seminars in Radiation Oncology*, 24(3):196–199.

Pruessmann, K. P., Weiger, M., Scheidegger, M. B., and Boesiger, P. (1999). SENSE: sensitivity encoding for fast MRI. *Magnetic resonance in medicine*, 42(5):952–62.

Ranjan, A. and Black, M. J. (2017). Optical Flow Estimation Using a Spatial Pyramid Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2720–2729. IEEE.

Ravishankar, S. and Bresler, Y. (2011). MR Image Reconstruction From Highly Undersampled k-Space Data by Dictionary Learning. *IEEE Transactions on Medical Imaging*, 30(5):1028–1041.

Sahiner, B., Pezeshk, A., Hadjiiski, L. M., Wang, X., Drukker, K., Cha, K. H., Summers, R. M., and Giger, M. L. (2019). Deep learning in medical imaging and radiation therapy. *Medical Physics*, 46(1):e1–e36.

Schlemper, J., Caballero, J., Hajnal, J. V., Price, A. N., and Rueckert, D. (2018). A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503.

Schlemper, J., Oksuz, I., Clough, J. R., Duan, J., King, A. P., Schnabel, J. A., Hajnal, J. V., and Rueckert, D. (2019). dAUTOMAP: Decomposing AUTOMAP to Achieve Scalability and Enhance Performance . In *Proceedings of the International Society of Magnetic Resonance in Medicine*, number 27.

Seegoolam, G., Schlemper, J., Qin, C., Price, A., Hajnal, J., and Rueckert, D. (2019). Exploiting Motion for Deep Learning Reconstruction of Extremely-Undersampled Dynamic MRI. *Lecture Notes in Computer Science*, 11767:704–712.

Stemkens, B., Tijssen, R. H. N., de Senneville, B. D., Lagendijk, J. J. W., and van den Berg, C. A. T. (2016). Image-driven, model-based 3D abdominal motion estimation for MR-guided radiotherapy. *Physics in Medicine and Biology*, 61(14):5335–5355.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Wiesinger, F., Boesiger, P., and Pruessmann, K. P. (2004). Electrodynamics and ultimate SNR in parallel MR imaging. *Magnetic Resonance in Medicine*, 52(2):376–390.

Winkel, D., Bol, G. H., Kroon, P. S., van Asselen, B., Hackett, S. S., Werensteijn-Honingh, A. M., Intven, M. P. W., Eppinga, W. S. C., Tijssen, R. H. N., Kerkmeijer, L. G. W., de Boer, H. C. J., Mook, S., Meijer, G. J., Hes, J., Willemsen-Bosman, M., de Groot-van Breugel, E. N., Jürgenliemk-Schulz, I. M., and Raaymakers, B. W.

(2019). Adaptive radiotherapy: The Elekta Unity MR-linac concept. *Clinical and translational radiation oncology*, 18:54–59.

Zachiu, C., Denis de Senneville, B., Moonen, C., and Ries, M. (2015a). A framework for the correction of slow physiological drifts during MR-guided HIFU therapies: Proof of concept. *Medical Physics*, 42(7):4137–4148.

Zachiu, C., Papadakis, N., Ries, M., Moonen, C., and Denis de Senneville, B. (2015b). An improved optical flow tracking technique for real-time MR-guided beam therapies in moving organs. *Physics in Medicine and Biology*, 60(23):9003–9029.

Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R., and Rosen, M. S. (2018). Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492.

Zijlema, S. E., Tijssen, R. H. N., Malkov, V. N., van Dijk, L., Hackett, S. L., Kok, J. G. M., Lagendijk, J. J. W., and van den Berg, C. A. T. (2019). Design and feasibility of a flexible, on-body, high impedance coil receive array for a 1.5 T MR-linac. *Physics in Medicine & Biology*, 64(18):185004.

Appendix A.

High-quality ground-truth motion estimates are required for training SPyNET. To determine which motion estimation is best suited for the dataset used in this work, a preliminary study was conducted comparing the motion estimation quality of optical flow (Zachiu et al., 2015b) and Elastix (Klein et al., 2010). Both methods were compared and evaluated based on registration performance. This was measured by the structural similarity (SSIM) metric (Wang et al., 2004), the mean-squared-error (MSE) between the reference image and the registered image, and evaluation of the error spectrum plot (ESP) (Kim and Haldar, 2018) of the registered images compared to the reference image were calculated over the entire image.

Optical flow and Elastix DVFs were computed for all fully-sampled cine MRIs in the training dataset used in this work. Optical flow was computed as described in section 2.4.1 with $\beta = 0.6$. Elastix DVFs‖ were computed on four resolution levels using rigid, affine, and deformable motion estimation using B-splines. For rigid and affine motion was estimated using the mutual information metric. For deformable motion estimation, mutual information was used with weight 1, and a transform bending energy penalty was added with weight 2. For every cine MRI, 100 dynamic/reference image pairs were randomly sampled to ensure representative measurements. The average SSIM, MSE, and ESP were computed over 8100 dynamic/reference image pairs.

It was found that optical flow yielded an average SSIM of 0.920 ± 0.045, which was significantly higher than the average SSIM of Elastix registrations 0.899 ± 0.053 (Wilcoxon, $p < 0.001$). The average MSE was 3.63 ± 1.86 for optical flow, which was significantly lower than the MSE of Elastix (Wilcoxon, $p < 0.001$), which was 5.08 ± 2.45. The averaged ESP is shown in Figure A1. It can be observed that for nearly all frequencies, the error of optical flow is lower than for Elastix, except for the very highest frequencies.

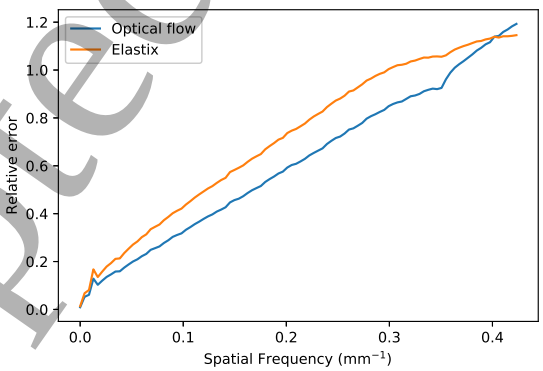

Figure A1: The average error spectrum plot of optical flow and Elastix registrations.

Based on these results have selected optical flow as ground-truth for evaluation and learning target for deep learning models.

‖ The exact parameter files can be found here: `http://elastix.bigr.nl/wiki/index.php/Par0060`

Appendix B.

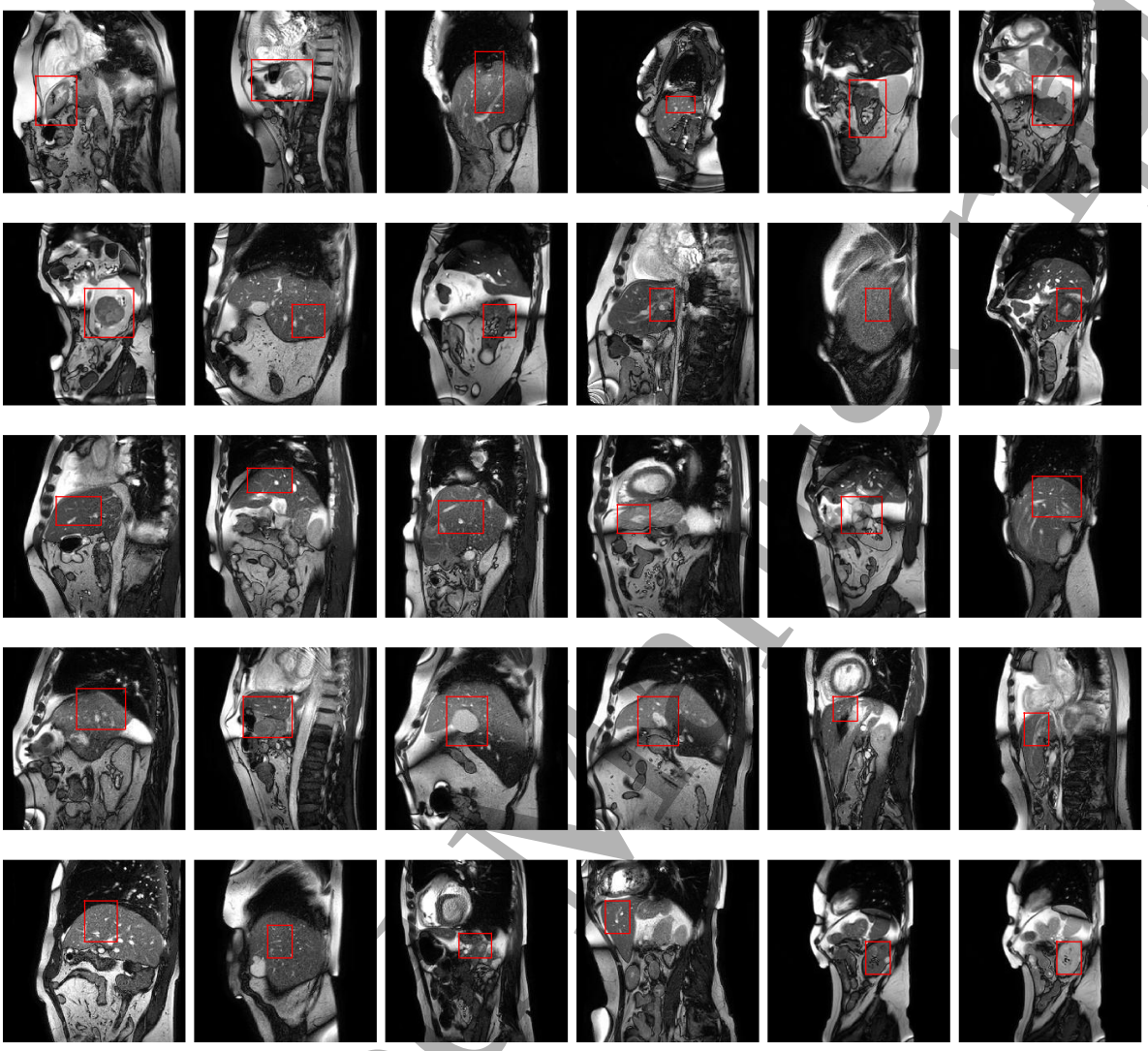

Figure B1: Manually generated regions-of-interest (ROIs) of the 30 patients used in the test set. These ROIs were used for the RMSE computation in Table 2 and the Bland-Altman plots in Figure 6. The ROIs were generated to include relevant structures and have an average size of 1010 ± 442 mm^2 or $4.1 \pm 1.8\%$ of the image.