Received: 17 August 2023 | Revised: 30 November 2023 | Accepted: 4 December 2023

DOI: 10.1002/mp.16918

RESEARCH ARTICLE

MEDICAL PHYSICS

AI-based motion artifact severity estimation in undersampled MRI allowing for selection of appropriate reconstruction models

Laurens Beljaards[1] | **Nicola Pezzotti**[2,3] | **Chinmay Rao**[1] | **Mariya Doneva**[4] | **Matthias J. P. van Osch**[1] | **Marius Staring**[1]

[1]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands

[2]Cardiologs, Philips, Paris, France

[3]Faculty of Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

[4]Philips Research, Hamburg, Germany

Correspondence
Laurens Beljaards, Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands.
Email: L.R.Beljaards@lumc.nl

Funding information
Dutch Research Council (NWO); Philips Research; ROBUST: Trustworthy AI-Based Systems for Sustainable Growth, Grant/Award Number: KICH3.LTP.20.006; Dutch Ministry of Economic Affairs and Climate Policy (EZK)

Abstract

Background: Magnetic Resonance acquisition is a time consuming process, making it susceptible to patient motion during scanning. Even motion in the order of a millimeter can introduce severe blurring and ghosting artifacts, potentially necessitating re-acquisition. Magnetic Resonance Imaging (MRI) can be accelerated by acquiring only a fraction of k-space, combined with advanced reconstruction techniques leveraging coil sensitivity profiles and prior knowledge. Artificial intelligence (AI)-based reconstruction techniques have recently been popularized, but generally assume an ideal setting without intra-scan motion.

Purpose: To retrospectively detect and quantify the severity of motion artifacts in undersampled MRI data. This may prove valuable as a safety mechanism for AI-based approaches, provide useful information to the reconstruction method, or prompt for re-acquisition while the patient is still in the scanner.

Methods: We developed a deep learning approach that detects and quantifies motion artifacts in undersampled brain MRI. We demonstrate that synthetically motion-corrupted data can be leveraged to train the convolutional neural network (CNN)-based motion artifact estimator, generalizing well to real-world data. Additionally, we leverage the motion artifact estimator by using it as a selector for a motion-robust reconstruction model in case a considerable amount of motion was detected, and a high data consistency model otherwise.

Results: Training and validation were performed on 4387 and 1304 synthetically motion-corrupted images and their uncorrupted counterparts, respectively. Testing was performed on undersampled in vivo motion-corrupted data from 28 volunteers, where our model distinguished head motion from motion-free scans with 91% and 96% accuracy when trained on synthetic and on real data, respectively. It predicted a manually defined quality label ('Good', 'Medium' or 'Bad' quality) correctly in 76% and 85% of the time when trained on synthetic and real data, respectively. When used as a selector it selected the appropriate reconstruction network 93% of the time, achieving near optimal SSIM values.

Conclusions: The proposed method quantified motion artifact severity in undersampled MRI data with high accuracy, enabling real-time motion artifact detection that can help improve the safety and quality of AI-based reconstructions.

KEYWORDS
motion corruption, motion artifact severity estimation, accelerated MRI, MRI reconstruction, deep learning

1 | INTRODUCTION

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging modality essential for visualizing the internal anatomy of a patient in the clinic, based on which a diagnosis can be made. While other modalities such as computed tomography (CT) can also acquire anatomical images in a non-invasive manner, MRI does not expose the subject to harmful ionizing radiation. However, a major drawback of MRI is that the process of acquiring the necessary k-space data is time-consuming, and requires subjects to lie still for extended periods of time, which can be especially challenging for young or very sick subjects. Small movements can already introduce severe blurring and ghosting artifacts,[1] necessitating re-acquisition. One study reports that 20% of the scans in the investigated hospital had to be reacquired as a result of motion artifacts, with an associated cost estimated at $115,000 per scanner every year.[2]

If the time required to traverse k-space during acquisition can be reduced, the scan will last shorter, thereby reducing the chance that the acquired image contains motion artifacts.[1] Acceleration of MRI can be achieved by parallel imaging methods (which utilize multicoil acquisition followed by coil-combination) or by compressed sensing methods (which utilize incoherent undersampling followed by sparsity promoting reconstruction). In both cases, only a fraction of k-space is acquired and a reconstruction technique leveraging prior knowledge is used to fully reconstruct the image. Recent research employing Artificial intelligence (AI) based reconstruction techniques has been successful,[3,4] but generally assumes an ideal setting without intra-scan motion.

Whereas acceleration of MRI acquisition lowers the risk of motion, residual motion artifacts may still result in images with reduced diagnostic quality. This problem is exacerbated by the fact that, when a lower percentage of k-space is sampled, the relative impact of motion becomes higher. In the era of AI-based reconstructions, however, such motion-corrupted data may confuse the network, giving rise to reconstructions that appear of sufficient quality but contain so-called hallucinations. Examples of hallucinations are the omission of existing pathology[3] or the wrongful creation of structures. As AI hallucinations may take a regular form (anatomy or pathology), they may go unnoticed and affect the subsequent clinical decision-making in contrast to traditional motion artifacts that are easily recognised by a radiologist. It is therefore of growing importance to reliably detect motion, preferably in a real-time and automated fashion. Data-driven motion detection is preferred in terms of simplicity and flexibility over magnetic resonance imaging (MR) navigators that come with increased complexity and acquisition times, as well as over external trackers that require additional hardware.[1,5]

Motion can be estimated in terms of motion parameters[6] on an inter-scan basis by for example, estimating deformation vector fields[7,8] and on an intra-scan basis by for example, aligned reconstruction,[9,10] which jointly searches for an uncorrupted multi-shot reconstruction and rigid-body motion parameters, with recent research inserting deep learning based components.[11,12] Besides estimating the motion parameters, retrospective motion estimation can also focus on estimating motion in terms of artifacts in the image, which is useful as a more direct metric of image quality, and is the focus of this paper. Attempts have been made to relate motion trajectories acquired with tracking systems to image quality, such as integrating motion based on head speed[13] and comparing that to known values of motion-corrupted cases.[14] Convolutional neural networks (CNNs) have been used to classify fully sampled MRI images as motion-free or corrupted,[15] to estimate motion between artificially corrupted images and the corresponding uncorrupted images in terms of structural similarity (SSIM),[16] or to generate probability maps for motion artifacts in MRI.[17,18]

While research has been done on estimating *motion parameters* of undersampled data, the aforementioned *motion artifact* estimation approaches all work on fully sampled images, whereas differentiation between motion and undersampling artifacts is challenging, and the focus of this paper. Moreover, undersampled imaging has become the clinical standard, which elevates the importance of differentiating motion artifacts from undersampling artifacts, that can have similar appearance. Our primary contributions are as follows:

1. We propose a deep-learning based regressor that can accurately estimate the severity of motion artifacts in brain MRI scans. Its ability to detect motion artifacts in *undersampled* acquisitions is critical for potential application in the clinic.
2. We synthesize motion-corrupted raw MR data from motion-free data, for training with realistic motion-corrupted and corresponding uncorrupted ground truth image pairs. We demonstrate on a prospectively motion-corrupted test set that synthetic motion-corrupted data can be used effectively during training in the common case no labeled data with in vivo intra-scan motion is available.

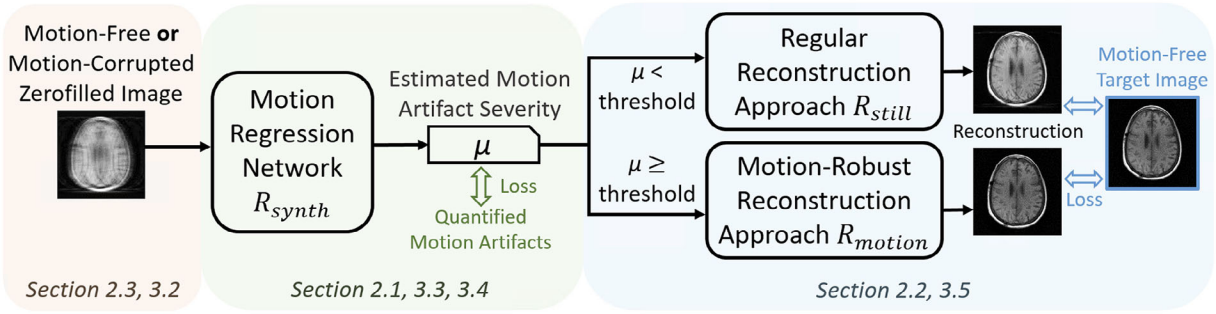

FIGURE 1 Overview of our framework. The reconstruction models also receive the coil sensitivity maps as input, which are omitted from this figure.

TABLE 1 Overview of the different models.

Model	Type	Arch.	Loss	Training Data	Test Data	Unders.
$\mathcal{R}_{\text{synth}}$	Regressor	*g*	Equation (1)	Synthetic & Still	Prospective & Still	1x to 8x
$\mathcal{R}_{\text{VGG19}}$	Regressor	VGG19	Equation (1)	Synthetic & Still	Prospective & Still	1x to 8x
$\mathcal{C}_{\text{mrart}}$	Classifier	*g*	-	Prospective & Still	Prospective & Still	1x to 8x
$\mathcal{R}_{\text{still}}$	Reconstructor	*f*	Equation (2)	Still	Prospective & Still	4x
$\mathcal{R}_{\text{motion}}$	Reconstructor	*f*	Equation (2)	Synthetic & Still	Prospective & Still	4x

Note: *Arch* means architecture, *unders* means undersampling factor. *Synthetic* refers to the synthetic retrospectively motion-corrupted training set, *Still* to motion-free data, and *Prospective* to the test set with real prospective motion.

3. Our model is able to detect motion corruption during acquisition before all k-space lines are sampled, enabling the possibility to alert the MR technician early, before the scan is fully completed. It could also prove valuable as a safety mechanism to prevent low-quality input data being used by AI-based approaches. Another use case is to leverage the model in a reconstruction framework, for example in an optimization strategy that uses the estimated motion artifact severity as a quality heuristic, or for the grouping criteria of adjacent k-space shots when little motion is detected between those shots in an approach like DISORDER,[10] or as stopping criterion in a model-based reconstruction.
4. We introduce and evaluate one potential use-case for the regressor: a deep-learning based reconstruction framework that falls back on a motion-robust solution when a considerable amount of motion is detected. This can improve quality if the motion-robust solution performs less well on motion-free cases, or improve speed if it is slower than a regular reconstruction approach.
5. We show that our approach accurately estimates motion artifact severity on prospectively and retrospectively motion-corrupted in vivo MRI data of the brain with an undersampling factor between $1-8\times$. Additionally, we investigate the effects of rigid-body motion on AI-based reconstructions by training reconstruction models with and without motion-corrupted data. We show an improvement in performance on a mixed set of motion-free and motion-corrupted data as a result of our selector framework.

2 | METHODS

An overview of our framework is given in Figure 1 and of our models in Table 1. The motion corruptor can use one or more motion-free datasets to generate a large amount of training pairs of motion-corrupted images with corresponding motion-free ground truth. Either an undersampled motion-corrupted image or an undersampled motion-free image is provided to the networks during training and inference. The framework contains a motion-robust and a high data consistency reconstruction network that are trained to reconstruct either motion-corrupted data or still data, respectively. The severity of motion artifacts as estimated by the motion artifact regressor determines whether the high data consistency or motion-robust reconstruction is deployed.

2.1 | Motion artifact regression network

For estimating the amount of motion in a scan, we use a CNN architecture g with weights ϕ called $\mathcal{R}_{\text{synth}}$ that takes an undersampled motion-corrupted 2D zero-filled image slice $\mathbf{x}$ as input and returns the predicted motion artifact amount μ. We focused on a 2D image acquisition protocol as this constitutes the majority of scans acquired in radiological practice. We propose a loss that quantifies the average pixel-wise difference between the motion-corrupted zero-filled image and the motion-free image, thereby optimizing

$$\arg\min_{\phi} \|g_{\phi}(A^H A C\mathbf{x}) - \|A^H A C\mathbf{x} - A^H A\mathbf{x}\|_1\|_1, \qquad (1)$$

where the inner L1-norm operator provides a scalar representing the mean pixel-wise difference between the images, and the measurement operator A multiplies the image with the coil sensitivities, applies the Fourier transform, and finally applies a mask that zeroes out the k-lines that were not measured. A^H is the Hermitian transpose of A, and $A^HA\mathbf{x}$ applies k-space masking to $\mathbf{x}$. This loss aims to approximate the intensity of motion artifacts in the motion-corrupted image, by isolating the effect of the motion-corruption operator C on the uncorrupted image. We reason that comparing the zero-filled images rather than the fully sampled images in the loss may lead to more stable training as the network input is undersampled as well. For the motion-free scans, C is the identity function and thus the target is 0. We define $\mathcal{R}_{\text{synth}}$ as a CNN with seven blocks, each consisting of two convolutional with corresponding leaky rectified linear unit (ReLU) layers, followed by a max pooling layer. Each block has half the width and height but double the number of feature layers compared to the previous block. The final block ends with a linear layer instead of a max pooling layer. $\mathcal{R}_{\text{synth}}$ uses 9M parameters. We compare this architecture against VGG19,[19] a popular architecture in medical imaging, trained in the same way on the same data. We used a VGG19 model with pre-trained weights (IMAGENET1K_V1), where after the original output layer we added one fully connected layer that returns the estimated motion artifact severity. This model $\mathcal{R}_{\text{VGG19}}$ uses 144M parameters.

To mitigate the effect of training variance, we employ an ensemble that combines 21 instances of the same network design that were trained with a different seed and order of data. The median prediction of the 21 networks is considered to be the prediction of the ensemble. For testing we selected the network that performed the best on the validation set.

2.2 | Reconstruction selector

We propose a reconstruction 'selector' framework that utilizes the regressor $\mathcal{R}_{\text{synth}}$ to estimate the level of motion artifacts in the image data to be reconstructed. If substantial motion is detected, a motion-robust approach is used for reconstruction, otherwise a regular reconstruction approach is used. This can improve quality when the motion-robust solution performs less well on motion-free cases, or improve speed when it is slower than a regular reconstruction approach. The regular reconstruction model $\mathcal{R}_{\text{still}}$ is trained to reconstruct motion-free k-space data, while the motion-robust model $\mathcal{R}_{\text{motion}}$ receives an undersampled corrupted image slice as input during training, with the target being the corresponding fully sampled uncorrupted image. Both reconstruction models use a modified version of the Adaptive-CS-Network architecture,[20] which is a deep-learning unrolled iterative reconstruction scheme consisting of a sequence of blocks that each apply a reconstruction and a data consistency operation. Specifically, we lowered the number of blocks to ten for faster training. The data consistency enforces similarity between the measured k-space data and reconstructed k-space[21] as follows:

$$\mathbf{r}_{i+1} = \mathbf{x}_i - \lambda_i A^H(A\mathbf{x}_i - y), \tag{2}$$

where $\mathbf{x}_i$ is the output of the reconstruction step in block i, $\mathbf{r}_{i+1}$ is the 'data consistent' residual image, and λ_i is a learned data consistency modifier for block i, allowing the data consistency to be imposed less strongly if beneficial for performance. The training loss for reconstruction architecture f with weights θ, which differ per reconstruction model, is described by:

$$\arg\min_{\theta} w\left\| |f_\theta(A^HAC\mathbf{x})| - |\mathbf{x}| \right\|_1 + \sum_i \left\| |f_\theta(A^HAC\mathbf{x}_i)| - |\mathbf{x}| \right\|_1. \tag{3}$$

The first term is the loss for the final predicted image and the second term is the loss for the intermediate image predicted by block i, compared against the fully sampled ground truth image. The first term is weighted stronger via w, which we set to 50.

2.3 | Synthesizing motion-corrupted K-Space data

We simulate motion-corrupted MRI acquisitions with a linear interleaved scanning protocol. We synthesize motion-corrupted scans by applying a 3D rigid motion pattern to a given still image over a series of timesteps to simulate the motion of the subject in the scanner. During each timestep, we apply a 3D shift and rotation in image space, and convert to k-space to sample relevant k-space lines according to a linear interleaved multi-slice pattern. Finally, we combine the sampled lines from each timestep into one motion-corrupted k-space. This process is also illustrated in Figure 2. The coil sentitivity maps were kept unchanged. We used 524 rigid-body head motion patterns acquired on a scanner with an optical tracking system from a study where participants were instructed to perform shaking or nodding motion.[22] 393 patterns were used for training and 131 for validation. We multiplied each measured motion pattern by a uniformly randomly selected factor between 0 and 2 in order to generate a variety of training images with different amounts of motion. For motion-corrupted validation images, the patterns were either strengthened or weakened by a random factor of up to 2. The motion patterns were only used for training data generation, our approach does not require them during inference.

2.4 | K-Space undersampling patterns

The regressor and reconstruction networks are fed images reconstructed from zero-filled k-space data

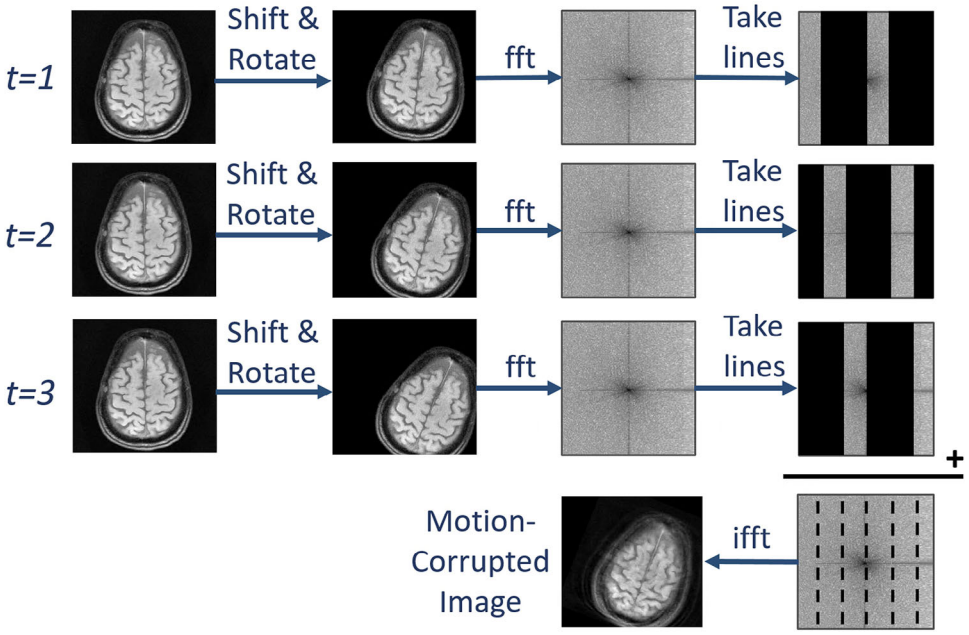

FIGURE 2 Overview of the motion-corruption process. It starts with a still brain image and a motion pattern, from which a motion-corrupted k-space and image are generated. For illustration purposes, the motion is exaggerated and the number of timesteps reduced.

that was retrospectively undersampled using Cartesian masks, with k-space lines set to zero in the phase encoding direction. As a default undersampling mask, we use the uniformly random distributed sampling pattern as used for the 2019 FastMRI challenge.[3] Given an undersampling factor R and a center fraction parameter c that determines the fraction of lines guaranteed to be sampled from the center of k-space, the remaining lines are sampled with probability $p = (\frac{1}{R} - c)/(1 - c)$. For some experiments we also used an undersampling mask based on Poisson disk sampling[23] that reflects better undersampling patterns as used by clinical MRI scanners. It precludes the occurrence of large unsampled gaps in k-space that may be present in completely random sampling approaches. The sampling probability is higher near the center (low frequencies), and lower near the edges (high frequencies). Our implementation incorporates a center fraction parameter that guarantees a sufficient number of centermost k-space lines are sampled.

3 | EXPERIMENTS AND RESULTS

3.1 | Training and implementation details

The motion artifact regression models were trained for 9×10^3 iterations with a batch size of 64 on an NVIDIA Quadro RTX 6000 GPU, while the reconstruction models were trained for 3×10^6 iterations. For the regression models we selected the retrospective undersampling factors per slice random uniformly from the range of 1× to 8×, with a corresponding central fraction between 11% and 4%, so that the models learn to estimate motion artifact severity on scans with every possible reasonable acceleration factor. Even the low accelerations are interesting as the amount of motion still needs to be correctly estimated and be discerned from undersampling artifacts. For reconstruction we used an acceleration of 4× with 8% central fraction since the low accelerations are considered to be trivial and 4× acceleration is clinically the most relevant at the moment.

3.2 | Data

We used multicoil brain T1, T2 and FLAIR scans from the NYU FastMRI brain dataset[24] as motion-free images and to generate retrospectively 3D-motion-corrupted images. We focused on a 2D image acquisition protocol as this constitutes the majority of scans acquired in radiological practice. Since 2D acquisition protocols are subject to through-plane motion as well, we used 3D motion. We discarded images with width or height smaller than 320 voxels. For training we used 4267 uncorrupted scans, and the same scans again as basis for 4267 retrospectively motion-corrupted training scans. 1389 of the training scans are T1-weighted, 2668 are T2-weighted and 210 are FLAIR. For validation, we used 1304 uncorrupted scans and 1304 derived motion-corrupted scans (427 T1, 809 T2, 68 FLAIR).

For prospectively motion-corrupted data we used the motion-related artifacts MR-ART dataset.[25] This dataset consists of 3D T1 brain images from 148 volunteers,

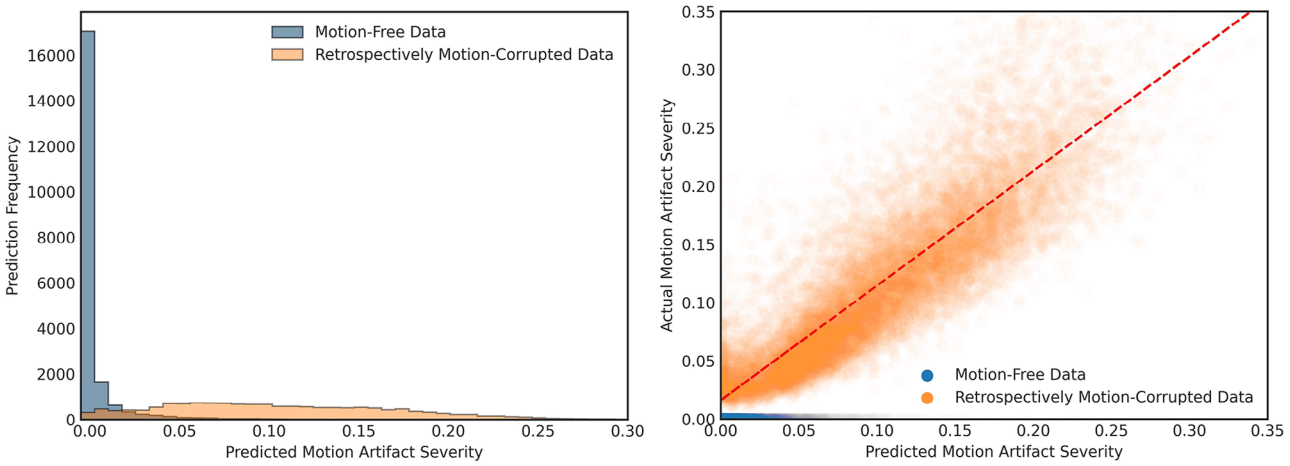

FIGURE 3 Histogram and scatterplot of motion severity predictions during the evaluation of $\mathcal{R}_{\text{synth}}$ on retrospectively motion-corrupted undersampled NYU FastMRI challenge data. The trend line for predictions on motion-corrupted data is $0.984x + 0.016$.

who were asked to lie still, nod five times ('Head Motion 1'), and nod ten times ('Head Motion 2') during three separate scans, respectively. For the motion-free volunteer task, 119/28/1 image volumes were labeled as 'Good'/'Medium'/'Bad' quality images by neuroradiologists. This changed to 7/59/75 for the first head motion task and 3/22/122 for the second head motion task. Twenty-eight sets of three scans each were used as test set. We resampled the MR-ART images from 1×1 mm to 0.6875×0.6875 mm followed by cropping to 320×320 pixels to match the resolution of the FastMRI challenge dataset. We discarded slices above the top of the head since those contain no information, as well as any slices $>$ 85 mm lower to avoid slices where the face had been removed for anonymization purposes and to match the NYU FastMRI data that only contains the top of the head.

3.3 | Performance of motion artifact regressor on retrospectively motion-corrupted data

The motion artifact regression model $\mathcal{R}_{\text{synth}}$ was trained on 4267 and 120 motion-free and 4267 and 120 derived synthetically motion-corrupted NYU and MR-ART cases, respectively. While prospectively corrupted MR-ART data was available, only retrospectively corrupted data was used during training as we wanted to investigate the effectiveness of training exclusively on synthetically generated data. For validation, we used 1304 uncorrupted scans and 978 motion-corrupted scans.

While our approach predicts the motion amount as a scalar, for evaluation we instead report the ability to differentiate between still and motion scans based on the predicted motion amount, since the quantified motion artifacts is a non-interpretable number and to allow for a comparison against other approaches. Figure 3 displays a histogram of the predictions of $\mathcal{R}_{\text{synth}}$ on retrospectively motion-corrupted NYU FastMRI data. Given a classification threshold separating the two peaks in the histogram at the lowest place in between, the model has an accuracy of 93.1%, with a false positive rate of 6.6% (incorrectly predicting a high amount of motion on a still scan) and a false negative rate of 7.3% (incorrectly predicting less motion than the threshold on a scan affected by motion). These results demonstrate that the regressor can accurately detect motion and differentiate between still and retrospectively motion-corrupted images.

3.4 | Performance of motion artifact regressor on prospectively motion-corrupted data

To assess whether the results on retrospectively motion-corrupted data are indicative of performance on prospectively motion-corrupted scans, we also evaluate the model $\mathcal{R}_{\text{synth}}$ on 28 volunteers (84 scans in total) from the MR-ART dataset. While the regressor is able to quantify motion artifacts as a scalar, only three discrete clinical artifact score labels by neuroradiologists ('good' $<$ 'medium' $<$ 'bad') and three task labels (still $<$ head motion 1 $<$ head motion 2) are available on the prospectively motion-corrupted data, as described in Section 3.2. We therefore evaluated the model based on the correspondence between its motion estimations and the artifact or task labels. First, we provided slices from the three different tasks per volunteer to the model and checked whether the difference in estimated motion artifact severity between the slices was in line with their labels, provided of course that the labels were different, that is, were not both 'Bad'. For a given slice pair (there are three possible pairs per slice per volunteer), the

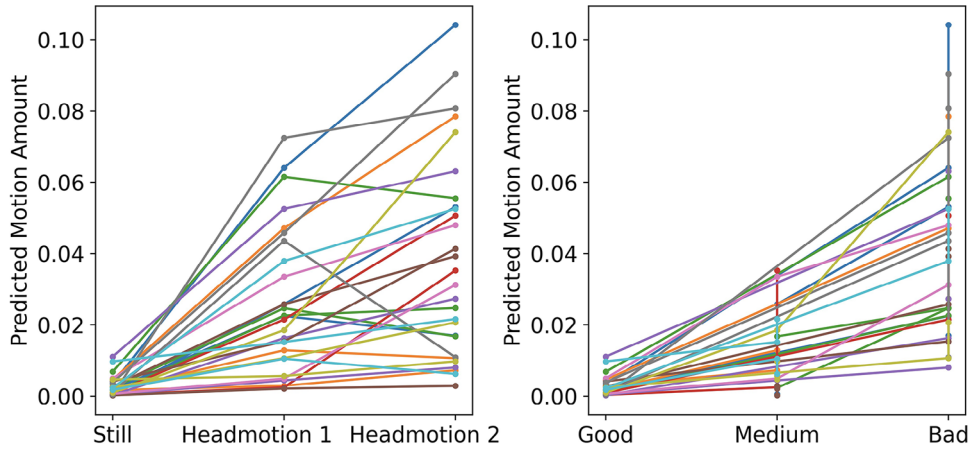

FIGURE 4 Predictions of our model $\mathcal{R}_{\text{synth}}$ on MR-ART data for each set of three scans per volunteer, with each scan grouped either by motion task (left) or artifact label (right). Each colored line corresponds to a different volunteer. Lines are vertical if the same artifact label was given to multiple tasks of the same volunteer. The data point for each scan was obtained by averaging the predictions for all its slices, with each slice having a random undersampling factor.

TABLE 2 Performance of the different models on MR-ART data with a random undersampling factor between 1 – 8×.

Data label:	Good Artifact score			Medium Artifact score			Bad Artifact score			3 Label-	2 Label-
Prediction	Good↑	Med↓	Bad↓	Good↓	Med↑	Bad↓	Good↓	Med↓	Bad↑	accuracy	accuracy
$\mathcal{R}_{\text{synth}}$	88%	12%	0%	30%	48%	22%	0%	14%	86%	76%	88%
$\mathcal{R}_{\text{VGG19}}$	76%	8%	16%	35%	22%	43%	0%	0%	100%	71%	83%
$\mathcal{C}_{\text{mrart}}$	96%	4%	0%	13%	57%	30%	0%	6%	94%	85%	95%

model correctly orders the artifact score labels 98.1% of the time. If instead of the artifact label we consider the task (still < head motion 1 < head motion 2), the model correctly predicted the ordering only 89.7% of the time. If we only compared the still task against head motion task 1 and 2, that is, not including the comparison of head motion task 1 against 2, this increased to 98.0%. In other words, estimating which task was performed is challenging since the artifacts caused by head motion task 1 and 2 are hard to distinguish from each other. Figure 4 illustrates the average predictions of $\mathcal{R}_{\text{synth}}$ for each set of three scans per volunteer on a per-volume basis. The model predicted an increasing amount of motion 100% of the time as the label of a volunteer progressed from Good to Medium or from Medium to Bad. When looking at predictions between different volunteers, it can be seen that the model sometimes estimated a scan to have more severe motion artifacts than a scan of another volunteer despite it having a better label.

We also evaluated the model on the prospectively motion-corrupted data using the motion amount predicted by our model to bin scans into one of three artifact score categories (Good, Medium or Bad quality) based on two thresholds, which were selected to maximize accuracy on 120 MR-ART cases not included in the test set. The results on the prospectively motion-corrupted test set are shown in Tables 2 and 3.

The VGG19 architecture did not perform as well as $\mathcal{R}_{\text{synth}}$ on the test set using the same training conditions and data, potentially because its architecture is much larger (namely seven times more parameters), making it more susceptible to overfitting. We used pre-trained weights for $\mathcal{R}_{\text{VGG19}}$ as it improved the accuracy by about 2%.

We also trained a model on MR-ART artifact labels specifically ($\mathcal{C}_{\text{mrart}}$) on 120 training cases. As this work focuses on training with synthetic data ($\mathcal{R}_{\text{synth}}$), this experiment is not intended to be a competitive comparison, but it serves to demonstrate what performance can be reached when training to predict the labels on prospectively motion-corrupted data. The model correctly ordered the artifact score labels 99.85% of the time for a given slice pair per subject, for example given a 'medium' and 'bad' slice, the model correctly predicted almost always more motion for the 'bad' slice. It achieved an accuracy on inter-subject separability between Good and Medium/Bad volumes of 95.2%, though that required reducing the task of the model to a classification task and training on prospectively motion-corrupted data, both of which were not the goal of this paper.

In an alternative experiment where $\mathcal{R}_{\text{synth}}$ was trained and evaluated on fully sampled data instead of undersampled data, it correctly ordered the artifact score labels 98.5% of the time for a given slice pair per subject,

TABLE 3 Performance of the different models on MR-ART data with a random undersampling factor between 1 – 8×.

Task:	Still task			Head motion task 1			Head motion task 2			3 Task-	2 Task-
Prediction	Still↑	HM1↓	HM2↓	Still↓	HM1↑	HM2↓	Still↓	HM1↓	HM2↑	accuracy	accuracy
$\mathcal{R}_{\text{synth}}$	93%	7%	0%	14%	39%	46%	7%	21%	71%	68%	91%
$\mathcal{R}_{\text{VGG19}}$	93%	7%	0%	25%	57%	18%	11%	39%	50%	67%	86%
$\mathcal{C}_{\text{mrart}}$	96%	4%	0%	4%	39%	57%	4%	11%	86%	74%	96%

HM1 and HM2 stand for head motion task 1 and 2.

TABLE 4 Performance in SSIM of the individual reconstruction models and the selector framework.

Dataset:	Still Data	Motion Data
Metric	[median] $\mu \pm \sigma$ SSIM	[median] $\mu \pm \sigma$ SSIM
Fully Sampled	[1.000] 1.000 ± 0.000	[0.762] 0.742 ± 0.161
Zero-Filled	[0.761] 0.750 ± 0.077^a	[0.648] 0.629 ± 0.108^a
$\mathcal{R}_{\text{still}}$	[0.915] 0.892 ± 0.094^a	[0.773] 0.746 ± 0.131^a
$\mathcal{R}_{\text{motion}}$	[0.909] 0.887 ± 0.095^a	[0.837] 0.805 ± 0.119^a
$\mathcal{R}_{\text{selector}}$	[0.914] 0.891 ± 0.096	[0.835] 0.803 ± 0.118

Note: The reconstruction models received 4× undersampled data, thus the fully sampled motion-corrupted data should not be seen as a baseline.
[a] Denotes a significant difference from $\mathcal{R}_{\text{selector}}$.

and achieved an accuracy on inter-subject separability between Good and Medium/Bad volumes of 88.1%, and between still data and head motion volumes of 91.7%.

3.5 | Reconstruction models and selector performance

The reconstruction models were initially trained for 1.7M iterations with one slice per iteration on uncorrupted brain T1, T2 and FLAIR images from the NYU FastMRI dataset,[24] and subsequently finetuned for 1M iterations on either 4267 uncorrupted cases (model $\mathcal{R}_{\text{still}}$) or both 4267 uncorrupted and 4267 derived retrospectively motion-corrupted cases (model $\mathcal{R}_{\text{motion}}$). The retrospectively motion-corrupted dataset was used as opposed to the MR-ART dataset as only the former contained enough data to train a reconstruction model on.

Table 4 shows the results of the two reconstruction models. The introduction of motion caused a large decrease in reconstruction quality of $\mathcal{R}_{\text{still}}$. In comparison, the model $\mathcal{R}_{\text{motion}}$ that was fed motion-corrupted data during training displayed better performance when evaluated on motion-corrupted data, although the reconstructions appear less sharp than the uncorrupted fully sampled images. On still data, the performance of $\mathcal{R}_{\text{motion}}$ decreased compared to $\mathcal{R}_{\text{still}}$ by 7.1%, relative to the target SSIM of 1. When inspecting the learned data consistency modifiers that allow the data consistency to be imposed less strongly, we measured a decreased weighing in the first block from 1.00 for $\mathcal{R}_{\text{still}}$ to 0.65 for $\mathcal{R}_{\text{motion}}$, and of the remaining blocks from an average of 0.71 to 0.13. The change in data consistency strength indicates that it is beneficial for the motion-induced models to have more freedom to compensate for motion in the measured k-space, and may explain the lower performance of $\mathcal{R}_{\text{motion}}$ on motion-free data. Example reconstructions can be seen in Figure 5.

We investigated a way to leverage the advantages of both the standard and motion-robust reconstruction approaches, by creating a motion-adaptive reconstruction framework based on a model selection mechanism that does not compromise on quality or data consistency for cases without motion. Using the amount of motion artifacts as predicted by the developed regressor, we select between a motion-robust versus a conventional reconstruction model. If $\mathcal{R}_{\text{synth}}$ quantifies the severity of motion artifacts to be above 0.025, we considered the scan to be motion-corrupted and it was thus fed into $\mathcal{R}_{\text{motion}}$ rather than $\mathcal{R}_{\text{still}}$. Table 4 shows the results of the selector framework. The selector framework comes very close to the performance of $\mathcal{R}_{\text{still}}$ on motion-free data and performed much better on motion-corrupted data, for which it almost matched the performance of $\mathcal{R}_{\text{motion}}$. The gold standard is for the selector framework to perform as well as $\mathcal{R}_{\text{still}}$ on still data and as well as $\mathcal{R}_{\text{motion}}$ on motion-corrupted data. The fact that the framework got close to the optimal values indicates that the regressor model was most of the time able to correctly provide $\mathcal{R}_{\text{still}}$ with motion-free scans and $\mathcal{R}_{\text{motion}}$ with scans containing motion artifacts.

On motion data, in the cases where the regressor provided accurate/inaccurate results and selected the optimal/suboptimal network, the difference between the reconstructions of $\mathcal{R}_{\text{motion}}$ and $\mathcal{R}_{\text{still}}$ was 0.0657/0.0235 SSIM on average. This indicates that classification errors on motion data are made more often when the reconstructions are similar. On still data the difference was 0.0054/0.0062, indicating that classification errors on still data are made more often when the reconstructions are different.

3.6 | The effect of the sampling scheme on performance

We investigated the performance of different sampling strategies, under the assumption that sampling the same set of k-lines for different cases may make the undersampling artifacts more predictable and thereby motion artifacts easier to identify. Acceleration during

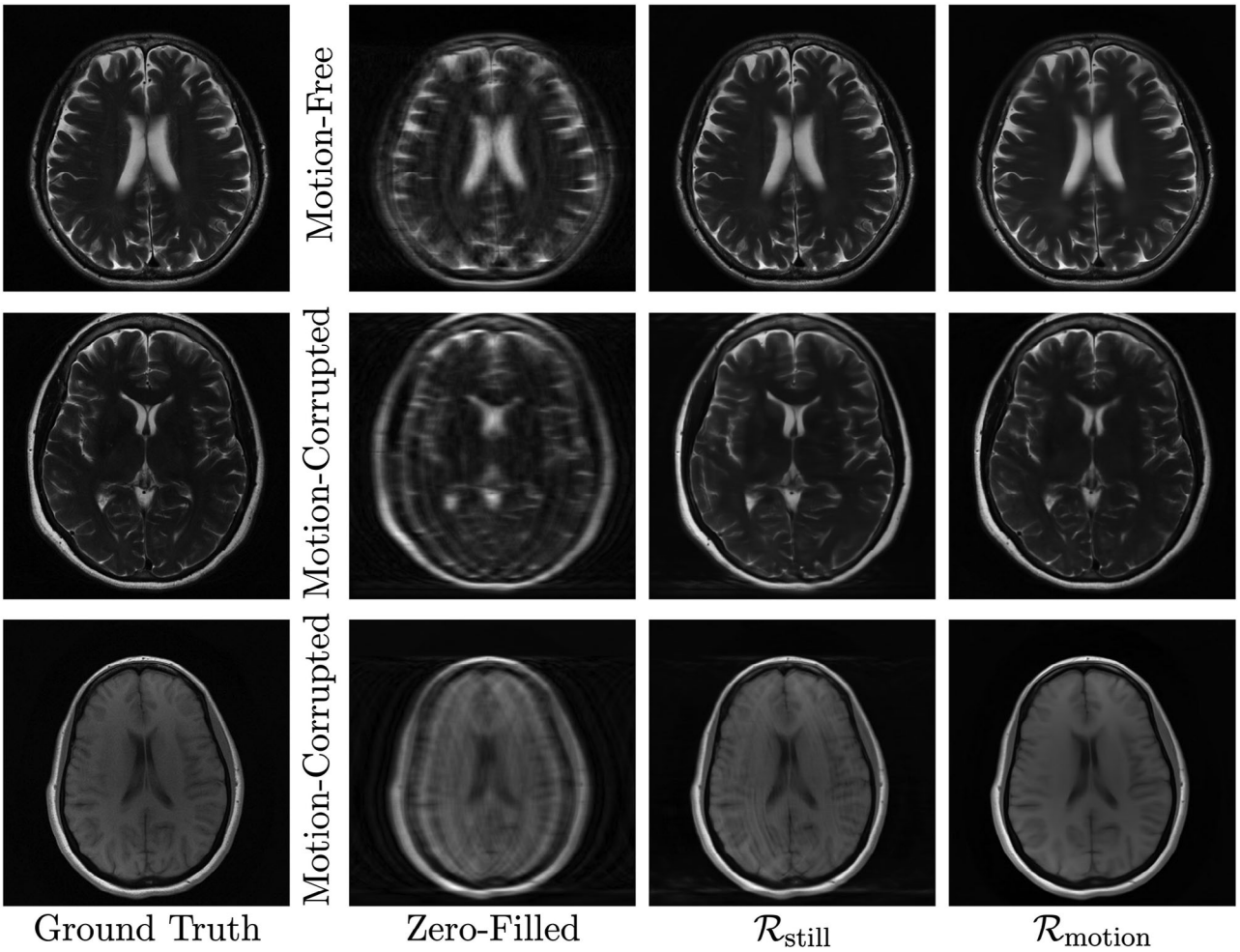

FIGURE 5 From left to right: Fully sampled uncorrupted image, the 4× undersampled zero-filled image (network input) that is uncorrupted in the first row and motion-corrupted in the other rows, the reconstruction by $\mathcal{R}_{\text{still}}$ and the reconstruction by $\mathcal{R}_{\text{motion}}$. The first row shows a case where $\mathcal{R}_{\text{still}}$ performed better, as $\mathcal{R}_{\text{motion}}$ blurred structure away in the top left and did not reconstruct detailed structures at the center top. The second row of images shows a case where $\mathcal{R}_{\text{motion}}$ performed better, as the reconstruction by $\mathcal{R}_{\text{still}}$ propagated some of the motion artifacts. The bottom images display a case where both models performed suboptimally, as the anatomy is smoothed away at the top and the bottom.

training and evaluation was fixed at 4×. Using random uniform probability undersampling masks, the model achieved an accuracy of 76% when predicting the exact label (Good/Med/Bad). We compared this random approach, which samples a different set of k-lines for each case, to a similar masking strategy that always samples the same set of k-lines for every scan. This approach achieved an accuracy of 75%. The Poisson mask approach described in Section 2.4 achieved an accuracy of 79%, allowing the network to better estimate motion artifact severity compared to using random uniform masking. However, note that we still used random uniform masking in the other experiments of this paper to adhere to the fastMRI-challenge setup.

4 | DISCUSSION AND CONCLUSION

We developed a deep-learning based regressor that can accurately estimate the severity of motion artifacts in undersampled brain MRI scans. To the best of our knowledge, the existing retrospective motion *artifact* estimation approaches require the full set of k-space data to be available.[15-18] We investigated motion artifact severity estimation on accelerated MRI data, which introduces undersampling artifacts on top of the motion artifacts that can have similar appearances and are thus challenging to be distinguished from each other. Our model is able to detect motion corruption during acquisition before all k-space lines are sampled, enabling the possibility to alert the MR technician early, before the scan is fully completed.

We simulated motion-corrupted MRI acquisitions based on uncorrupted MRI data, according to a linear interleaved scanning protocol, though different protocols can be implemented as well. When trained to quantify the severity of motion artifacts on undersampled motion-free data and undersampled synthetically motion-corrupted data, the model was able to separate data from the two classes with an accuracy of over

93%. While this regressor was trained exclusively on retrospectively motion-corrupted data, it was still able to distinguish prospectively motion-corrupted nodding data from still data 91% of the time, indicating that our motion-corruption framework generalized well to real world data. The regressor ordered the artifact score labels (Good/Medium/Bad) correctly 98% of the time for each set of scans of a given subject. When compared against VGG19, we found that our architecture performed better, potentially because the higher number of parameters of VGG19 makes it more susceptible to overfitting.

When investigating different sampling strategies, the use of a particular Poisson disk sampling pattern improved the performance compared to using random uniform sampling. We believe that using the same sampling pattern between all training and testing cases caused the undersampling artifacts to be more predictable, making it easier to distinguish them from motion artifacts. However, using such a single sampling mask instance will introduce a bias, when a different sampling scheme would be used, that is, a compromise between generalizability and performance.

Presence of motion severely impacts reconstructions of fully sampled data by a regular AI-based approach. By training the reconstruction model also on motion-corrupted cases, it learns to deviate from the measured k-space, which can partially alleviate the negative effects of motion and improve reconstruction quality on motion-corrupted data. However, the learned data consistency that allows the model to deviate from the measured k-space and thereby compensate for the motion, also reduces how strongly the reconstruction is based on the measured data when no motion is present and will therefore negatively impact reconstructions of still data.

We investigated a way to leverage the advantages of both the standard and motion-robust reconstruction approaches, by creating a motion-adaptive reconstruction framework based on a model selection mechanism that does not compromise on quality or data consistency for cases without motion. This is done by selecting the optimal reconstruction network based on our motion artifact severity estimator. One can also imagine a use-case where a motion-robust approach performs as well as a regular approach, but is much more computationally expensive. In such case, the selector can be a time saving measure by only performing motion-robust reconstruction on cases with visible motion artifacts. We show a significant improvement in reconstruction performance on a mixed set of motion-free and motion-corrupted data as a result of our selector framework. For motion-corrupted cases on which the regressor made a mistake (i.e., predicted too little motion corruption), the difference in SSIM between the two reconstruction models was on average three times as small compared to when the regressor did correctly estimate much motion corruption. The cases where the difference in SSIM is larger are likely to be high motion cases on which $\mathcal{R}_{\text{still}}$ performs much worse. These high motion cases are easier for the regressor to identify correctly. Thus, classification mistakes by the regressor are most often made on low motion cases where the choice in reconstruction model is not very impactful. This is further substantiated by the fact that the selector achieves close to optimal performance.

A limitation of the used datasets is that the validation set uses raw k-space in vivo data but with simulated motion, and the test set uses in vivo image data with real prospective motion, albeit obtained by instructing participants to move, that is, not natural motion. As basis for the retrospectively motion-corrupted training and validation sets, we used data with a 2D acquisition protocol. We took the anisotropic voxel sizes into account when performing rotation and shift calculations. The lower resolution in the z-direction can make the simulation less realistic, although we believe the effect to not be very detrimental since the performance was good on the 3D test set with real motion. For future work, the performance of 3D and 2D reconstruction models on motion-corrupted data could be compared to investigate the impact of through-plane motion, as through-plane motion requires the 2D reconstruction models to hallucinate to fill the gaps. We also would like to train and evaluate a reconstruction model on prospectively motion-corrupted cases, though a challenge is the availability of data and the fact that the still and motion-corrupted scans need to be perfectly aligned. Another direction for future research could be to more explicitly compensate for motion in k-space by integrating the quantified motion artifacts in the data consistency term of the reconstruction. Further investigations could be towards the effect of different motion patterns on model performance such as finetuning on nodding patterns, and investigating whether it can be impactful to take the effect of motion on the coil sensitivity maps into account during synthetic motion-corruption and motion compensation.

In conclusion, the proposed motion detector showed a very high accuracy on retrospectively as well as prospectively motion-corrupted MRI data, and a motion synthesis framework can be used effectively during training in the common case when no labeled data with real intra-scan motion is available. This enables, among others, the use of our method as a safety mechanism against AI hallucinations, as a prompt for re-scanning, or as a component in a motion-robust reconstruction framework.

ACKNOWLEDGMNTS

This publication is part of the project ROBUST: Trustworthy AI-based Systems for Sustainable Growth with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO), Philips

Research, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023. We thank Christophe Schülke, Elwin de Weerdt, Jakob Meineke and Jochen Keupp for their guidance and constructive discussions.

CONFLICT OF INTEREST STATEMENT

The authors have no relevant conflicts of interest to disclose.

REFERENCES

1. Zaitsev M, Maclaren J, Herbst M. Motion Artefacts in MRI: a complex problem with many partial solutions. *J Magn Reson Imaging: JMRI*. 2015;42:887-901.
2. Andre J, Bresnahan B, Mossa-Basha M, et al. Toward quantifying the prevalence, severity, and cost associated with patient motion during clinical MR examinations. *J Am Coll Radiol: JACR*. 2015;12:689-695.
3. Knoll F, Murrell T, Sriram A, et al. Advancing machine learning for MR image reconstruction with an open competition: overview of the 2019 fastMRI challenge. *Magn Reson Med*. 2020;84:3054-3070.
4. Muckley MJ, Riemenschneider B, Radmanesh A, et al. Results of the 2020 fastMRI Challenge for Machine Learning MR Image Reconstruction. *IEEE Trans Med Imaging*. 2021;40:2306-2317.
5. Maclaren J, Herbst M, Speck O, Zaitsev M. Prospective motion correction in brain imaging: a review. *Magn Reson Med*. 2013;69:621-636.
6. Spieker V, Eichhorn H, Hammernik K, et al. Deep learning for retrospective motion correction in MRI: a comprehensive review. *IEEE Trans Med Imaging*. 2023.
7. Terpstra M, Maspero M, Bruijnen T, Verhoeff J, Lagendijk J, Berg C. Real-time 3D motion estimation from undersampled MRI using multi-resolution neural networks. *Med Phys*. 2021;48:6597-6613.
8. Huttinga N, Berg C, Luijten P, Sbrizzi A. MR-MOTUS: model-based non-rigid motion estimation for MR-guided radiotherapy using a reference image and minimal k-space data. *Phys Med Biol*. 2020;65:015004.
9. Cordero-Grande L, Teixeira RP, Hughes E, Hutter J, Price A, Hajnal J. Sensitivity encoding for aligned multishot magnetic resonance reconstruction. *IEEE Trans Comput Imaging*. 2016;2:266-280.
10. Cordero-Grande L, Ferrazzi G, Teixeira RPAG, et al. Motion corrected MRI with DISORDER: distributed and incoherent sample orders for reconstruction deblurring using encoding redundancy. *Magn Reson Med*. 2020;84:713-726.
11. Hossbach J, Splitthoff DN, Cauley S, et al. Deep learning-based motion quantification from k-space for fast model-based magnetic resonance imaging motion correction. *Med Phys*. 2023;50:2148-2161.
12. Haskell MW, Cauley SF, Bilgiç B, et al. Network accelerated motion estimation and reduction (NAMER): convolutional neural network guided retrospective motion correction using a separable motion model. *Magn Reson Med*. 2019;82:1452-1461.
13. Todd N, Josephs O, Callaghan MF, Lutti A, Weiskopf N. Prospective motion correction of 3D echo-planar imaging data for functional MRI using optical tracking. *NeuroImage*. 2015;113:1-12.
14. Castella R, Arn L, Dupuis E, et al. Controlling motion artefact levels in MR images by suspending data acquisition during periods of head motion. *Magn Reson Med*. 2018;80:2415-2426.
15. Fantini I, Rittner L, Yasuda C, Lotufo R. Automatic detection of motion artifacts on MRI using Deep CNN. In: *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. 2018:1-4.
16. Sciarra A, Chatterjee S, Dünnwald M, et al. Reference-less SSIM Regression for Detection and Quantification of Motion Artefacts in Brain MRIs. 2022.
17. Iglesias JE, Lerma-Usabiaga G, García-Peraza-Herrera LC, Martinez S, Paz-Alonso PM. Retrospective Head Motion Estimation in Structural Brain MRI with 3D CNNs. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2017.
18. Küstner T, Liebgott A, Mauch L, et al. Automated reference-free detection of motion artifacts in magnetic resonance images. *MAGMA*. 2018;31:243-256.
19. Simonyan K, Zisserman A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. ICLR; 2015.
20. Pezzotti N, Yousefi S, Elmahdy MS, et al. An adaptive intelligence algorithm for undersampled knee MRI reconstruction. *IEEE Access*. 2020;8:204825-204838.
21. Zhang J, Ghanem B. ISTA-Net: iterative shrinkage-thresholding algorithm inspired deep network for image compressive sensing. *Proc IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2018.
22. Eichhorn H, Frost R, Kongsgaard A, et al. *Evaluating the performance of markerless prospective motion correction and selective reacquisition in a general clinical protocol for brain MRI*. PsyArXiv; 2022.
23. Bridson R. Fast poisson disk sampling in arbitrary dimensions. In: *ACM SIGGRAPH 2007 Sketches*. SIGGRAPH '07. Association for Computing Machinery; 2007:22-es.
24. Knoll F, Zbontar J, Sriram A, et al. FastMRI: a publicly available raw k-Space and DICOM Dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiol Artif Intell*. 2020;2:e190007.
25. Nárai Á, Hermann P, Auer T, et al. Movement-related artefacts (MR-ART) dataset of matched motion-corrupted and clean structural MRI brain scans. *Sci Data*. 2022;9:630.

How to cite this article: Beljaards L, Pezzotti N, Rao C, Doneva M, van Osch MJP, Staring M. AI-based motion artifact severity estimation in undersampled MRI allowing for selection of appropriate reconstruction models. *Med Phys*. 2024;1-11. https://doi.org/10.1002/mp.16918