Reinforcement Learning for Sampling on Temporal Medical Imaging Sequences

Zhishen Huang
Amazon
Seattle, WA 98109
zhishen.huang@colorado.edu

Abstract

Accelerated magnetic resonance imaging resorts to either Fourier-domain subsampling or better reconstruction algorithms to deal with fewer measurements while still generating medical images of high quality. Determining the optimal sampling strategy given a fixed reconstruction protocol often has combinatorial complexity. In this work, we apply double deep Q-learning and REINFORCE algorithms to learn the sampling strategy for dynamic image reconstruction. We consider the data in the format of time series, and the reconstruction method is a pre-trained autoencoder-typed neural network (U-Net). We present a proof of concept that reinforcement learning algorithms are effective to discover the optimal sampling pattern which underlies the pre-trained reconstructor network (i.e., the dynamics in the environment). The code for replicating experiments can be found at https://github.com/zhishenhuang/RLsamp.

1 Introduction

Magnetic resonance imaging (MRI) is a non-radioactive, non-invasive medical imaging process to visualize human organs and tissues for diagnostic purposes. MRI scanners sequentially collect measurements in the frequency domain (or k-space), from which an image is reconstructed. A central challenge in MRI is its time-consuming sequential acquisition process as the scanner needs to densely sample the underlying k-space for accurate reconstruction, and during the scanning process patients are expected to remain still. The physical nature of MRI process makes it difficult to image objects in motion such as hearts, compromises image quality by motion artifacts, and incurs patients' discomfort and safety concerns. Hence, reconstructing high-quality images from limited measurements is desirable. There are two core parts in the accelerated MRI pipeline: a sampling pattern deployed to collect the (e.g., limited/undersampled) data in k-space and a corresponding reconstruction method (reconstructor) that also enables recovering any missing information. Reconstruction methods can be motivated by mathematical theory of compressed sensing [5, 4] or data-driven methodology. Several deep learning based approaches [17, 16, 13, 9] have been shown to demonstrate high reconstruction quality and robust performance to various sources of artifacts and hyperparameter settings.

The experimental design of subsampling schemes is more difficult to determine due to the combinatorial nature of this optimization problem. For a set of images $\{\mathbf{x}_i\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^{m \times n}$, we aim to find the optimal sampling pattern in the k-space to cater for the best reconstruction outcome. Let $\mathcal{S}$ be the operation of applying a sampling mask on the k-space information, $\mathcal{G}$ be the operator of reconstructing the image from k-space measurements, and $\mathcal{F}[\mathbf{x}]$ be the 2D-Fourier transform. With

a choice of loss function $\mathcal{L}(\cdot, \cdot)$, the optimization problem is formulated as

$$\min_{\mathcal{S}} \sum_{i=1}^{N} \mathcal{L}(\mathcal{G} \circ \mathcal{S} \circ \mathcal{F}[\mathbf{x}_i], \mathbf{x}_i)$$

$$\text{s.t. } \mathcal{S} \text{ satisfies designated sampling budget} \tag{1.1}$$

A general sampling operator $\mathcal{S}$ can vary with respect to each input image $\mathbf{x}_i$. As a simple example, one family of sampling pattern for medical images is one-dimensional Cartesian sampling in k-space. With one-dimensional sampling mask, the sampling takes place with respect to one dimension of the input image and meanwhile the other dimension is fully sampled. This sampling pattern can be coded into MRI machines, and its performance is corroborated by the compressed sensing theory. If the sampling happens to the second dimension, then $\mathcal{S} \circ \mathcal{F}[\mathbf{x}] = \text{FFT}[\mathbf{x}] \cdot \text{diag}(\mathbf{m})$, where $\mathbf{m} \in \{0,1\}^n$ is a binary vector, $\text{diag}(\mathbf{m})$ is an $n \times n$ matrix with $\mathbf{m}$ on its diagonal. The sampling budget in this case can be phrased as $\sum_{i=1}^{n} m_i \leq b$ for some $b \in \mathbb{N}^+$.

In this work, we consider the data in the form of a time series $\{\mathbf{x}_t\}$. We aim to build a sampling policy for a series of images that contains time evolution information. Specifically, the sampling policy is realized through a policy function $\pi(\cdot | \widehat{\mathbf{x}}_{t_k-\alpha : t_k})$ given reconstruction history in the recent α steps. As t increases, the sampling policy π informs the next frequency location to sample. This is different from precedent study on adaptive sampling where the target under investigation is static.
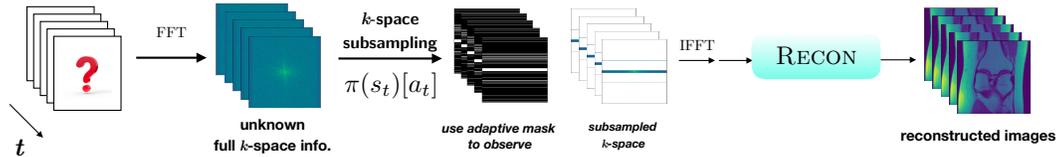
Figure 1: Experiment pipeline: sampling and reconstruction

We leverage deep Q-learning [22] and policy gradient method [21] respectively to show that a sampling policy for collecting limited measurements in the k-space adapted to the underlying image reconstruction method and training dataset can be learned. The major novelty of this work is that the sampling policy is built with respect to a time series of images (such as a serial heartbeat) instead of a set of static images. We also point out that this study is mostly a proof of concept that in the MRI application, RL is effective in learning the sampling strategy that optimizes the performance of the reconstructor (the environment).

1.1 Related work

Early attempts to learn data-adaptive sampling patterns mostly revolve around static target objects and the parameters of the sampling pattern are jointly updated along with those of reconstructors. [25] considers one-dimensional Cartesian masks and uses a neural network to evaluate the importance of each unseen row in the frequency domain, so that at inference time, the evaluator network selects the row with highest importance score to acquire. [10] encodes sampling patterns in a set of distributions, each of which shares the same dimension as the subsampled dimension and is parametrized by a logit. A categorical distribution governs the applicability of each sampling distribution, and all logits and the categorical distribution hyperparameters are co-trained with unrolled ISTA reconstructors [7]. This is similar to the LOUPE approach [2], where the sampling mask is directly parameterized as Bernoulli random variables, and [24], which uses a careful thresholding operation to handle gradient updates on parameters characterizing one-dimensional Cartesian masks. [8] introduces an alternating training framework to co-train the reconstructor and a mask predictor to achieve object-level adaptiveness with more explicit control over the mask predictor training process.

The sequential decision making approach has been adopted to solve the optimal sampling problem in recent studies. [3] explore the policy gradient method to train a policy network that guides a one-dimensional sampling choice with respect to pre-trained reconstructors which are tuned to denoise aliased images blurred by uniformly random sampling masks. [15] have applied double Q-learning on constructing adaptive sampling policies for static magnetic resonance images, and the reconstructor in their study is pre-trained in the same manner as in [3]. [11] leverages Monte Carlo tree search

to explore mask options, and trains the policy network with tree search selections. Meanwhile, the reconstructor network is alternatively updated along with the policy network in the training process.

2 Methodology

We formalize the sequential selection of the sampling mask for MRI as a partially observable Markov decision process (POMDP). The state at acquisition step t of this POMDP corresponds to the tuple consisting of recent α reconstruction results $s_t = (\widehat{\mathbf{x}}_t, \cdots, \widehat{\mathbf{x}}_{t-\alpha})$ and the history of actions (i.e., the current sampling mask $\mathbf{M}_{t-1}$ for the time series of images). An action a_t is to select a particular k-space frequency (row or column) to observe. After an action is taken, the sampling mask $\mathbf{M}_t$ underlying the measurement operation $\mathcal{S}_t$ has the a_t-th diagonal element changed from 0 to 1 compared to $\mathbf{M}_{t-1}$ for $\mathcal{S}_{t-1}$. The reward at each acquisition step is $r_t(s_t, a_t) = \max\big(\mathcal{L}(\mathcal{G} \circ \mathcal{S}_t \circ \mathcal{F}[\mathbf{x}_t], \mathbf{x}_t) - \mathcal{L}(\mathcal{G} \circ \mathcal{S}_{t-1} \circ \mathcal{F}[\mathbf{x}_t], \mathbf{x}_t), 0\big)$.

We consider two RL algorithms to build policies in this work: deep double Q-learning and the policy gradient method. Both of these two policies are motivated from the Bellman optimality condition while taking different approximations.

2.1 Deep double Q-learning (DDQN)

In deep Q-learning, we use a neural network as an evaluating function to determine the value of each action given the current state. A greedy policy ensues by selecting the action with the largest estimated value in the action space. The double Q-learning framework keeps two copies of the Q-network during the training process: one $Q_{\boldsymbol{\theta}}$ for online action selection while exploring new trajectories, the other, the target network Q_{eval}, is the same as the online network except that its parameters are copied every τ steps from the online network and kept fixed on all other steps.

We have used the experience replay technique to store observed state transitions and corresponding rewards in memory $\mathcal{M}$. The target network is used for offline evaluation of memorized states. The training of deep Q learning is set up as a self-supervised problem where the loss function is to minimize the temporal difference error over data sampled from a replay memory buffer. Specifically, the loss function to minimize is formulated as

$$L(\boldsymbol{\theta}) = \sum_{\substack{(r_{t+1}, s_{t+1}, s_t, a_t) \\ \in \mathcal{M}}} \big\| \underbrace{R_{t+1} + \gamma Q_{\text{eval}}(s_{t+1})[a_{\text{eval}}^*]}_{\text{target}} - \underbrace{Q_{\boldsymbol{\theta}}(s_t)[a_t]}_{\text{online}} \big\|_1 \qquad (2.1)$$
$$\text{s.t. } a_{\text{eval}}^* = \arg\max_a Q_{\text{eval}}(s_{t+1})[a]$$

2.2 Policy gradient method

Policy gradient methods aim to directly maximize the expected return of a policy parametrized by $\boldsymbol{\theta}$ on the POMDP. In this study, we use the REINFORCE algorithm [21] to train a deep policy network that maximizes the expected sum of rewards given the designated sampling budget.

In the REINFORCE algorithm, we first collect a whole sequence of state-action pairs up to the time horizon T in a sampling episode, then compute the Q-value $q(s, a)$ based on the partial aggregation of observed state-action data, and finally construct a gradient-based update to policy parameters $\boldsymbol{\theta}$. With the log-ratio trick [21], we maximize the following utility functions:

$$J(\boldsymbol{\theta}) = \sum_{t=0}^{T-1} \log \pi_{\boldsymbol{\theta}}(s_t)[a_t] \cdot G_t = \sum_{t=0}^{T-1} \log \pi_{\boldsymbol{\theta}}(s_t)[a_t] \cdot \bigg(\sum_{k=t+1}^{T} \gamma^{k-1} r_k \bigg), \qquad (2.2)$$

where $a_t = \arg\max_a \pi_{\boldsymbol{\theta}}(s_t)$.

3 Experiments

3.1 Dataset

We test the RL sampling schemes on the OCMR data [6] which is an open cardiovascular MRI dataset containing multiple cardiac cine series. We use 62 series for training, and 32 series for

testing. Each series in the selected OCMR data is fully sampled, and every frame is dimension 384×144. We aggregate all coils so that the dataset in this study can be considered as single-coiled. The length of each series in the training set varies between 16 and 28, and in the testing set between 16 and 27. There are in total 1519 frames in the training time series and 642 frames in the testing time series.

3.2 Pre-trained reconstructors

A modest but viable option of $\mathcal{G}$ is the Unet [18], which was originally proposed for image segmentation but later was shown to be useful for reconstruction/de-aliasing purposes. To train a Unet in a supervised way as a reconstructor, one needs to set up a training dataset where the aliased images are generated in a certain way. As mentioned in the Related Work section, precedent work on building sampling policies relies on pre-trained reconstructors, where the aliased images in the training dataset are subsampled by uniformly random masks in the k-space.

In this work, we consider 1D sampling strategies in experiments. We define the *uniformly random* mask as a sampling pattern where the lowest b frequencies in the dimension to be subsampled are always collected while the rest of the sampling budget are uniformly at random allocated across the unobserved high-frequency domain. Here b is a hyperparameter, given the sampling budget. We define the *low-frequency* masks as the sampling pattern where all sampling budget is allocated to the lowest frequencies in the dimension to be subsampled. We define the *energy-distribution-based* masks as the sampling pattern where the lowest b frequencies are always sampled, while the rest of the sampling budget are used according to the energy distribution of each high frequency in the training images.

Before getting into the policy building step, we train three Unet reconstructors with respect to three separate dataset whose aliased images are generated by each of the aforementioned masks respectively. Every training image pair in the dataset is in the format of (aliased image $\widetilde{\mathbf{x}}_i^{\Xi}$, corresponding ground truth $\mathbf{x}_i$), where $\widetilde{\mathbf{x}}_i^{\Xi} = \text{IFFT} \circ \mathcal{S}_{\Xi} \circ \text{FFT}[\mathbf{x}_i]$. Each pre-trained Unet reconstructor intends to de-alias the crude IFFT images blurred by a particular sampling scheme Ξ: $\mathcal{G}_{\Xi}(\widetilde{\mathbf{x}}_i^{\Xi})$. We point out that unlike static dataset where energy-distribution-based masks usually outperforms uniformly random masks and low-frequency masks [8], with respect to the OCMR dataset, the pre-trained reconstructors perform the best with the low-frequency mask in terms of both normalized rooted mean squared error (NRMSE) and the structural similarity index measure (SSIM) [23] as Table 1 shows.

Subsampling schemes Ξ	NRMSE:$= \frac{\|\mathbf{x}_{\text{recon}} - \mathbf{x}_*\|_2}{\|\mathbf{x}_*\|_2}$	SSIM
Uniformly random masks	0.6884	0.4380
Low-frequency masks	**0.3668**	**0.7456**
Energy-distribution-based masks	0.6323	0.4816

Table 1: Performances of pre-trained reconstructors with respect to aliased images which undergo different sampling schemes. The acceleration ratio is 6-fold with the $b = 8$ lowest base frequencies to start and 16 high frequencies to sample. The full size of the subsampled image dimension is 144.

3.3 Problem setup

In this work, we aim to solve the following optimization problem on input time series data $\{\mathbf{x}_i\}_{i=1}^T$:

$$\min_{\mathbf{m}} \sum_{i=1}^T \mathcal{L}(\mathcal{G}_{\Xi} \circ \mathcal{S}_{\text{diag}(\mathbf{m})} \circ \mathcal{F}[\mathbf{x}_i], \mathbf{x}_i)$$

$$\text{s.t.} \begin{cases} \mathcal{S}_{\text{diag}(\mathbf{m})}[\mathbf{y}] = \mathbf{y} \cdot \text{diag}(\mathbf{m}) \\ \sum_{i=1}^n m_i \leq b + s \\ m_i \in \{0, 1\} \end{cases} \tag{3.1}$$

where $\mathcal{G}_{\Xi}$ is a pretrained Unet with respect to mask pattern Ξ listed in Table 1, s is a pre-designated sampling budget, $\mathbf{y} = \mathcal{F}[\mathbf{x}_i]$ and $\mathbf{y} \in \mathbb{C}^{m \times n}$, $\text{diag}(\mathbf{m})$ is an $n \times n$ diagonal matrix with vector $\mathbf{m}$ in its diagonal, and the loss function is $\mathcal{L}(\mathbf{x}, \mathbf{x}_*) = \frac{\|\mathbf{x} - \mathbf{x}_*\|_2}{\|\mathbf{x}_*\|_2}$.

3.4 Hyperparameter settings

Acceleration ratio In accelerated MRI problem, we define the *acceleration ratio* as $\frac{\text{the full size of } k\text{-space dimension to be subsampled}}{\text{the actual amount of frequencies to observe}}$. In this study, we focus on 6-fold acceleration case, where there are 8 lowest base frequencies to start with in the observation set and we continue to sample additional 16 high frequencies to complete an episode of measurement before applying the reconstructor. The total size of the full k-space is 384×144, where the second dimension is to be sampled.

Pre-trained reconstructors The Unet we use as reconstructor has 6 downsampling and upsampling layers, and has two channels of input, each of which corresponds to the real and imaginary part of the data collected in the Fourier domain . The loss function we use to train the Unet is $\frac{\|\mathbf{x}_{\text{recon}} - \mathbf{x}_*\|_2}{\|\mathbf{x}_*\|_2} + 5 \cdot \left(1 - \text{SSIM}(\mathbf{x}_{\text{recon}}, \mathbf{x}_*)\right)$. We collect each image slice in the time series as the ground truth in the training dataset for sampling policy building, and apply various sampling patterns to generate corresponding aliased images. For reconstructor tuning, with respect to each training dataset generated by different sampling patterns $\boldsymbol{\Xi}$, we train the Unet for 300 epochs, with learning rate as 10^{-5}, ℓ_2 regularization weight 10^{-5} in the Adam optimizer, and the batchsize 10.

Policy network We use a convolutional neural network coupled with fully connected feed-forward layers as the policy network. The policy network sees the reconstruction images in the previous 8 time steps, and output the probability distribution on the remaining unobserved high frequencies to sample.

DDQN algorithm setup We set the size of the memory buffer as 100. We use the Adam optimizer with 10^{-5} as the constant learning rate. We run DDQN training for 600 epochs with the discount factor set as 0.8 and batch size 3.

REINFORCE algorithm setup We add an *entropy regularization* term [12] to the loss function in order to encourage the learning process to explore larger trajectory space, and we set the weight of this penalty term to be 0.01. We use the Adam optimizer with 10^{-5} as the constant learning rate. We run REINFORCE training for 600 epochs with discount factor set as 0.1 and batch size 3.

3.5 Validation for the proof of concept

3.5.1 Convergence results

The following figures show the convergence of the reinforcement learning training paradigm. Figures 2a and 2b show that the masks predicted by the policy network trained by DDQN and REINFORCE, respectively, demonstrate improved performance in terms of the image reconstruction error compared to the fixed pre-trained reconstructor. Both DDQN and REINFORCE can recover the sampling scheme under which the reconstructor is trained. In this case, the underlying sampling scheme is low-frequency sampling policy.

Figures 3 and 4 show the evolution of predicted masks returned by the policy network. Both policies converge to the low-frequency sampling policy, while the DDQN paradigm takes more iterations to discover ultimate target.

Subsampling schemes/Policies	NRMSE$:= \frac{\|\mathbf{x}_{\text{recon}} - \mathbf{x}_*\|_2}{\|\mathbf{x}_*\|_2}$	SSIM
DDQN trained policy	0.3948	**0.7522**
REINFORCE trained policy	**0.3943**	0.7521
Uniformly random masks	0.5638	0.5793
Low-frequency masks	0.3948	**0.7522**
Energy-based masks	0.6160	0.4825

Table 2: Reconstruction accuracy on the testing dataset by different policies. Note that the set of testing images are different from that in Table 1 in the sense that only frames in time series after the initial base observations are taken into account. In other words, for each time series in a test file, we consider reconstructions on frames $\mathbf{x}_{\text{base}} + 1, \cdots, \mathbf{x}_{\text{end}}$ where "base" is the count of base frequencies we use as basic observations.

(a) DDQN RMSE convergence curve **(b)** REINFORCE RMSE convergence curve
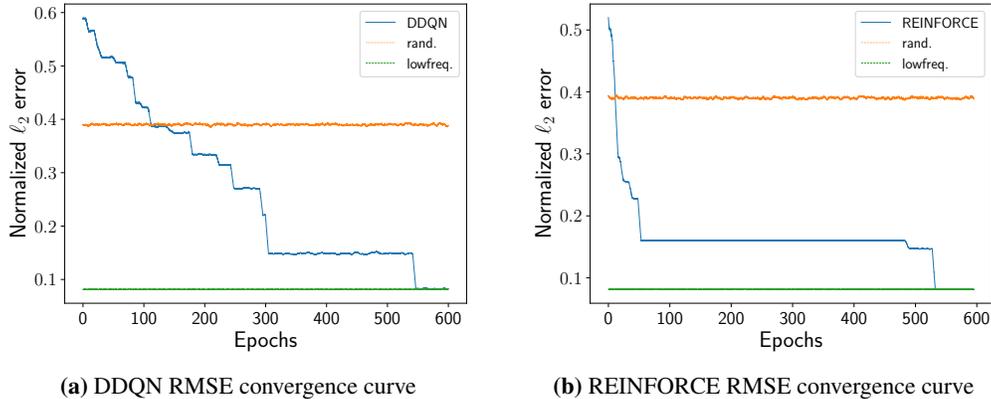
Figure 2: Convergence curve of image reconstruction error (normalized ℓ_2 error) during training.
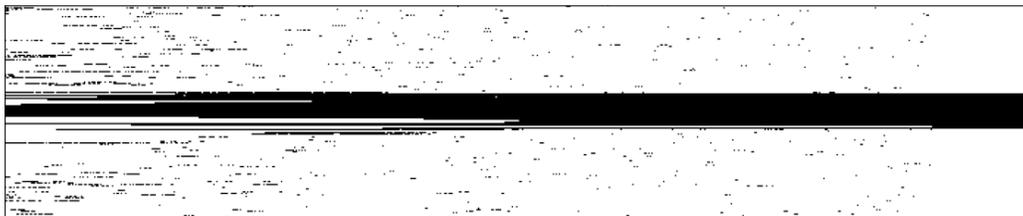
Figure 3: Mask evolution history as DDQN training proceeds. The horizontal axis shows the direction of time, and the vertical axis is the frequency grid of the subsampled dimension in k-space. We put the low frequency in the middle of the vertical axis and the high frequencies on the upper and lower end of the vertical axis. Each vertical cross section is a sampling mask at a given iteration step returned by the policy network trained through DDQN algorithm.

3.5.2 Reconstruction examples

In Figure 5, we show a set of reconstruction examples in the testing set as sampling proceeds. The time series characterizes a cardiac cycle where the diastole is followed by the systole. We show the ground truth images in the cardiac cycle, their corresponding reconstruction images, and error maps. When generating reconstructions on the testing time sequences, we first assemble the sampling masks according to policies under comparison and the designated sampling budget before applying corresponding reconstructors. After masks are generated regarding a testing time series, they remain fixed for the rest of frames in that time series. The sampling masks used for measurement in k-space starts from the base $b = 8$ low frequencies, and garner additional $s = 16$ high frequencies according to sampling policies. The mask used for each frame in Figure 5 consists of both base low frequencies and policy-informed high frequencies.

4 Conclusion

In this work, we present a synthetic study where with prior knowledge on the performance of candidate reconstructors, reinforcement learning techniques can effectively uncover the underlying mask patterns with which the reconstructor is trained to denoise undersampled dynamic cardiac cine sequences. This study also shows that such dynamic MRI datasets like OCMR data lead to new challenges as the conventional reconstructors and energy-based sampling schemes do not demonstrate the same performance as in the static setting.

Future work   Several aspects of using RL algorithms can be further investigated:

1. How can we accelerate the training process or improve the reconstruction quality through more sophisticated RL optimization practice [19] or leverage prior information of data distribution in the training process?

6
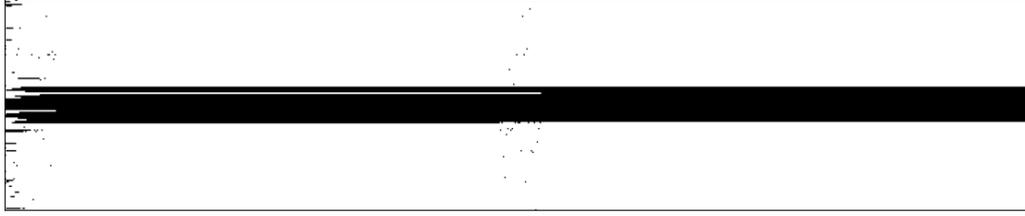
Figure 4: Mask evolution history as REINFORCE training proceeds. The horizontal axis shows the direction of time elapses, and the vertical axis is the frequency grid of the subsampled dimension in k-space. We put the low frequency in the middle of the vertical axis and the high frequencies on the upper and lower end of the vertical axis. Each vertical cross section is a sampling mask at a given iteration step returned by the policy network trained through REINFORCE algorithm.

2. Do reconstructors such as MoDL [1] or unrolling based reconstructors [14, 20] with higher complexity improve the reconstruction accuracy for the dynamic dataset?

3. How can we develop a self-supervised training process to either simultaneously or alternatively train the policy network and the reconstructor network such that the final reconstruction performance can outperform the starting baseline?

Acknowledgments and Disclosure of Funding

References

[1] Hemant K. Aggarwal, Merry P. Mani, and Mathews Jacob. MoDL: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, 2019.

[2] Cagla D. Bahadir, Alan Q. Wang, Adrian V. Dalca, and Mert R. Sabuncu. Deep-learning-based optimization of the under-sampling pattern in MRI. *IEEE Transactions on Computational Imaging*, 6:1139–1152, 2020.

[3] Tim Bakker, Herke van Hoof, and Max Welling. Experimental design for MRI by greedy policy search. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18954–18966. Curran Associates, Inc., 2020.

[4] Stephen Becker. *Practical Compressed Sensing: Modern data acquisition and signal processing*. PhD thesis, California Institute of Technology, 2011.

[5] Emmanuel J. Candes and Michael B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[6] Chong Chen, Yingmin Liu, Philip Schniter, Matthew Tong, Karolina Zareba, Orlando Simonetti, Lee Potter, and Rizwan Ahmad. OCMR (v1.0)–open-access multi-coil k-space dataset for cardiovascular magnetic resonance imaging, 2020.

[7] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, ICML 2010 - Proceedings, 27th International Conference on Machine Learning, pages 399–406, 2010. 27th International Conference on Machine Learning, ICML 2010 ; Conference date: 21-06-2010 Through 25-06-2010.

[8] Zhishen Huang and Saiprasad Ravishankar. Single-pass object-adaptive data undersampling and reconstruction for MRI. *IEEE Transactions on Computational Imaging*, 8:333–345, 2022.
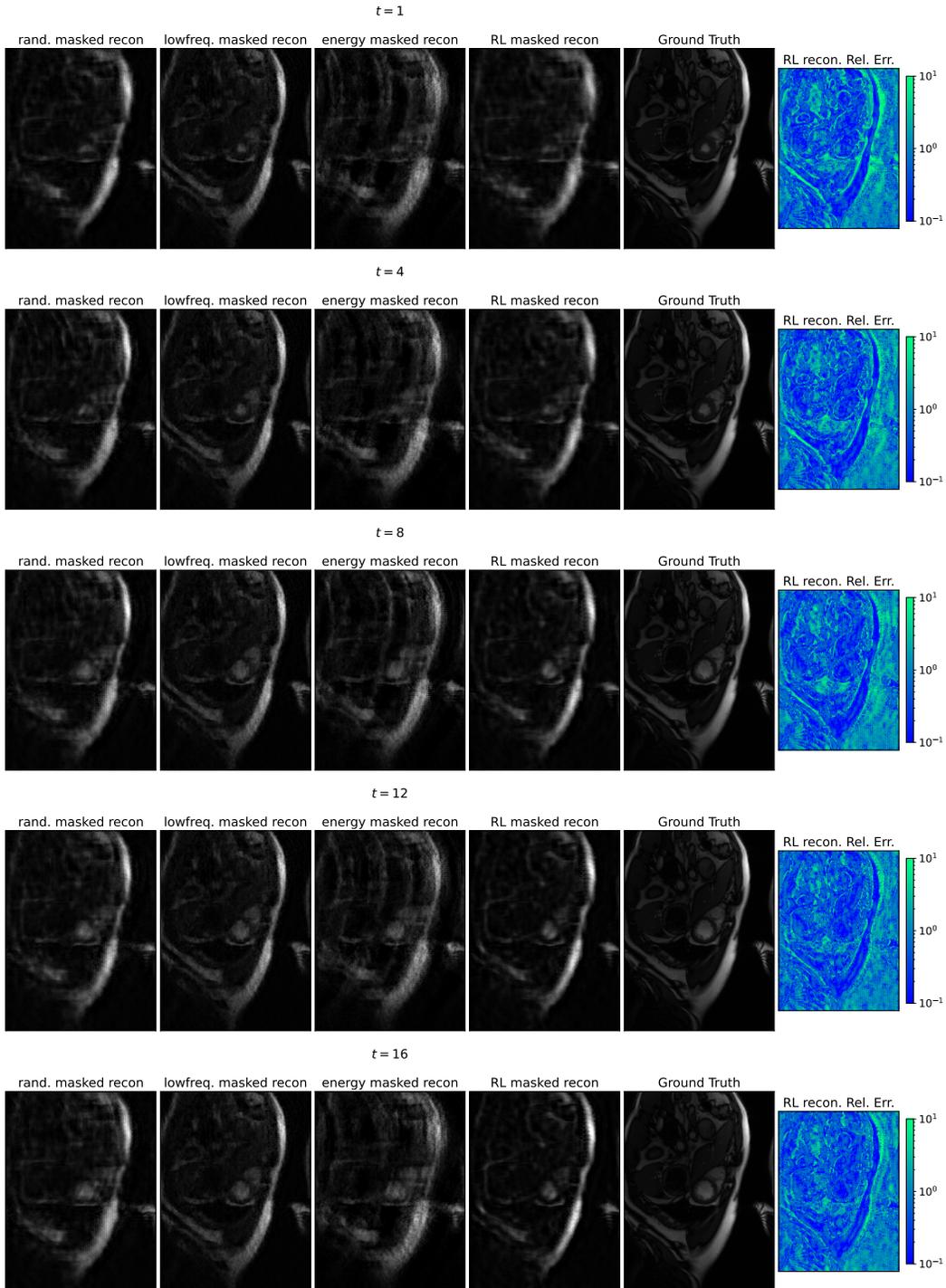
Figure 5: An example for a cardiac cycle that contains reconstruction (left column), ground truth (central column) and error maps (right column).

[9] Zhishen Huang, Siqi Ye, Michael T. McCann, and Saiprasad Ravishankar. Model-based reconstruction with learning: From unsupervised to supervised and beyond, 2021.

[10] Iris A.M. Huijben, Bastiaan S. Veeling, and Ruud J.G. van Sloun. Learning sampling and model-based signal recovery for compressed sensing MRI. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8906–8910, 2020.

[11] Kyong Hwan Jin, Michael Unser, and Kwang Moo Yi. Self-supervised deep active accelerated MRI. *CoRR*, abs/1901.04547, 2019.

[12] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR.

[13] Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

[14] Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020.

[15] Luis Pineda, Sumana Basu, Adriana Romero, Roberto Calandra, and Michal Drozdzal. Active MR k-space sampling with reinforcement learning. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 23–33, Cham, 2020. Springer International Publishing.

[16] Saiprasad Ravishankar, Bihan Wen, and Yoram Bresler. Online sparsifying transform learning—part I: Algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):625–636, 2015.

[17] Saiprasad Ravishankar, Jong Chul Ye, and Jeffrey A. Fessler. Image reconstruction: From sparsity to data-adaptive methods and machine learning. *Proceedings of the IEEE*, 108(1):86–109, 2020.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[19] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 07–09 Jul 2015. PMLR.

[20] Nir Shlezinger, Jay Whang, Yonina C. Eldar, and Alexandros G. Dimakis. Model-based deep learning. *Proceedings of the IEEE*, pages 1–35, 2023.

[21] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018.

[22] Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[23] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[24] Tomer Weiss, Sanketh Vedula, Ortal Senouf, Oleg Michailovich, Michael Zibulevsky, and Alex Bronstein. Joint learning of cartesian under sampling andre construction for accelerated MRI. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8653–8657, 2020.

[25] Zizhao Zhang, Adriana Romero, Matthew J. Muckley, Pascal Vincent, Lin Yang, and Michal Drozdzal. Reducing uncertainty in undersampled MRI reconstruction with active acquisition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2049–2053, 2019.