LATENT-SPACE DISENTANGLEMENT WITH UNTRAINED GENERATOR NETWORKS ALLOWS TO ISOLATE DIFFERENT MOTION TYPES IN VIDEO DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Isolating different types of motion in video data is a highly relevant problem in video analysis. Applications can be found, for example, in dynamic medical or biological imaging, where the analysis and further processing of the dynamics of interest is often complicated by additional, unwanted dynamics, such as motion of the measurement subject. In this work, it is shown that a representation of video data via untrained generator networks, together with a specific technique for latent space disentanglement that uses minimal, one-dimensional information on some of the underlying dynamics, allows to efficiently isolate different, highly non-linear motion types. In particular, such a representation allows to freeze any selection of motion types, and to obtain accurate independent representations of other dynamics of interest. Obtaining such a representation does not require any pre-training on a training data set, i.e., all parameters of the generator network are learned directly from a single video.

1 INTRODUCTION

Processing motion information in a time series of images is a classical but still very active research topic in computer vision and computational imaging, with a plethora of applications ranging from autonomous driving to biological and medical imaging. In this context, one can separate three (strongly interconnected) directions of research: i) Motion reconstruction, which aims at reconstructing dynamic image data from incomplete or indirect measurements, with applications for instance in dynamic magnetic resonance (MR) imaging or dynamic positron-emission tomography (PET), see Otazo et al. (2015); Bustin et al. (2020); Rahmim et al. (2009) for examples. ii) Motion estimation, which aims to estimate and represent motion between different frames. This is one of the most classical problems in computer vision and often addressed, for instance, via optical flow estimation, see for instance Fortun et al. (2015) for a review. iii) Motion correction, which aims to correct for motion, typically via registration techniques. The latter are again a well-established but still very active research topic, in particular in the context of medical imaging such as MR imaging, PET, computed tomography (CT) Oliveira & Tavares (2014); Kyme & Fulton (2021) but also in applications like dynamic fluorescence microscopy Kumar et al. (2013).

Of course, this separation into three direction of research is somewhat artificial and close connections between them exist: Having estimated motion fields available is crucial for motion correction, motion correction is strongly interconnected with image reconstruction, and adequately reconstructed time series data is the basis of classical motion estimate techniques.

In the past years, deep learning based methods have enabled a significant progress in all of the above research direction connected to motion in time series of images, see Bustin et al. (2020), Tu et al. (2019) and Fu et al. (2020) for recent review papers on deep learning techniques for reconstruction, motion estimation and motion correction, respectively.

A task that is related but still different to the above-described research directions in time series analysis is the *isolation* of different types of motion. By this, we mean the following general problem setting: Given a video with different types of motion, synthesize a new video that does not show all types of motion, but only a subset of motion-types that is relevant for further processing. A generic area of applications where this problem is relevant is medical imaging, e.g., MR imaging or

PET. Here, a concrete example is a video showing cardiac motion together with additional motion resulting from breathing or patient movement, and the goal is to obtain a video showing only the isolated cardiac motion for further analysis.

While, at first glance, the task of isolating motion is strongly related to image registration techniques, the latter do not apply here since, even in the case where motion fields that register each frame of the time series to a representative template are available, it is still a highly non-trivial problem to decompose such motion fields into different components corresponding to different types of motion. As consequence, to the best knowledge of the authors, a generic method capable of isolating different types of motion in video data does not exist so far.

The goal of this paper is to show that untrained generator networks, together with a specific technique for latent space disentanglement, can fill this gap. More specifically, we consider the optimization of a generator network to represent a given time series of images, where different latent space variables are forced to independently explain the different types of motion. The latter is achieved by incorporating one-dimensional information on all but one of the different motion types present in the video. As we show in our numerical experiments, obtaining such a representation of a time series of images allows, in a second step, to freeze any selection of motion types, and to obtain accurate independent representations of the other dynamics of interest.

The use of untrained generator networks for image representation was popularized by Ulyanov et al. (2020), which also partially inspired our work. Since the appearance of Ulyanov et al. (2020), many works employing the deep image prior for image reconstruction in various applications Gong et al. (2018); Baguer et al. (2020) as well as works on the analysis of this type of approaches Heckel & Soltanolkotabi (2019); Jagatap & Hegde (2019); Dittmer et al. (2020); Habring & Holler (2022) have been published. More recently, also works that employ the deep image prior for representation and reconstruction of dynamic MR data appeared, see for instance Hyder & Asif (2020); Yoo et al. (2021). Existing works, however, focus on reconstruction and do not employ a specific latent-space disentanglement as proposed here, which is main ingredient for not only representing but also isolating different motion types.

Latent space disentanglement is in turn an active research topic in the context of GANs, see for instance Chen et al. (2016) for a seminal work on using latent space disentanglement to learn interpretable representations, Tulyakov et al. (2018) for a work on decomposing motion and content in videos, and Liu et al. (2022) for a work that employs latent space disentanglement for semantic face editing. But again, also in the context of GANs, to the best knowledge of the authors, a method capable of isolating different motion types in video data as the one proposed here does not exist.

As prototypical application of the technique for unsupervised motion isolation introduced in this paper, we provide experiments where we isolate cardiac motion from other motion types such as respiratory motion or motion due patient movement, using both phantom and real MR image data with partially simulated motion. The one dimensional information on some of the underlying motion types in this case corresponds to a scalar describing the cardiac- or breathing state, a signal which in practice can easily be obtained for instance by simultaneous electrocardiogram (ECG) or chest-displacement measurements.

In addition to providing new possibilities for motion isolation, the advantage of the method presented here is its general applicability in different medical imaging modalities an beyond, and the fact that, contrary to classical registration techniques, the approach is rather generic and no explicit modeling of motion types of any form is required. A disadvantage is that the motion isolation is only implicit via latent-space variables, and that explicit motion information in the form of motion fields does not become available with this technique.

2 METHOD

We introduce the proposed approach in a general setting (including possibly indirect observations of the image data) first, and then specify its concrete application to isolating respiratory and cardiac motion in dynamic images that were obtained with MR imaging.

Consider a linear, discretized dynamic imaging inverse problem with data $(y_t)_{t=1}^T \subset \mathbb{R}^M$ and linear operators $A_t \in \mathcal{L}(\mathbb{R}^{N_1 \times N_2}, \mathbb{R}^M)$ of the form

$$y_t = A_t x_t, \quad t = 1, ..., T, \tag{1}$$

where the unknown is a sequence of images $(x_t)_{t=1}^T$ with $x_t \in \mathbb{R}^{N_1 \times N_2}$ for each t. Following the basic idea introduced in Ulyanov et al. (2020) for static images, we consider each frame $x_t \in \mathbb{R}^{N_1 \times N_2}$ to be the output of a generator $x_t = G_\theta(z_t)$, whose parameters we want to learn only from the given data. In particular, we consider generator networks $G_\theta : \mathbb{R}^q \to \mathbb{R}^{N_1 \times N_2}$ of the form

$$G_\theta(z_t) = \theta_L^1 * \sigma_{L-1}(\theta_{L-1}^1 * (...\sigma_1(\theta_1^1 * z_t + \theta_1^2)...) + \theta_{L-1}^2) + \theta_L^2, \tag{2}$$

with time-independent parameters $\theta_j^i \in \Theta$, for $1 \leq i \leq 2$ and $1 \leq j \leq L$ and pointwise nonlinearities $\sigma_1, \ldots, \sigma_{L-1}$. The generator network maps the latent space $\mathbb{R}^q$ to the image (frame) space $\mathbb{R}^{N_1 \times N_2}$. Assuming the dynamic image sequence $(x_t)_{t=1}^T$ contains $m \in \{2, 3, \ldots\}$ independent types of motion, we split the latent variable $z \in \mathbb{R}^q$ into a time-independent part $z^0 \in \mathbb{R}^{q-m}$ and m time-dependent variables $(z_t^i)_{t=1}^T$ with $z_t^i \in \mathbb{R}$, $i = 1, \ldots, m$. We further assume that all the time dependent variables $(z_t^i)_{t=1}^T$ except $(z_t^1)_{t=1}^T$ are given as $(\hat{z}_t^i)_{t=1}^T$ from some one-dimensional a-priori information on the state of the respective types of motion (e.g. from electrocardiograms or chest-displacement measurements).

We then reconstruct the image sequence $(x_t)_{t=1}^T$ together with the network parameters θ and the time-dependent variable $(z_t^1)_{t=1}^T$ via solving

$$((\hat{z}_t^1)_{t=1}^T, \hat{\theta}) \in \arg\min_{(z_t^1)_{t=1}^T, \theta} \frac{1}{T} \sum_{t=1}^T \|y_t - A_t G_\theta(\hat{z}^0, z_t^1, \hat{z}_t^2 \ldots, \hat{z}_t^m)\|_2^2, \tag{3}$$

where $\|\cdot\|_2$ is the Euclidean norm (but can be a more general loss) and $\hat{z}^0 \in \mathbb{R}^{q-m}$ is a randomly initialized, fixed static latent variable. Once a solution is obtained, we do not only obtain the reconstructed images sequence $(\hat{x}_t)_{t=1}^T$ via $\hat{x}_t = G_{\hat{\theta}}(\hat{z}^0, \hat{z}_t^1, \ldots, \hat{z}_t^m)$, but, more importantly, can generate image sequences $(\hat{x}_t^i)_{t=1}^T$ for $i = 1, \ldots, m$, which we expect to contain only the ith type of motion with all others being fixed, via

$$\hat{x}_t^i = G_{\hat{\theta}}(\hat{z}^0, \hat{z}_{h_1}^1, \ldots, \hat{z}_{h_{i-1}}^{i-1}, \hat{z}_t^i, \hat{z}_{h_{i+1}}^{i+1}, \ldots, \hat{z}_{h_m}^m), \tag{4}$$

where $h_1, \ldots, h_{i-1}, h_{i+1}, \ldots, h_m$ are fixed reference frames.

As a concrete example, in this paper we consider the application of this general approach to isolating cardiac motion from respiratory motion in dynamic images obtained from MRI, where one-dimensional information about the respiratory state (e.g. from measurements of the chest displacement) is available. In this case, $m = 2$, and the latent variable $z_t \in \mathbb{R}^q$ at time t is decomposed as $z_t = (z^0, z_t^1, z_t^2)$ with z_t^2 known (and given as $(\hat{z}_t^2)_{t=1}^T$). As our focus is on motion isolation rather than reconstruction, we further assume the reconstructed dynamic image sequence to be available, i.e., A_t is the identity, noting that a generalization of our approach to reconstruction does not pose any conceptual difficulties. In summary, this yields the following optimization problem

$$((\hat{z}_t^1)_{t=1}^T, \hat{\theta}) \in \arg\min_{(z_t^1)_{t=1}^T, \theta} \frac{1}{T} \sum_{t=1}^T \|y_t - G_\theta(\hat{z}^0, z_t^1, \hat{z}_t^2)\|_2^2. \tag{5}$$

Algorithmic strategy To solve the minimization problem (5), we use Pytorch Paszke et al. (2017) and the ADAM optimizer Kingma & Ba (2014) with default settings. To achieve a good minimization of the loss, and in particular stability w.r.t varying random initializations (see also Subsection 3.3 below), we iteratively reduce the learning rate after a fixed number of epochs, track the network parameters and latent variables that achieve the minimal loss, and export those parameters and variables as the optimal solution (instead of the variables of the last iterate). Note that this does not cause much computational overhead due to the rater small dimensionality of our network (see Section 3 for details).

3 Numerical Experiments

In this section the results of several experiments concerning dynamic images with respiratory and cardiac motion are presented to illustrate the behavior of the new method. In particular, we present

a synthetic phantom example and two semi-synthetic examples where real dynamic cardiac MR images were enriched with synthetic respiratory motion. Note that a synthetization of some of the motion types is necessary in order to have a ground-truth with isolated motion available.

Results for two additional real dynamic cardiac MR images are provided in the appendix. In addition, the supplementary material contains videos for all results. For all experiments, we assume one-dimensional information, henceforth referred to as motion triggers, about the respiratory motion to be given. For the phantom data, we include also experiments with given motion-triggers for both motion types as reference scenario. For the MR images, we include additional experiments where only a perturbed version of the breathing-motion trigger is available.

To assess the quality of the motion isolation, we compute the relative error norms ($\mathbf{E}_h^1$ and $\mathbf{E}_h^2$) of the dynamic images $\hat{x}^1 = (\hat{x}_t^1)_{t=1}^T$, $\hat{x}^2 = (\hat{x}_t^2)_{t=1}^T$ containing just one reconstructed kind of motion, where

$$\mathbf{E}_h^1 = \|\hat{x}^1 - x_{\text{true}}^1\|_2 / \|x_{\text{true}}^1\|_2, \quad \mathbf{E}_h^2 = \|\hat{x}^2 - x_{\text{true}}^2\|_2 / \|x_{\text{true}}^2\|_2, \tag{6}$$

and x_{true}^i is the ground truth showing only the ith type of motion, i.e., cardiac motion for $i = 1$ and respiratory motion for $i = 2$. Here, the subscript h refers to the frame at which the other motion state is fixed, see (4). Since we are not aware of any comparable method for isolating different motion types in videos, we put the obtained error into context by providing also the idealized phantom experiments with information on all motion triggers as reference.

In principle, as described in Section 2, our method allows to freeze one kind of motion at any state, and generate images containing only the second kind of motion, as long as a sufficient mixing of motions was observed. In practice, the choice of the freezing frame h has an impact on the performance of the single motion reconstruction (though for the phantom at an overall rather low error regime). Figure 6c provides an example for this by plotting the errors $\mathbf{E}_h^1$ and $\mathbf{E}_h^2$ as a function of the frame h that is fixed for the other kind of motion. In our experiments, we always show results for fixing the motion state that provides the best performance with respect to the ground truth.

For all experiments shown in the paper, we repeated the experiment 20 times with 20 different seeds, and show the result whose performance w.r.t. the error measure $\mathbf{E}_h^1$ is closest to the median performance. For a discussion on the stability of our method w.r.t varying seeds see Subsection 3.3. Quantitative error measures for all experiments are provided in Table 1. The supplementary material further contains videos showing the best result that was achieved over the 20 seeds for each experiment. The source code to reproduce all experiments is available in the supplementary material. All our experiments were conduced on a workstation with an AMD Ryzen 7 3800X 8-Core Processor and 32 GB of memory, using a Nvidia RTX 3090 GPU with 24 GB of memory. The smallest largest experiment considered here (solving (5) with phantom and real cine MR data, respectively) took around 19 seconds and 3.8 minutes, respectively.

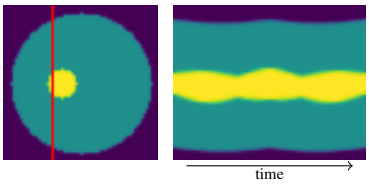
3.1 SYNTHETIC DATA

The first test problem, consisting of 80 frames with 64×64 pixels, corresponds to a synthetic example displaying two nested disks and is represented in Figure 1. A more compact representation of this dynamic image can be observed in Figure 21a, where a vertical slice of the image (marked in red on the left image), is displayed over time (and can be seen on the right image) clearly showing temporal changes. This representation will be used throughout the paper.

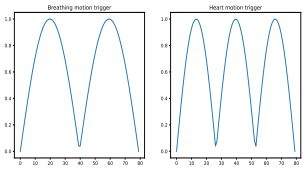
In this example, three cardiac motion cycles are simulated by dilation of the internal disk while two simulated breathing motion cycles are represented by shearing of the whole image. Note that the size of the frames over time is maintained constant. The ground truth one-dimensional motion information, i.e., the motion trigger, that was used to parametrize the different types of motions is displayed in Figure 2b. Note that this is the information that defines one (or both in a reference scenario) of the latent variables $(z^1(t))_{t=1}^T$ and $(z^2(t))_{t=1}^T$.

The generator modeling the solution for this example, as defined in Equation (2), corresponds to a standard deep convolutional neural network (CNN) with 5 layers, where transpose two-dimensional convolutions are used for all convolutions in (2) and no biases are used. The network parameters are given as follows. Number of channels: $[64, 128, 64, 32, 16, 1]$, (square) kernel size: $[4, 4, 4, 4, 4]$, stride: $[1, 2, 2, 2, 2]$, padding: $[0, 1, 1, 1, 1]$, activation functions: $[\text{Tanh}, \text{LeakyReLU}, \text{Tanh}, \text{LeakyReLU}, \text{Tanh}]$. In total, the network has 3.03360×10^5 parameters.

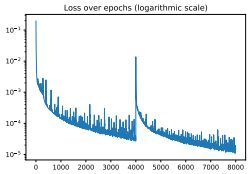
Figure 1: Selected phantom frames displaying different phases of respiratory and cardiac movement.



(a) Compact video representation.



(b) Ground truth motion triggers.



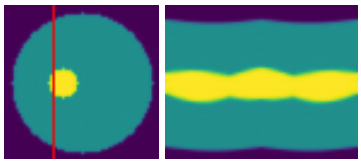
(c) Loss function history.

Figure 2: Alternative representation of the time evolution for the synthetic data and loss function plot.

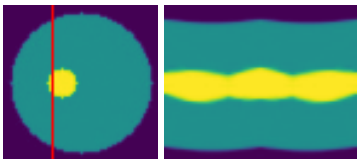
The latent space $Z \in \mathbb{R}^{64}$ is split into 62 static components and 2 dynamic components. Blocks of $[4000, 4000]$ epochs with learning rates $[0.01, 0.005]$ are used for the Adam optimizer. An example plot of the loss value history throughout the iterations can be found in Figure 2c. The latent variables are always initialized randomly from a uniform distribution on the interval $[0, 1)$, and the network weights are initialized orthogonally following Saxe et al. (2013).

The dynamic image reconstruction is performed in two different scenarios. First, as reference scenario, we consider both motion triggers associated to the two types of movements (as displayed in Figure 2b) to be known. In this case, we use them as the temporal-dependent latent variables $(z_t^1)_{t=1}^T$ and $(z_t^2)_{t=1}^T$, and optimize only the network weights θ. Second, we assume that one of the dynamic components in the latent space, $(z_t^1)_{t=1}^T$ corresponding to the cardiac motion, is unknown. We then use the framework defined in Equation (5), and we optimize the network over the network weights θ and the components of the latent space $(z_t^1)_{t=1}^T$.

The reconstruction of the dynamic image in both cases is shown in Figure 3. Even though a representation of the given data with mixed motion is not our primary goal, it can still be observed that both representations are visually very similar to the ground truth, giving evidence that one-dimensional information on one of the movements is enough to disentangle the latent space without a significant degradation in representing the data.



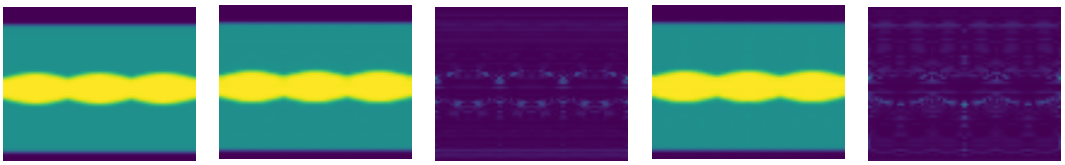
(a) Reconstruction for known $(z_t^1)_{t=1}^T$ $(z_t^2)_{t=1}^T$.



(b) Reconstruction for known $(z_t^2)_{t=1}^T$.

Figure 3: Synthetic video reconstruction with full motion.

Using the strategy described in Section 2, motion isolation is performed in the case where both motion triggers, $(z_t^1)_{t=1}^T$ and $(z_t^2)_{t=1}^T$, are known, and in the case where just $(z_t^2)_{t=1}^T$ is known. Figure 4 shows the isolated cardiac motion reconstructions, while Figure 5 shows the isolated respiratory motion reconstructions. Similarly to Figure 3, it can also be observed that single-motion reconstructions for either one or two known triggers are visually very close to the ground truth. This is especially meaningful for the reconstruction of cardiac motion in Figure 4d, where no prior information on the motion is used, supporting our hypothesis that latent space disentanglement is a meaningful tool for motion isolation.



(a) Ground truth. (b) Known z_t^1,z_t^2. (c) Error in (b). (d) Known z_t^1. (e) Error in (d).

Figure 4: Compact representation of the synthetic video reconstructions with only cardiac motion. Error images are upscaled by a factor of 10.

(a) Ground truth. (b) Known z_t^1,z_t^2. (c) Error in (b). (d) Known z_t^1. (e) Error in (d).

Figure 5: Compact representation of the synthetic video reconstructions with only respiratory motion. Error images are upscaled by a factor of 10.

3.2 CARDIAC MR IMAGES

After exploring the potential of the method on synthetic data, we test it on cine MR images, comprising images with a four-chamber view and datasets with a short-axis view. We provide results for two datasets in this section, and results for two additional datasets in the supplementary material. Table 1 provides error measures for all experiments. All real data used in this work comes from the datasets that were made available by the organizers of the ISMRM reconstruction challenge 2014[1]. In all experiments considered here, the original videos were obtained via a sum-of-squares reconstruction from fully sampled MR data, and contain a 2D slice of the entire thorax showing the beating heart in one region. For our experiments, we simulated three heartbeats by concatenating the single-heartbeat-videos three times, and simulated two respiratory cycles with vertical (resp. horizontal) shearing motion for the four-chamber (resp. short-axis) view. After obtaining videos showing a slice of the entire thorax with three heartbeats and two breaths, we cropped the videos to a region of interest around the heart, see the top rows of Figures 12 and 14 for the original data that is used as the input to our method. The final data consists of 99 frames with spatial resolution 100×100 for the four-chamber view, and of 81 frames with spatial resolution 70×70 for the short-axis view.

In all experiments with real data, the generator is a standard deep convolutional neural network (CNN) with 7 layers, ReLU activation functions in the first and last layer, LeakyReLUs in the middle layers and no biases. The latent space $\mathbb{R}^{100}$ is split into 98 static components and 2 dynamic components. Network parameters shared by all experiments are given as: Number of channels: $[100, 640, 320, 160, 80, 40, 20, 1]$, stride: $[2, 2, 2, 2, 2, 2, 1]$. To account for the different image dimensions, the shape of the remaining parameters differs slightly: (Square) kernel size: $[4, 4, 4, 4, 4, 4, 3]$ (four-chamber), $[4, 4, 4, 4, 4, 5, 4]$ (short-axis), padding: $[0, 2, 0, 2, 2, 1, 1]$ (four-chamber), $[0, 2, 2, 2, 2, 1, 1]$ (short-axis). In total, the networks have 5.388980×10^6 (four-chamber) and 5.396320×10^6 (short-axis) parameters.

The network is optimized according to Equation (5), where $\hat{z}^0$ and $(z_t^1)_{t=1}^T$ are randomly initialized from a uniform distribution on the interval $[0, 1)$, and the network weights are initialized according to the Pytorch self-initialization. For both experiments, blocks of $[4000, 4000, 4000, 4000, 4000]$ epochs with learning rates $[0.01, 0.008, 0.005, 0.003, 0.001]$ are used for the Adam optimizer.

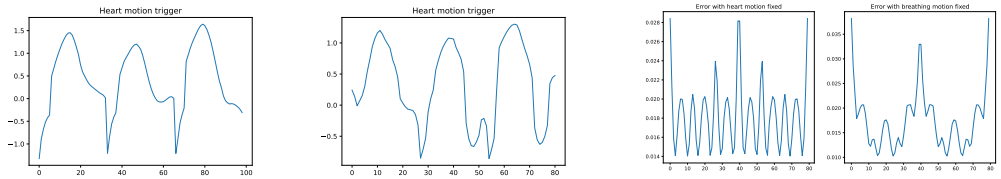
Motion isolation is performed on the MR images as explained in Section 2. The learned generator and the reconstructed dynamic latent space variables are used to freeze one type of motion while maintaining the dynamics of the other type of motion. The latent space variables associated to the cardiac motion are shown in Figures 6a and 6b for the four-chamber and short-axis view, respec-

[1] challenge.ismrm.org. Permission to use the data in research was obtained from the organizers per mail.

tively. Note that these plots have a physical interpretation associated to the heart's activity, and were completely unknown before performing the optimization.

The results of the motion isolation experiments are shown in Figures 7 and 8 for the four-chamber view experiment and in Figures 9 and 10 for the short-axis view experiment. In Figures 7 and 9, the top row shows selected frames of the given image sequence, containing both cardiac and respiratory motion. The bottom row shows a generated image sequence with only cardiac motion. Figures 8 and 10 show a reference frame with a marked slice (first column), the dynamics of the slice over time for the ground truth (second and fifth column) and for the generated image sequence (third and sixth column), and a difference image (third and seventh column). Columns two to four show the isolated cardiac motion, columns five to seven show the isolated breathing motion.

It can be observed that in both cases and for both types of motion, the isolation of motion works well, and the different structures of the motion are clearly visible in the slice-based visualization of the generated images. This is also confirmed by the quantitative values provided in Table 1. We should note that some artifacts are visible in the breathing motion isolation. In our experience, those are mostly related to having obtained a sub-optimal solution of the minimization problem (recall that we provide results corresponding to the median of the performance of our method). In cases where a favorable random initialization leads to an improved convergence of the methods these artifacts are reduced, see for instance the videos showing the best result that was achieved over the 20 seeds as provided in the supplementary material.



(a) Four-chamber view. (b) Short-axis view. (c) E_h^1 and E_h^2 as in (6) as a function of h.

Figure 6: (a),(b) Motion triggers for the two real MRI reconstructed videos with only cardiac motion. (c) Error w.r.t. fixing different frames for the second phantom experiment.

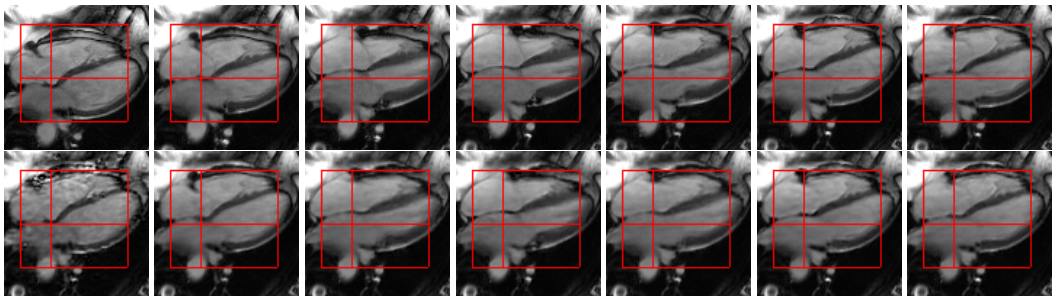


Figure 7: Four-chamber view. First row: Representative frames of the original video. Second row: Reconstructed video with only cardiac motion.

3.3 Stability and Extensions

To study the stability of the method with respect to the initialization of the parameters, the solution of the optimization problem and the subsequent motion isolation are repeated for 20 different seeds for each of the experiments in this paper, see Table 1 for an evaluation. It can be observed that the method is rather stable on average, with few (in practice 1-2 per 20 seeds) negative outliers. In our experiments, negative outliers were always connected to the loss of the final result still being comparatively high. Consequently, in application, these outliers can be detected by a high value of the loss (without knowing the ground truth), and the method can be re-run with a different initialization in such cases.

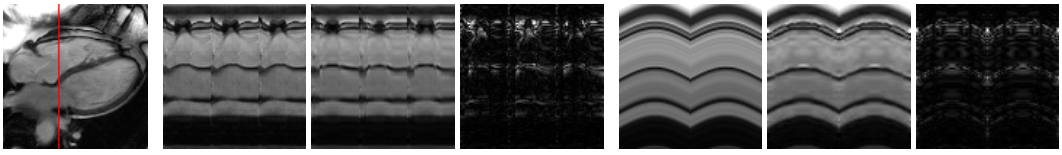


Figure 8: Four-chamber view. First image: Reference frame with marked slice, second (resp. fifth) image: slice over time for ground truth with only cardiac (resp. respiratory) motion, third (resp. sixth) image: slice over time for reconstructed images with only cardiac (resp. respiratory) motion, fourth and seventh image: difference between ground truth and reconstructed (upscaled by a factor of two).

Figure 9: Short-axis view. First row: Representative frames of the original video. Second row: Reconstructed video with only cardiac motion.

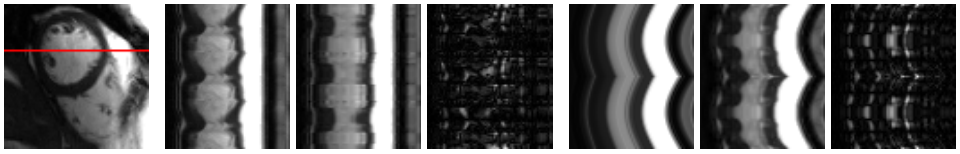


Figure 10: Short-axis view. First image: Reference frame with marked slice, second (resp. fifth) image: slice over time for ground truth with only cardiac (resp. respiratory) motion, third (resp. sixth) image: slice over time for reconstructed images with only cardiac (resp. respiratory) motion, fourth and seventh image: difference between ground truth and reconstructed (upscaled by a factor of two).

To further evaluate stability of the method with respect to errors in the provided motion triggers, we repeated the two real data experiments shown in the paper with perturbed motion triggers (the time-position where the motion trigger is sampled being perturbed by additive Gaussian noise with a standard deviation of 50% of the length of one timestep). While details and qualitative results are provided in the appendix, a quantitative evaluation of this experiment is provided in the last four lines of Table 1, showing that allowing for perturbation in the motion trigger does not degrade reconstruction performance.

In addition, we also refer to the appendix for an extension of the phantom experiment with tree types of motion (cardiac, respiratory and motion due to body movement), with the results provided there confirming that such an extension is possible in principle.

Table 1: Error in isolating of different kinds of motion for the experiments considered in this paper, repeating them for 20 different seeds. MAD denotes the median absolute deviation.

	Median	MAD	Mean	Std. dev.
Phantom example, z^1, z^2 known - $\mathbf{E}_h^1$ (cardiac)	4.86e-03	2.40e-04	4.89e-03	4.58e-04
Phantom example, z^1, z^2 known - $\mathbf{E}_h^2$ (resp.)	6.65e-03	3.83e-04	6.51e-03	8.81e-04
Phantom example, z^2 known - $\mathbf{E}_h^1$ (cardiac)	9.93e-03	1.05e-03	1.03e-02	1.97e-03
Phantom example, z^2 known - $\mathbf{E}_h^2$ (respiratory)	1.20e-02	1.09e-03	1.22e-02	1.39e-03
Four-chamber view - $\mathbf{E}_h^1$ (cardiac)	1.01e-01	2.25e-02	1.21e-01	4.84e-02
Four-chamber view - $\mathbf{E}_h^2$ (respiratory)	8.29e-02	1.94e-02	9.34e-02	3.88e-02
Short-axis view - $\mathbf{E}_h^1$ (cardiac)	8.07e-02	1.21e-02	9.21e-02	2.83e-02
Short-axis view - $\mathbf{E}_h^2$ (respiratory)	7.21e-02	1.41e-02	8.29e-02	3.20e-02
Four-chamber view, example 2 - $\mathbf{E}_h^1$ (cardiac)	9.02e-02	7.55e-03	1.07e-01	4.84e-02
Four-chamber view, example 2 - $\mathbf{E}_h^2$ (resp.)	7.52e-02	1.14e-02	9.24e-02	5.90e-02
Short-axis view, example 2 - $\mathbf{E}_h^1$ (cardiac)	1.36e-01	3.41e-02	1.42e-01	3.84e-02
Short-axis view, example 2 - $\mathbf{E}_h^2$ (respiratory)	1.05e-01	2.50e-02	1.14e-01	3.27e-02
Four-chamber view (perturbed) - $\mathbf{E}_h^1$ (cardiac)	9.87e-02	1.55e-02	1.20e-01	4.08e-02
Four-chamber view (perturbed) - $\mathbf{E}_h^2$ (resp.)	8.06e-02	1.04e-02	9.85e-02	3.25e-02
Short-axis view (perturbed) - $\mathbf{E}_h^1$ (cardiac)	7.78e-02	5.32e-03	7.89e-02	1.30e-02
Short-axis view (perturbed) - $\mathbf{E}_h^2$ (respiratory)	6.58e-02	4.67e-03	6.88e-02	1.27e-02

CONCLUSIONS AND OUTLOOK

This paper introduces a new method for motion isolation based on the joint optimization of an untrained generator network over both the network parameters and the latent codes. Assuming one-dimensional information on all but one of the motions is known, motion isolation is achieved through latent space disentanglement. Feasibility of this method was shown for isolating respiratory and cardiac motion in dynamic MR images, but the proposed method is general and can conceptually be used in many applications, e.g., in bio-medical imaging, biology, bio-mechanics or physics.

A limitation of the method, resulting from non-convexity, is its dependence on initializations, which is counteracted here via loss-based restarting strategies. Further, no explicit motion information is made available by our methods. While this might be considered as limitation, it also comes with the advantage that no explicit modeling of the underlying motion types is necessary.

Our work shows a great potential of latent space disentanglement on untrained generators for video data. It opens the door to more advanced disentanglement schemes (e.g., based on modifications of the loss function or additional constraints on the latent space variables). Moreover, we expect the proposed method to be very well suited as image prior for dynamic inverse problems (e.g., in tomography, super-resolution) where the reconstructed solution displays different kinds of independent motion.

4 REPRODUCIBILITY STATEMENT

The source code and data to reproduce all experiments of the main part of the paper are available in the supplementary material. Further, after acceptance of the paper, all data and source code necessary to reproduce the experiments of the paper will be published as GIT repository.

REFERENCES

Daniel Otero Baguer, Johannes Leuschner, and Maximilian Schmidt. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Problems*, 36(9): 094004, 2020.

Aurélien Bustin, Niccolo Fuin, René M Botnar, and Claudia Prieto. From compressed-sensing to artificial intelligence-based cardiac MRI reconstruction. *Frontiers in cardiovascular medicine*, 7: 17, 2020.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, 29, 2016.

Sören Dittmer, Tobias Kluth, Peter Maass, and Daniel Otero Baguer. Regularization by architecture: A deep prior approach for inverse problems. *Journal of Mathematical Imaging and Vision*, 62(3): 456–470, 2020.

Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1–21, 2015.

Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, 65(20):20TR01, 2020.

Kuang Gong, Ciprian Catana, Jinyi Qi, and Quanzheng Li. PET image reconstruction using deep image prior. *IEEE transactions on medical imaging*, 38(7):1655–1665, 2018.

Andreas Habring and Martin Holler. A generative variational model for inverse problems in imaging. *SIAM Journal on Mathematics of Data Science*, 4(1):306–335, 2022.

Reinhard Heckel and Mahdi Soltanolkotabi. Denoising and regularization via exploiting the structural bias of convolutional generators. *arXiv preprint arXiv:1910.14634*, 2019.

Rakib Hyder and M Salman Asif. Generative models for low-dimensional video representation and reconstruction. *IEEE Transactions on Signal Processing*, 68:1688–1701, 2020.

Gauri Jagatap and Chinmay Hegde. Algorithmic guarantees for inverse imaging with untrained network priors. *Advances in neural information processing systems*, 32, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ankur N Kumar, Kurt W Short, and David W Piston. A motion correction framework for time series sequences in microscopy images. *Microscopy and Microanalysis*, 19(2):433–450, 2013.

Andre Z Kyme and Roger R Fulton. Motion estimation and correction in SPECT, PET and CT. *Physics in Medicine and Biology*, 66(18), 2021.

Kanglin Liu, Gaofeng Cao, Fei Zhou, Bozhi Liu, Jiang Duan, and Guoping Qiu. Towards disentangling latent space for unsupervised semantic face editing. *IEEE Transactions on Image Processing*, 31:1475–1489, 2022.

Francisco PM Oliveira and Joao Manuel RS Tavares. Medical image registration: A review. *Computer methods in biomechanics and biomedical engineering*, 17(2):73–93, 2014.

Ricardo Otazo, Emmanuel Candes, and Daniel K Sodickson. Low-rank plus sparse matrix decomposition for accelerated dynamic mri with separation of background and dynamic components. *Magnetic resonance in medicine*, 73(3):1125–1136, 2015.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural information processing systems*, 2017.

Arman Rahmim, Jing Tang, and Habib Zaidi. Four-dimensional (4D) image reconstruction strategies in dynamic PET: Beyond conventional independent frame reconstruction. *Medical physics*, 36(8): 3654–3670, 2009.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Zhigang Tu, Wei Xie, Dejun Zhang, Ronald Poppe, Remco C Veltkamp, Baoxin Li, and Junsong Yuan. A survey of variational and CNN-based optical flow techniques. *Signal Processing: Image Communication*, 72:9–24, 2019.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1526–1535, 2018.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *International Journal of Computer Vision*, 128:1867–1888, 2020.

Jaejun Yoo, Kyong Hwan Jin, Harshit Gupta, Jerome Yerly, Matthias Stuber, and Michael Unser. Time-dependent deep image prior for dynamic mri. *IEEE Transactions on Medical Imaging*, 40 (12):3337–3348, 2021.

A Appendix

A.1 Results on two additional datasets

In this section of the appendix, we present results as shown in the paper for two additional MR images, one with a four-chamber view (Figures 12 and 13) and one with a short-axis view (Figures 14 and 15). The estimated latent space variables associated to the cardiac motion are shown in Figure 11.

It can be observed that, again, in both cases the isolation of cardiac and respiratory motion works well with the proposed method. In particular for the short-axis view, as can be seen in Figure 15, the isolation of respiratory motion is even better than with the short-axis example shown in the paper. In this case, however, the reconstructed motion trigger for the cardiac motion is inaccurately reconstructed for the second heartbeat. Again, note that we present result corresponding to the median performance of the methods over repeating each experiment 20 times with 20 different seeds for the random initialization. In case of a favorable random initialization, convergence of the method and, consequently, also the corresponding results are improved, see for instance the videos provided in the supplementary material.

(a) Four-chamber view. (b) Short-axis view.

Figure 11: Motion triggers for the two real MRI reconstructed videos with only cardiac motion.

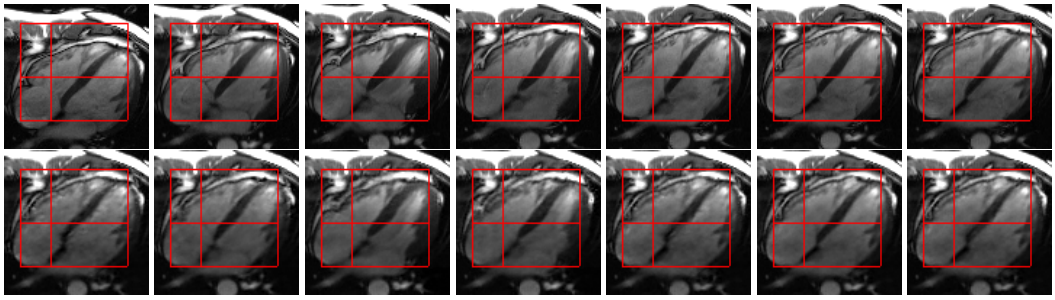


Figure 12: Four-chamber view. First row: Representative frames of the original video. Second row: Reconstructed video with only cardiac motion.

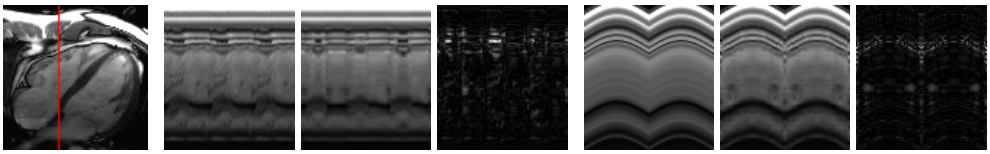


Figure 13: Four-chamber view. First image: Reference frame with marked slice, second (resp. fifth) image: slice over time for ground truth with only cardiac (resp. respiratory) motion, third (resp. sixth) image: slice over time for reconstructed images with only cardiac (resp. respiratory) motion, fourth and seventh image: difference between ground truth and reconstructed (upscaled by a factor of two).

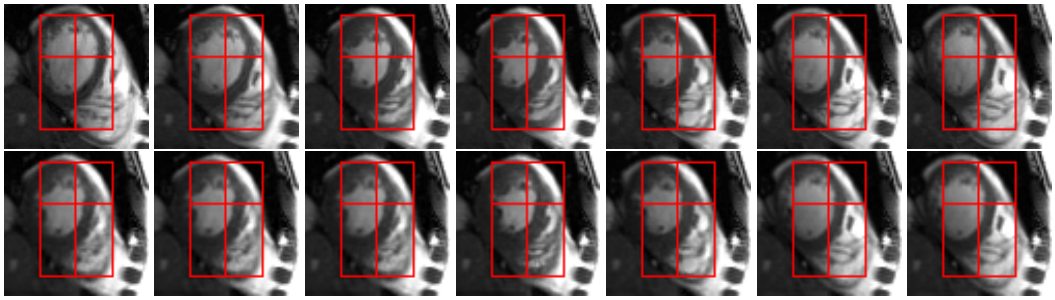


Figure 14: Short-axis view. First row: Representative frames of the original video. Second row: Reconstructed video with only cardiac motion.

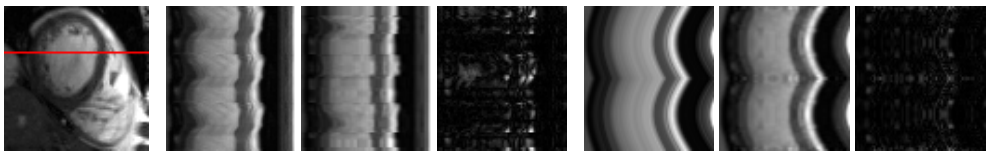
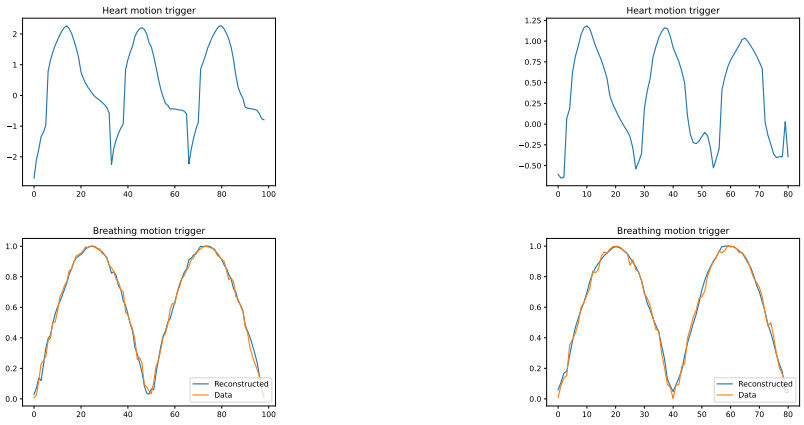


Figure 15: Short-axis view. First image: Reference frame with marked slice, second (resp. fifth) image: slice over time for ground truth with only cardiac (resp. respiratory) motion, third (resp. sixth) image: slice over time for reconstructed images with only cardiac (resp. respiratory) motion, fourth and seventh image: difference between ground truth and reconstructed (upscaled by a factor of two).

A.2 RESULTS OBTAINED WITH PERTURBED MOTION TRIGGERS

In this section, we repeat the two experiments of Section 3.2 of the paper, but instead of taking the true breathing motion trigger that created the data as given, we consider a perturbed version with the time-position where the motion trigger is sampled being perturbed by additive Gaussian noise with a standard deviation of 50% of the length of one timestep. Then we use this motion trigger (instead of the ground truth one) as input for our method, allowing deviations from the perturbed trigger that are penalized with an L^2 discrepancy. Quantitative results for this experiment can be found in the last four lines of Table 1 of the paper. Qualitative results can be found in Figures 16 to 20. In particular, Figure 16 shows the initially provided trigger for breathing motion in orange, and the reconstructed triggers in blue. As can be seen, the method is able so successfully remove perturbations in the trigger initially provided.

Figures 16 to 20 visualize the resulting image sequences. Comparing to the corresponding results of the paper, it can be observed that the reconstruction quality has not become worse due to this perturbation (only the initial frame 0 in Figure 17 (left frame in bottom line) suffers from artifacts).



(a) Four-chamber view.

(b) Short-axis view.

Figure 16: Motion triggers for the two real MRI reconstructed videos with only cardiac motion.

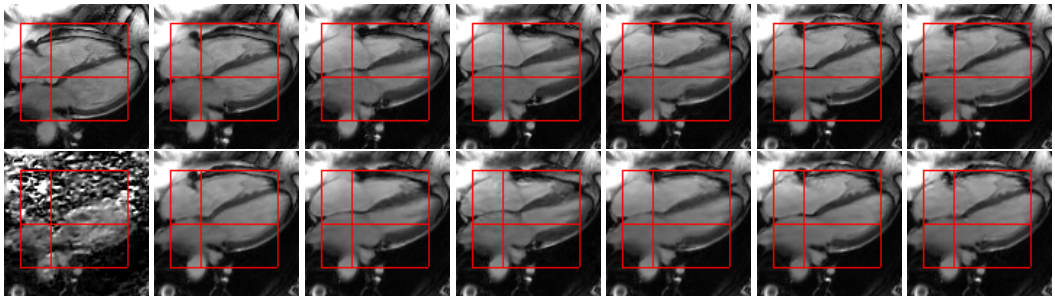


Figure 17: Four-chamber view, reconstructed using a perturbed respiratory motion trigger. First row: Representative frames of the original video. Second row: Reconstructed video with only cardiac motion.

A.3 EXPERIMENTS WITH A THIRD TYPE OF MOTION

In this section of the appendix, we provide results of an phantom experiment with three different simulated types of motion: Respiratory motion, cardiac motion and motion due to body movement. To this aim, we use the same data, network parameters and algorithmic setup as in the phantom experiment of the paper, assuming the motion triggers for respiratory and cardiac motion to be known, and the motion due to body movement to be completely unknown.

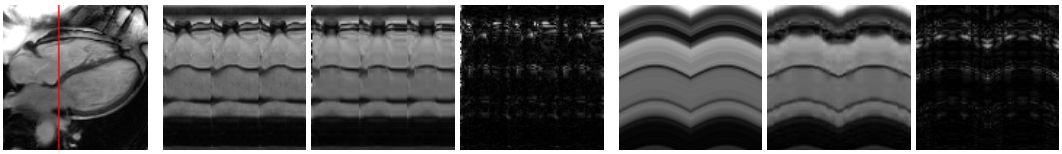


Figure 18: Four-chamber view, reconstructed using a perturbed respiratory motion trigger. First image: Reference frame with marked slice, second (resp. fifth) image: slice over time for ground truth with only cardiac (resp. respiratory) motion, third (resp. sixth) image: slice over time for reconstructed images with only cardiac (resp. respiratory) motion, fourth and seventh image: difference between ground truth and reconstructed (upscaled by a factor of two).

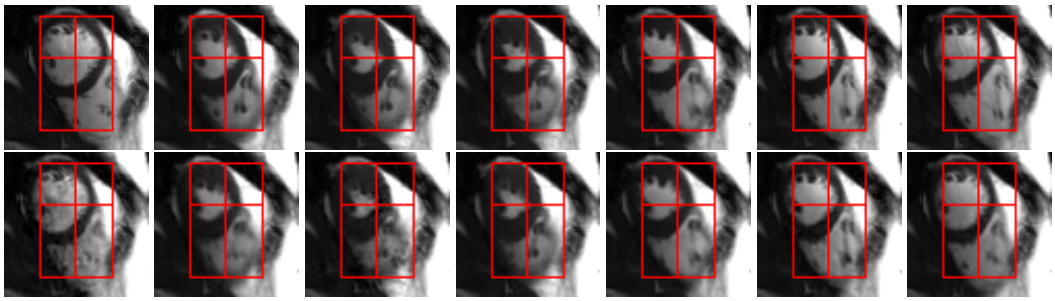


Figure 19: Short-axis view, reconstructed using a perturbed respiratory motion trigger. First row: Representative frames of the original video. Second row: Reconstructed video with only cardiac motion.

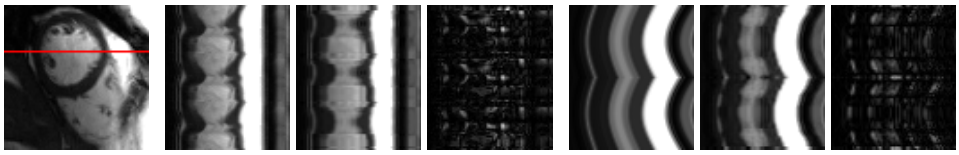
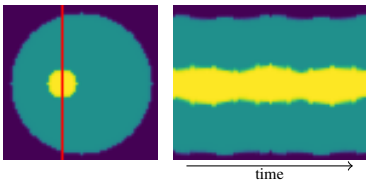
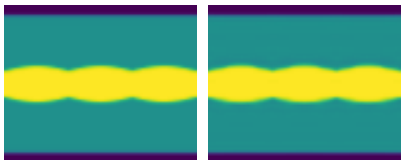


Figure 20: Short-axis view, reconstructed using a perturbed respiratory motion trigger. First image: Reference frame with marked slice, second (resp. fifth) image: slice over time for ground truth with only cardiac (resp. respiratory) motion, third (resp. sixth) image: slice over time for reconstructed images with only cardiac (resp. respiratory) motion, fourth and seventh image: difference between ground truth and reconstructed (upscaled by a factor of two).

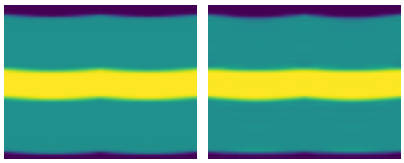
Results are provided in Figure 21. As can be seen there, the proposed method is capable of isolating all three types of motion successfully, providing results for all the types that are visually rather close to the ground truth.



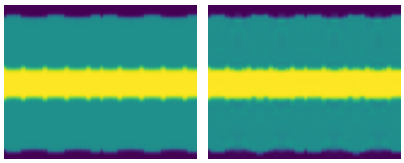
(a) Selected slice and slice representation of data.



(b) Only cardiac motion.



(c) Only respiratory motion.



(d) Only body motion.

Figure 21: Results for isolating three types of motion with phantom data. In subfigures (b),(c),(d), the left images shows the ground truth and the right image shows the reconstructed motion.