Towards Encoding 3D Abdominal MRI Acquisitions as Neural Fields

Basty, N., Rainer, G., Thomas, E.L., Bell, J.D. and Whitcher, B.

https://doi.org/10.1109/ISBI60581.2025.10980772

Towards Encoding 3D Abdominal MRI Acquisitions as Neural Fields

Basty, N., Rainer, G., Thomas, E.L., Bell, J.D. and Whitcher, B.

https://doi.org/10.1109/ISBI60581.2025.10980772

Towards Encoding 3D Abdominal MRI Acquisitions as Neural Fields

Nicolas Basty[§†], Gilles Rainer[⋆†], E. Louise Thomas[§†], Jimmy D. Bell[§†],

and Brandon Whitcher[§†]

†Orcino Health Ltd, London, United Kingdom

§Research Centre for Optimal Health, University of Westminster, London, United Kingdom

⋆ Imperial College London, London, United Kingdom

ABSTRACT

Medical imaging data is typically 3D, causing scan sizes and databases to grow cubically with resolution, unlike the quadratic growth in standard computer vision tasks. Com- pressing scan dimensionality is essential for deep learning, as raw data often exceeds GPU memory limits. Autoencoders are commonly used for data-specific non-linear compression, balancing compactness and fidelity. However, they are limited to the resolution of the training data. Inspired by Neural Fields, we propose an autoencoder with a fully-connected network as its decoder, and train it on the UK Biobank abdominal MRI dataset. Beyond more fidelity in the reconstruction, our encoding is a continuous function of 3D coordinates rather than 3D rasters like the original data, which enables our architecture to be utilized in a variety of applications such as super-resolution, in-painting and extrapolation. We show that this change of paradigm in representation leads to higher and better compression, with better properties, and enables the use of such imaging databases for deep learning in their compressed state.

Index Terms— Continuous function, Implicit representation, Latent space

1. INTRODUCTION

Current deep learning research in style transfer, multilabel segmentation, and super-resolution for high-resolution 3D medical imaging faces key challenges as these tasks require extensive GPU memory and large labeled datasets, both limited by data scarcity and privacy regulations. Developing high-quality synthetic data that mirrors real medical imaging is crucial to address these limitations, reduce privacy concerns, and enable more robust studies.

Although techniques like Generative Adversarial Net- works (GANs), often focused on style transfer [1], have been popular for data synthesis, they face challenges in

handling the vast amounts of data associated with volumetric medical imaging data. More recently, techniques like diffusion have emerged but are still limited in effectively representing and managing large heterogeneous datasets. Hence, there is an increasing demand for alternative approaches to efficiently represent such data at manageable sizes, facilitating analysis and training across large databases of medical scans.

Recent work on diffusion models in magnetic resonance imaging (MRI) of the brain utilized compressed latent codes [2] instead of the raw 3D data. Images of the brain follow a very regular structure and tools for segmentation and registration are widely available, which is not necessarily the case when it comes to other anatomical regions of interest. Whole-body imaging comes with a much larger degree of data variability, as well as potentially larger image volumes, despite lower resolution. As a result, fewer works dealing with whole-body imaging exist, one such example is Mensing et al. [3] in which the group uses conditional GANs to generate whole-body data, but are restrained by GPU memory issues, and improvements would rely on larger models and/or higher resolution data which would both come with higher GPU memory requirements.

An emerging encoding technique for 2D or 3D data, from the field of computer vision, is implicit representations [4, 5], which summarise imaging data as continuous functions, breaking away from standard imaging grids and enabling powerful compression. Standard Convolutional Neural Networks (CNNs) reason via the relationship between neighboring pixels (inherently tied to the training resolution) and translation-invariant filters, which is an appropriate representation for tasks like texturing and object detection or classification. However, CNNs still operate in the discrete pixel domain, with local filters, inherently limiting the reconstruction expressiveness and ability, and inheriting un- desirable grid artifacts. Coordinate-based fully-connected networks on the other hand have shown great capacity for encoding signals and by design inherit all the advantages of switching to a continuous domain. In early works, small neural networks such as a Multi-Layer Perceptron (MLP) were overfit to a single instance of the database, but subsequent works have investigated how to generalize across a database, either through the use of hypernetworks [6] or the use of an additional modulation MLP [7]. Implicit representations are also beginning to be applied to medical imaging data as shown in a recent review article [8], with a majority using them for image reconstruction.

In this research paper, we propose a novel architecture combining a 3D CNN-based encoder and a modulated SIREN-based decoder [7], to learn continuous representations of whole-body MRI in order to efficiently compress large volumes in preparation for tasks such as data synthesis. Through our investigation, we seek to contribute to the ongoing efforts to leverage deep learning for whole-body MRI.

2. METHODOLOGY

2.1. Image pre-processing

In our experiments, we used data from the UK Biobank abdominal MRI protocol [9], specifically the in-phase channel of the 3D Dixon MRI acquisition. The positioning of each participant during scanning is subject to random factors, including the accuracy of neck placement by the radiographer, centering, as well as the degree of alignment along the axis of the scanner. These random factors result in varying degrees of inclusion of arms and legs. The protocol covers 1.1m from the neck downwards. This results in clipping of the knees when the neck positioning is too high or when the subject is tall.

As for all encoding tasks, removing unwanted variance from the database is crucial. Standardizing the alignment of scans within the grid is of particular importance for coordinate-based models that operate in a canonical unit space. We want our model to spend its representational power on learning and explaining aspects of the data that are relevant to subsequent research tasks; shifts in position within the volume, for instance, are of no relevance to the database and should be minimized in the curating phase.

We hence shifted the data such that the centre of the hips is in the centre of the volume and the top of the image at the height of the middle point between the shoulders, guided using the bone-joint landmarks computed using [10]. The final image size was set to $128 \times 128 \times 192$, resulting in no sudden boundary at the top or bottom of the image.

2.2. Baseline CNN Auto-Encoder

Convolutional auto-encoders have established themselves as the standard non-linear dimensionality reduction technique for data that is arranged on regular grids, typically images. A convolutional encoder compresses the image into a latent code, which the convolutional decoder decompresses back into a grid of pixels. The training is self-supervised, since the output of the compression-decompression is optimized to match the input as closely as possible. At inference time, in- stances can be compressed by passing them through the encoder and stored in their compressed code version. Similarly, the latent space can be explored by interpolating between la- tent codes of known inputs, for example.

For 3D data, such as MRI scans, the convolutions use 3D filters. Because of the added dimension, the size of the data, the filters, gradients and intermediate representations, now grows cubically rather than quadratically, which limits the practicability of convolutional architectures. We propose to use a completely different paradigm for the decoder instead of using a symmetric architecture, where the decoder is com- posed of deconvolutions of the same size and stride as the encoder.

2.3. Proposed Model

We propose to replace the decoder with a fully-connected architecture parametrized on grid coordinates. A schematic of the architecture is shown in Figure 1. The encoder remains the same standard CNN but we replace the decoder with a modulated SIREN architecture [7].

Rather than outputting a 3D tensor of identical shape to the input, as is common with convolutional architectures, our decoder is parametrized on input coordinates (x,y,z) and outputs a single value for the given point. Points can be batched to predict the entire volume, but fundamentally the decoder learns the mapping from a position to its corresponding value, as a continuous function. The coordinates are fed into a SIREN [5] architecture which consists of fully- connected layers with sine activations. The latent code is fed to the modulator network, which is fully-connected with skip connections and ReLU activations. The modulator can adapt the SIREN's predicted density field to the relevant shape and appearance based on the scan's latent code by outputting element-wise multipliers for the internal SIREN activations at every layer.

During training, we minimise a combination of Kullback- Leibler (KL) divergence, on the latent codes, and mean squared error on the final output, compared to the high resolution ground truth. The computed KL divergence is used as a loss function on the latent space, which is commonly encountered in variational autoencoders (VAEs) or similar models [2].

For our proposed model, we compute the MSE loss on a sparse set of points (75,000, equivalent to 2.38% of the data contained in the ground truth volume) within the 3D unit cube, as it would be too memory-expensive to feed the entire volume forward and backward. Random sampling of the 3D volume spreads points equally in space, resulting in many samples of the background, the empty space around the body. Sampling randomly in polar coordinates (angle and radius from the center) results in samples concentrating more towards the center, but only provides samples within the unit sphere. We use this polar sampling in the XY-plane, and random sampling of the Z coordinate, then remap the XY coordinates to stretch the samples to cover the entire square volume, as shown on the right in Fig. 2. This is done by simply scaling the radius of each generated point by the length of the segment that intersects the square. The final sampling scheme is more appropriate for our body scans after centering. A new random set of points is generated at every training iteration.

2.4. Training Details

The final model was trained on an NVIDIA A5000 GPU with 24GB memory using data of 1,000 UK Biobank participants, for 450 epochs using the Adam optimizer with a learning rate of 10^{-4}, with a batch size of 4 and 75,000 random point samples per epoch. We used the same 3D convolutional encoder for both the baseline CNN and

our architecture, compressing the input to a latent vector of 3072 elements. Our modulated SIREN decoder has 5 hidden layers of 1024 neurons.

3. RESULTS AND ANALYSIS

We compare several reconstruction metrics on an unseen test set of 300 scans to highlight the differences between the CNN baseline and our modulated SIREN decoder. Both architectures are trained on the same data, for the same number of epochs, and compress the scans to the same length latent vector (3072 elements). This corresponds to a 4694× compression ratio compared to the original data or a 1024× compression with regards to the pre-processed, aligned and cropped scans. Table 1 shows that our model outperforms a standard CNN autoencoder in terms of data fidelity for the same task, however these quantitative metrics do not necessarily reflect the substantial visual improvement, as a lot of image volume is just background. Fig. 2 provides a visual comparison, where our model recovers more crisp details from the encoded scans versus the output from a standard CNN.

3.1. Latent Space Exploration – Participant Interpolation

Fig. 3 shows an example of linearly traversing the latent space between the projected codes of two participants. The intermediate reconstructions display details and a plausible evolution between the two participants, demonstrating that the latent space is well-behaved and the decoder has learned a realis- tic model of the human body.

3.2. Scan Extrapolation and Inpainting

Continuous sampling enables querying data anywhere in 3D space. This property could be used to extend the field-of-view for tall participants, or positioning errors, and therefore help homogenize scans across large databases. Fig. 4 shows 4 such examples, the second example shows clipped lungs being re- covered. Fig. 5 shows a motivating example when acquisition of parts of the scan have been omitted or corrupted. We initialize the latent code to zeros and optimize it, supervising the reconstruction with only half the input signal, and it manages to reconstruct a body that is very similar to the ground truth.

4. CONCLUSION

We describe a novel architecture combining CNNs and modulated Siren for implicit representations of abdominal MRI scans. In addition to showcasing improved quantitative and qualitative results for extreme compression over standard CNN autoencoders, which have shown great promise in highly heterogeneous data such as

the brain [2], implicit representations directly unlock exciting new applications on top of facilitating research through data compression.

5. ACKNOWLEDGMENTS

This research has been conducted using the UK Biobank Re- source under Application Number 105847.

6. COMPLIANCE WITH ETHICAL STANDARDS

The UK Biobank has approval from the North West Multi- Centre Research Ethics Committee (REC reference: 11/NW/0382), with informed consent obtained from all participants.

7. REFERENCES

[1] S. Dayarathna, K. T. Islam, S. Uribe, G. Yang, M. Hayat, and Z. Chen, “Deep learning based synthesis of MRI, CT and PET: Review and analysis,” Medical Image Analysis, p. 103046, 2023.

[2] W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, “Brain imaging generation with latent diffusion models,” in MICCAI Workshop on Deep Generative Models. Springer, 2022, pp. 117–126.

[3] D. Mensing, J. Hirsch, M. Wenzel, and M. Gunther, “3D (c) GAN for whole body MR synthesis,” in MICCAI Workshop on Deep Generative Models. Springer, 2022, pp. 97–105.

[4] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing scenes as neural radiance fields for view synthesis,” in ECCV, 2020.

[5] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, “Implicit neural representations with peri- odic activation functions,” Advances in Neural Information Processing Systems, vol. 33, pp. 7462–7473, 2020.

[6] V. Sitzmann, E. R. Chan, R. Tucker, N. Snavely, and G. Wetzstein, “MetaSDF: Meta-learning signed distance functions,” in Proc. NeurIPS, 2020.

[7] I. Mehta, M. Gharbi, C. Barnes, E. Shechtman, R. Ramamoorthi, and M. Chandraker, “Modulated peri- odic activations for generalizable local functional representations,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14214–14223.

[8] A. Molaei, A. Aminimehr, A. Tavakoli, A. Kazerouni, B. Azad, R. Azad, and D. Merhof, “Implicit neural representation in medical imaging: A comparative survey,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2381–2391.

[9] T. J. Littlejohns, J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, J. D. Bell, C. Boultwood, R. Collins, M. C. Conroy, N. Crabtree, N. Doherty, A. F. Frangi, N. C. Harvey, P. Leeson, K. L. Miller, S. Neubauer, S. E. Petersen, J. Sellors, S. Sheard, S. M. Smith, C. L. M. Sudlow, P. M. Matthews, and N. E. Allen, “The UK Biobank imaging enhancement of 100,000 participants: Rationale, data collection, management and future directions,” Nature Communications, vol. 11, no. 1, 2020.

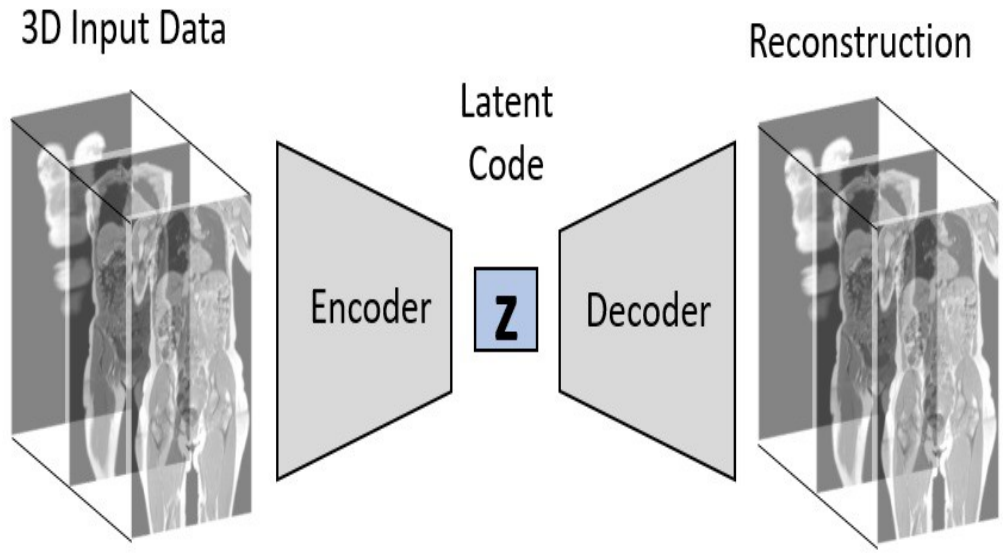
[10] N. Basty, Y. Liu, M. Cule, E. L. Thomas, J. D. Bell, and B. Whitcher, “Image processing and quality control for abdominal magnetic resonance imaging in the UK Biobank,” arXiv preprint arXiv:2007.01251, 2020.

Tables

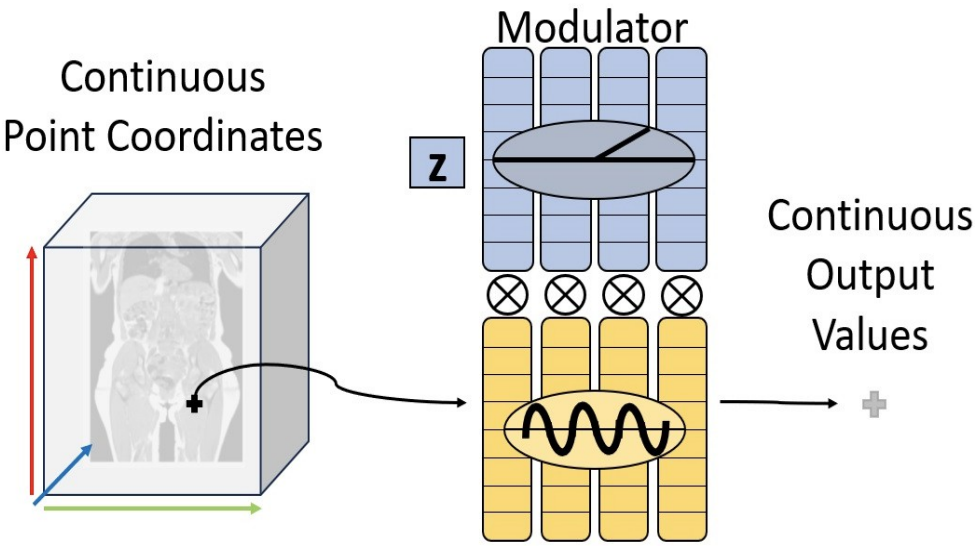
Table 1. Quantitative evaluation for 300 held out test datasets.

	MSE (↓)	SSIM (↑)	PSNR (↑)
Baseline CNN	0.242	0.723	19.45 (dB)
Ours	**0.218**	**0.731**	**20.03** (dB)

Figures



(a) Symmetric encoder-decoder



(b) Our custom decoder architecture

Fig. 1. Typically, the decoder architecture is the symmetric of the encoder and the output is a tensor of identical shape to the input (a). Our decoder instead learns the data as a continuous function of the 3D coordinates inside the MRI volume (b).

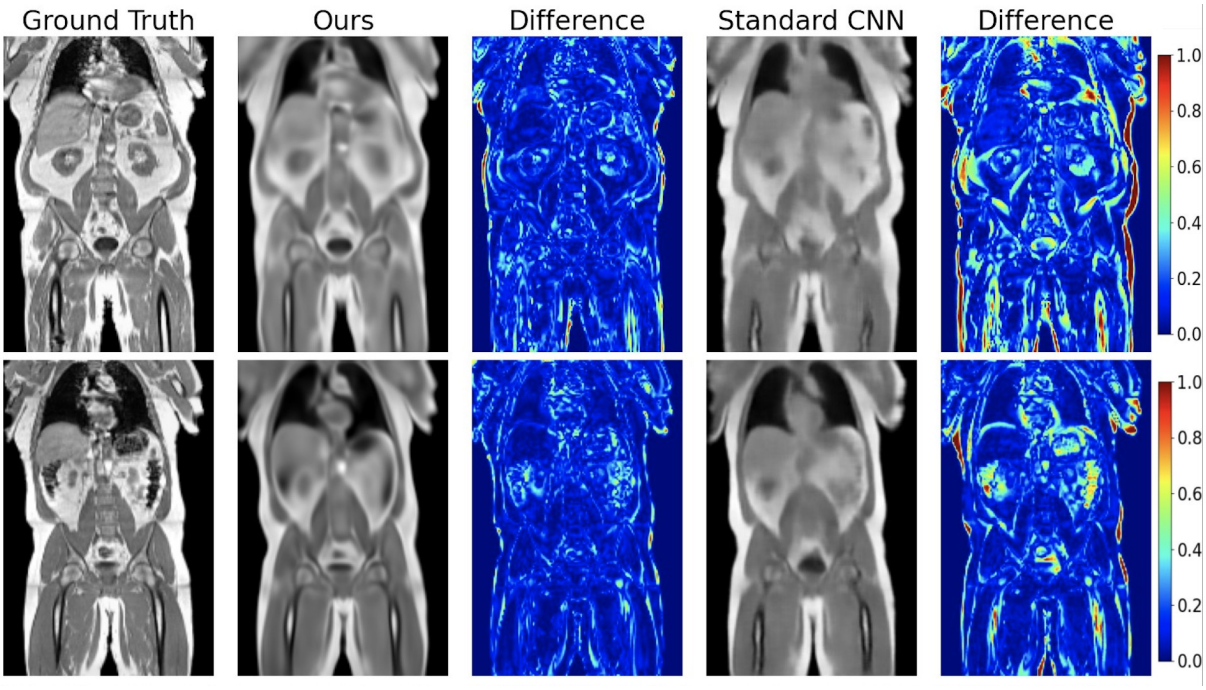


Fig. 2. Example reconstructions from the test set.

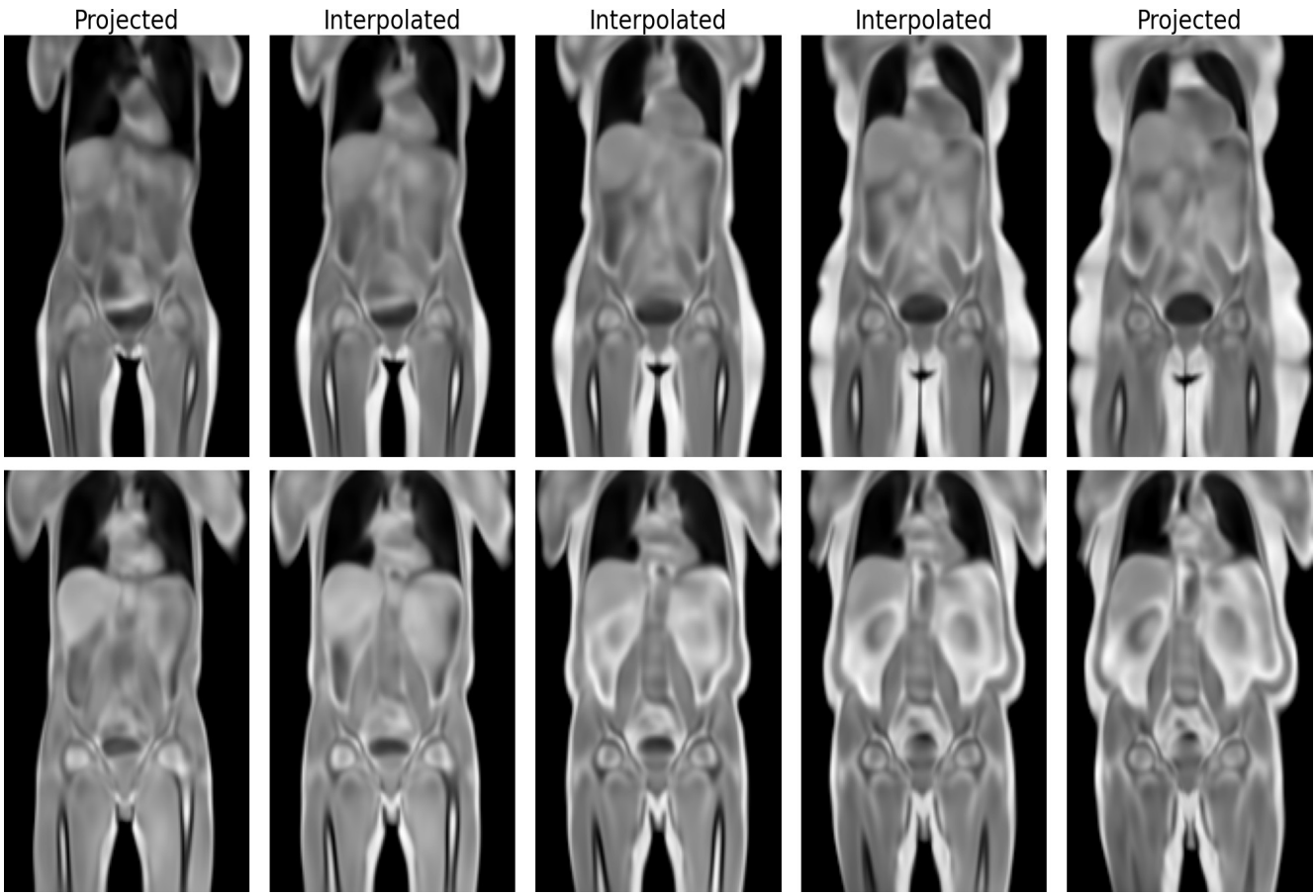


Fig. 3. Examples of linear interpolation in the latent space between 2 projections.

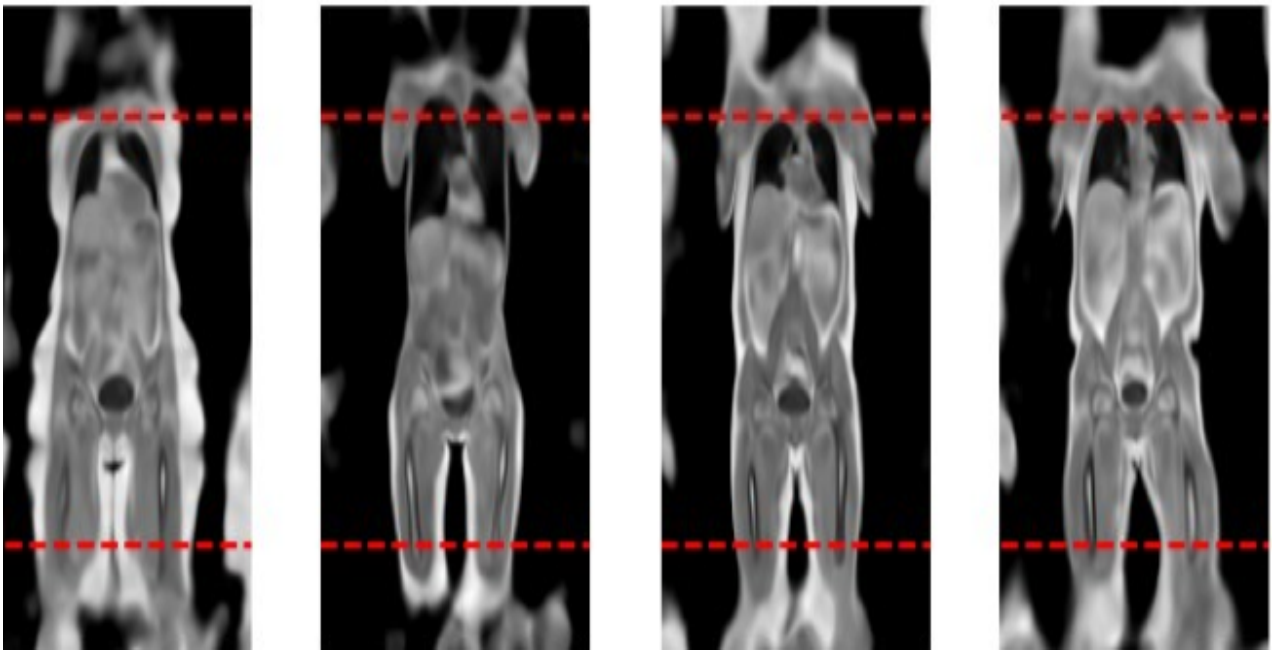


Fig. 4. Extrapolation outside the scan volume coordinates (highlighted by red dashed lines on top and bottom)

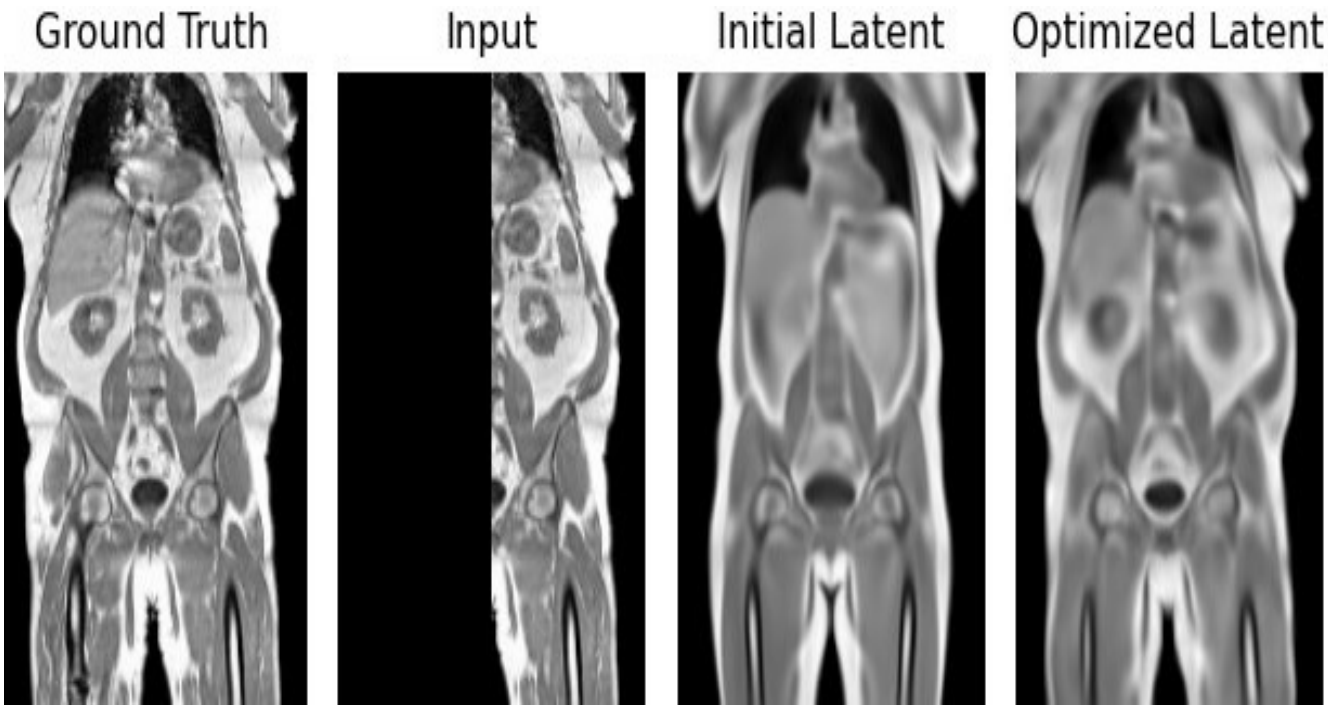


Fig. 5. Inpainting for missing data.