AN ASSESSMENT OF THE EFFECT OF MULTI-TASK LEARNING IN THE RECURRENT INFERENCE MACHINE FOR SPARSE MAGNETIC RESONANCE IMAGING RECONSTRUCTION.

T.M. Paquaij

July 2024

Master committee:
Committee Chair: Brune, C., prof.dr
Committee Member (UT): Wolterink, J.M., dr.
External Member: Simonis, F.F.J., dr.ir
Additional Member: Caan, M.W.A., dr.ir.

UNIVERSITY OF TWENTE.

Abstract

In recent decades, the increasing reliance on medical imaging for diagnostics has negatively impacted upon healthcare costs and efficiency. One of the reasons for this is that acquisition, reconstruction, image analysis, and diagnosis are traditionally separate steps, lengthening the diagnostic process. Accelerating these steps can save time and reduce costs. Accelerated magnetic resonance imaging (MRI) speeds up acquisition by minimising k-space samples, leading to aliased scans when linearly reconstructed. Deep learning methods, like recurrent inference machines (RIM), can reconstruct these scans faithfully using learned priors but currently exclude post-reconstruction tasks, missing the opportunity for improved performance through joint optimisation.

This research examines the impact of multi-task approaches on the reconstruction and segmentation of sparse MRI data. The multi-task learning for accelerated-MRI reconstruction and segmentation (MTLRS) model is the baseline for this evaluation. This cascade-structured model incorporates multiple RIMs and Attention U-net modules informing each other through hidden states. Various multi-task approaches are assessed using uncertainty and predictive performance metrics, compared to MTLRS without a multi-task connection (JOINT). This study uses the Stanford Knee MRI with Multi-task Evaluation (SKM-TEA) dataset, which includes 155 3D Cartesian sampled Double Echo Steady State (DESS) knee scans with meniscus and cartilage segmentation labels, undersampled with an $8 \times$ 2D undersampling Gaussian mask. Each multi-task approach is evaluated using reconstruction and segmentation metrics and an estimated quantitative T2 error metric. Uncertainty quantification is applied to identify variability within the unrolled network.

The multi-task approach using a module with spatially adaptive semantic guidance (**SASG**) significantly improves reconstruction metrics compared to the JOINT approach. The Tukey honest significance difference (Tukey HSD) test demonstrated that, at a 95% confidence level, $p < 0.001$ for all metrics. Additionally, uncertainty estimation of intermediate predictions shows faster convergence to lower uncertainty with the **SASG** approach, indicating a positive impact on reconstruction.

This research demonstrates that the **SASG** approach enhances the performance of the MTLRS model, which shows its validity across various unrolled multi-task reconstruction and segmentation networks. Furthermore, the proposed validation method could guide performance assessments of other unrolled network architectures.

Contents

1 Introduction

MRI is a powerful and noninvasive medical imaging modality widely used to visualise the internal structures of the human body. The increase in diagnostic imaging in current healthcare systems has a significant negative impact on both time and costs [1]. This has created a demand to accelerate the diagnostic pathway. Currently, acquisition, reconstruction, and medical image analysis are often treated as separate, sequential processes, which increases the length of time and limits the ability to introduce techniques for learning between tasks. Combining multiple MRI-related tasks could improve diagnostic performance owing to shared representation and combined inference [2].

MRI visualises internal structures by utilising magnetic relaxation properties captured in the spatial frequency domain known as k-space. During acquisition, k-space is affected by hardware limitations and noise, which the forward model does not fully describe. MRI reconstruction aims to produce an image from the acquired data while minimising noise effects. This process is often defined as an ill-posed inverse problem [3], meaning that an exact solution might be impossible. Instead, one minimises the difference between the exact and the recovered image.

MRI acquisition can be accelerated by reducing the number of k-space measurements, directly affecting the time needed. Reconstructing undersampled k-space data increases the difficulty of the inverse problem due to the limited measurements. Reconstruction models incorporating regularisation can be used for accelerated MRI reconstruction, as they are capable of learning from prior examples [4].

The RIM has been proposed as a deep learning solution to address inverse problems by directly learning the inference procedure. The RIM is an unrolled network that iteratively updates the initial representation using recurrent neural network (RNN) architectures [5]. It performs updates by evaluating the inverse problem's log-likelihood gradient without explicitly evaluating the prior. Unlike other unrolled networks, such as the End-to-End-variational network (E2E-VN) [6] and MoDL [7], the RIM learns the prior of the update instead of the whole scan [8].

The prior term is based on the available data used during the model's training phase. To create a generalised solution, a comprehensive representation of real-world data is needed. Introducing a related task implements additional data within the network's learning process, enhancing understanding of real-world situations.

A related task for image reconstruction is tissue segmentation. Tissue segmentation is crucial for detecting pathological abnormalities and contains valuable texture information. Furthermore, Adler et al. suggest that the optimisation formalism in segmentation and reconstruction have similarities, indicating that joint optimisation could improve both tasks [9]. Joint image reconstruction and segmentation have been proposed, but many lack the full advantages of combined inference. Models like RecSeg [10], SegNet [11], and IDSLR [12] do not focus on propagating the segmentation into the reconstruction unrolled network. Studies in other fields suggest that incorporating the segmentation output within the physics reconstruction network could be beneficial [13].

Karkalousos et al. presented a Multi-Task Learning with joint Reconstruction and Segmentation model (MTLRS) for accelerated MRI reconstruction and segmentation [14]. The model uses a cascade of RIMs and Attention U-nets for the joint reconstruction segmentation task. Their research showed significant performance increases when combining both networks within each cascade. The most competitive approaches, joint and multi-task, differ due to a connection between the segmentation output and the RIM's memory layers.

However, the evaluation technique showed the effect of the added multi-task connection to be insignificant. Only one multi-task technique was implemented, making it hard to determine whether the method or the evaluation technique fell short. Therefore, this study focuses on two research questions:

1. Which evaluation technique can indicate whether the multi-task connection significantly improves the MTLRS performance?
2. Which multi-task approach is most beneficial for the MTLRS model?

This study proposes an extensive evaluation method using three validation metrics and two internal uncertainty quantification methods to evaluate predicted performance and stability within the prediction formation.

Furthermore, four alternative multi-task approaches are proposed, presenting a selection of existing techniques used in various multi-task architectures. All four approaches will be compared with the joint and multi-task approaches suggested in the MTLRS paper on the Stanford Knee MRI with Multi-Task Evaluation (SKM-TEA) dataset [15].

2 Technical background

This chapter provides essential background information for understanding the principles used in this research. Due to the complexity of the data and models, a general understanding of physics and numerical definitions is needed.

2.1 Magnetic resonance imaging (MRI)

MRI is a technique for visualising tissues within the body by measuring the effects of magnetic relaxation, a property related to hydrogen bonds within tissue molecules. The MRI signal originates from hydrogen nuclei, each containing a single proton exhibiting a quantum mechanical property known as spin. Conceptually, spin can be visualised as a charge distribution rotating around an axis, creating a magnetic dipole moment aligned with the spin's direction (Figure 2A) [16].

Initially, the spatial orientation of individual proton dipole moments is random, resulting in zero net magnetisation. However, when a magnetic field (B_0) is applied, it influences the alignment of some proton dipole moments (Figure 2B). Although most remain randomly oriented, the alignment of a few protons creates a positive net magnetisation, which is the vector sum of all individual dipole moments, proportional to the applied magnetic field strength [17]. The proton spin precesses about the external field axis at a frequency known as the Larmor frequency.

The external magnetic field establishes an equilibrium net magnetisation. Applying a 90-degree radio-frequency (RF) pulse (B_1) perpendicular to the static magnetic field (B_0) excites the net magnetisation out of its equilibrium state. The RF pulse, which varies over time and rotates at the Larmor frequency, affects protons precessing at the Larmor frequency, rotating the magnetisation from the z-axis into the transverse plane, thereby creating transverse magnetisation (Figure 2C).

Following the RF pulse, relaxation occurs in two dimensions: longitudinal relaxation (T_1), where the net magnetisation returns to equilibrium, and transverse relaxation (T_2^*), where the transverse magnetisation dephases. These relaxation properties are tissue-dependent and are used to create different types of weighted MRI images (e.g., T1-weighted, T2-weighted), highlighting various tissue characteristics and providing distinct contrast in MRI signals.

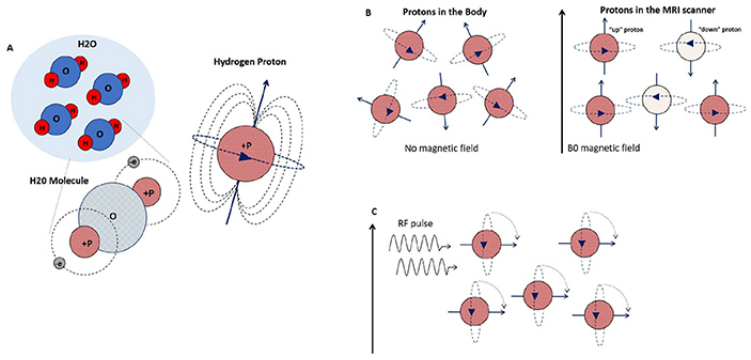


Figure 2: A: Schematic visualisation of the dipole moment and proton spins; B: Orientation of protons without a magnetic field vs. orientation with B_0 applied; C: Orientation of proton dipole moments and spins after an RF pulse [18].

2.1.1 K-space formalisation

In MRI, gradients are used to encode the spatial position of each signal. As visualised in figure 3, these gradients induce varying precession frequencies and phases within the homogeneous magnetic field. After excitation, proton spins exhibit unique precession frequencies and different phases based on their local

magnetic fields. Detecting these unique frequency and phase signatures allows the system to localise and distinguish signals from different locations, constructing a spatial frequency domain called k-space.

These gradients control movement through k-space, where the gradient amplitude determines the speed and the gradient direction determines the movement within k-space. Furthermore, these gradients make the selective acquisition of certain lines within k-space possible during acquisition.

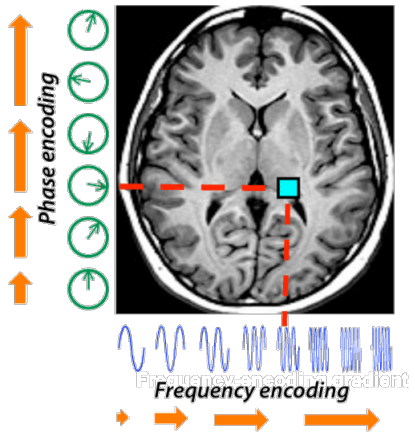


Figure 3: The employment of phase and frequency encoding, utilised for the localisation of magnetisation signals within the scan [19].

2.2 (Accelerated) MRI reconstruction

The acquired k-space represents the scanned object but is also affected by acquisition choices and inherent noise. Mathematically, the relation is described by the forward model shown in equation (1), where the measurement y (k-space) is a combination of the signal of interest x (scan) affected by the forward operator A and an additive noise vector n [20].

$$y = A(x) + n \tag{1}$$

2.2.1 Inverse problem

MRI reconstruction aims to recover faithfully the signal of interest while minimising the effects of noise by identifying appropriate parametric models. This inverse problem is often ill-posed, meaning multiple solutions may exist due to uncertainties in the measurements [3]. Instead of directly recovering the actual object, an objective function minimises the difference between the recovered and the actual object. The error of least squares is commonly used in MRI reconstruction, as it is well-suited for normally distributed noise. The inverse problem for MRI reconstruction is represented in (2), where $\hat{x}$ denotes the optimal solution, $A(x)$ is the forward operator applied to x, and y represents the noisy measurements.

$$\hat{x} = \arg\min_{x} \frac{1}{2}||A(x) - y||_2^2 \tag{2}$$

For Cartesian-sampled MRI, the forward operator $A(x)$ typically involves a Fast Fourier Transform (FFT) followed by sub-sampling of the Fourier coefficients [21]. The original object can be approximated with $\hat{x} = \frac{1}{N}A^*(y)$, assuming all conditions of fully Cartesian sampling are satisfied, including adherence to the Nyquist criterion [21]. Here, A^* is the adjoint operator, and N is the number of k-space measurements. This approximation can be evaluated using the inverse Fast Fourier Transform (iFFT).

Accelerated MRI focuses on increasing the speed of the MRI acquisition process by minimising the total k-space sampling time. Over the past decades, various sequence strategies have been proposed to accelerate MRI, reducing the time, by efficiently measuring in k-space [22]. This research, however, focuses on

techniques that target minimising the total number of k-space measurements, thereby decreasing the overall sampling time. This approach affects the Cartesian grid, making the iFFT inadequate for retrieving the actual image. Since parametric optimisation of the normal inverse problem can lead to noise amplification and overfitting, a new problem statement is formulated with an additive regularisation term [21] (3).

$$\hat{x} = \arg\min_{x} \frac{1}{2}||A(x) - y||_2^2 + \lambda R(x) \tag{3}$$

The variational model shown in (3) illustrates the interplay between the data consistency term $\frac{1}{2}||A(x)-y||_2^2$, which aims to recover the image $\hat{x}$ by aligning with the forward model, and the regularisation term $R(x)$. The regularisation term incorporates prior knowledge about the likely appearance of the actual image, helping to constrain the solution and mitigate the effects of noise and overfitting.

Various forms of regularisation can be employed. A commonly chosen regularisation term is the Total Variation (TV), which approximates the image gradient using finite differences [23]. Alternatively, neural networks can determine this regularisation term, often through trainable parameters within convolutional layers [4].

2.2.2 Deep learning accelerated MRI reconstruction

Deep learning accelerated MRI focuses on reconstructing sparsely sampled data into high-resolution scans using neural networks [24]. Compared to numerical approaches, such as compressed sensing (CS), deep learning avoids complicated optimisation-parameter tuning and performs faster during inference [25].

Deep learning accelerated MRI reconstruction can be performed using several approaches. Two distinct approaches are supervised and unsupervised learning, which differ based on the available data. In the case of unsupervised learning, the training does not rely on paired data with known targets. Instead, it aims to learn patterns and structures within the data [26]. In supervised learning, the training process uses paired data with known targets. A sampling mask is applied to create undersampled k-space data, simulating the real-world scenario of sparse sampling. The model is trained to reconstruct the target from the undersampled data, effectively learning the inverse of the undersampling process. The supervised reconstruction of sparsely sampled k-space can be seen as an extended forward model in which the undersampling mask is included, as shown in figure 4.

The inverse problem for supervised reconstruction is depicted in figure 4. This example shows the inverse problem for sensitivity-encoded (SENSE) k-space. It is important to mention that all available information is included inside the forward model, as described in (4). Here, the forward model describes the formulation of coil-specific and sparsely sampled k-space y_i, in which the linear forward operator A models the sub-sampling process of multi-coils data. P represents the sub-sampling mask selecting a fraction of the total number of measurements, $\mathcal{F}$ the FFT projecting the image into k-space, and ϵ is the expand operator aiming to transform the image of interest, x, into coil-specific images x_i. These coil-specific images are based on a specific coil's sensitivity map S_i.

$$\begin{aligned} y_i &= A(x) + n_i \\ A &= P \circ \mathcal{F} \circ \epsilon, \\ \epsilon(x) &= (S_0 \circ x, ..., Sc \circ x) = (x_0, ..., x_c) \end{aligned} \tag{4}$$

The adjoint operator A^* can then be given by (5), with $\mathcal{F}^{-1}$ the iFFT, P^T the transpose of the utilised mask and ρ representing the reduce operator, which transforms the multi-coil images into the actual scan. Here, H represents the Hermitian complex conjugate of the sensitivity maps.

$$\begin{aligned} A^* &= \rho \circ \mathcal{F}^{-1} \circ P^T, \\ \rho(x_0, ..., x_c) &= (\sum_{i=1}^{c} S_i^H \circ x_i), \end{aligned} \tag{5}$$

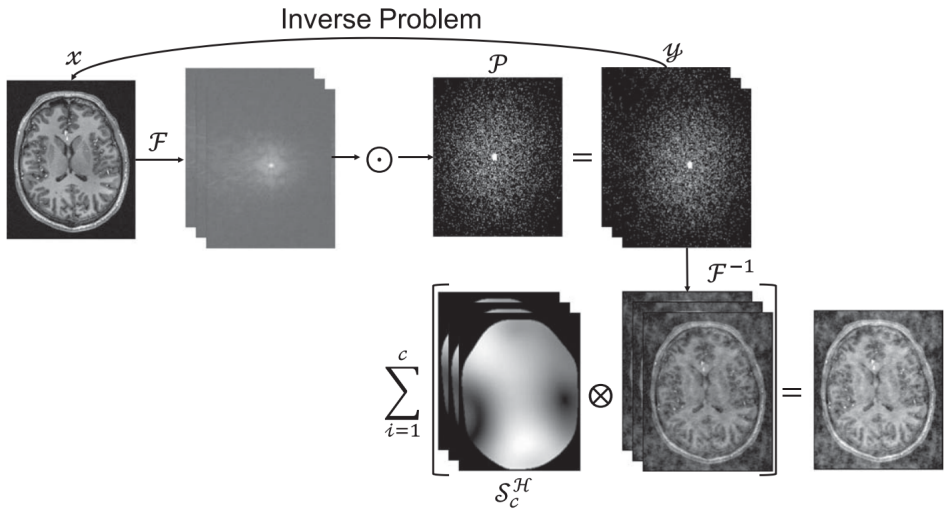


Figure 4: Schematic overview of the forward model for supervised MRI reconstruction of undersampled for sensitivity encoded MRI [27]. Here x reflects the target scan, $\mathcal{F}$ the FFT, which is the linear relation between the target and the Cartesian fully sampled k-space, P the undersampling mask, y the undersampled k-space, $\mathcal{F}^{-1}$ the iFFT, reflecting the linear relation between the undersampled k-space and the aliased scan. Lastly, S_c^H the Hermitian complex conjugate sensitivity maps decode the coil-specific reconstructed aliased scan.

However, when the adjoint operator is applied to the sparsely sampled k-space, the reconstructed image is affected by the undersampling mask, resulting in an aliased image. Therefore, a deep learning-driven parametric model can be utilised to retrieve the actual image.

2.2.3 Recurrent inference machine (RIM)

One approach for solving the supervised inverse problem for MRI reconstruction of sparsely sampled k-space is the RIM. The RIM is a generic physics-based unrolled model designed to solve inverse problems using a recurrent neural network [8]. The RIM aims to solve the inverse problem by optimising the Maximum A Posterior (MAP) estimate, as shown in equation (6). This approach is similar to the variational model presented in (3). However, while both approaches solve inverse problems, the MAP approach is inherently Bayesian, focusing on maximising the posterior probability [28]. The variational model utilises a deterministic approach and minimises the difference between observed and predicted data [29]. It is important to note that the functions used to determine posterior probability in the MAP approach are deterministic and similar to those in the presented variational model.

$$\hat{x} = \arg\max_{x} \left(\log p(y|x) + \log p(x) \right) \tag{6}$$

Equation (6) illustrates the MAP approximation, where $\hat{x}$, the best approximation, is determined based on the likelihood distribution of the forward model $p(y|x)$, and the parametric prior $p(x)$, which regulates the most likely appearance of the MR image.

The RIM is an unrolled network that refines the initial reconstruction using an iterative structure. Compared to alternative methods, the RIM primarily focuses on providing a gradient update to the initial prediction, gradually improving the estimate over each iteration. This iterative refinement allows the RIM to enhance progressively the reconstructed image's quality.

$$x_{t+1} = x_t + \gamma_t \nabla (\log p(y|x) + \log p(x))(x_t) \tag{7}$$

This recursive process is represented in (7), which determines the estimation of the next prediction, x_{t+1},

based on the current prediction x_t. The update is guided by a gradient, regulated by the learning rate γ_t, and the MAP approximation $\nabla(\log p(y|x) + \log p(x))(x_t)$ based on the current prediction.

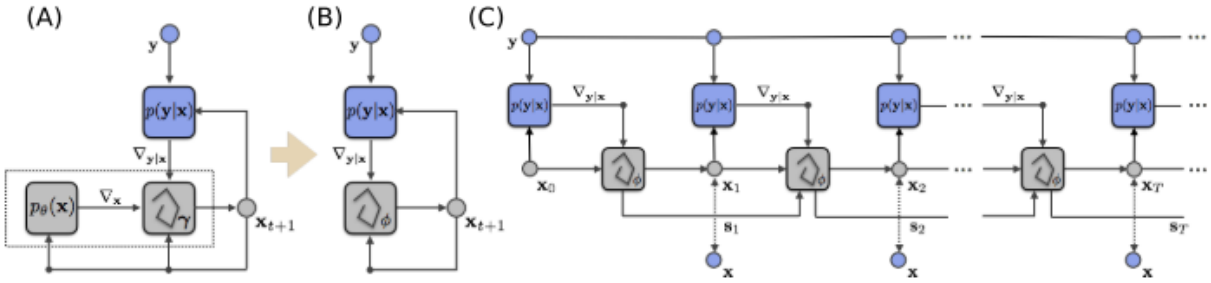


Figure 5: Schematic overview of the recurrent structure of MAP estimation: (A) The three boxes represent likelihood model $p(y|x)$, the parametric prior $p_{\theta(x)}$, and update function γ. In each iteration, likelihood and prior collect the current estimate of x_t to send a gradient to γ. γ then produces a new estimate of x. Grey boxes represent internal, data-independent modules, while blue boxes represent external, data-dependent modules. (B) Model simplification. Components $p(x)$ and γ are combined into one model with trainable parameters ϕ. The model recurrently produces new estimates through feedback from $p(y|x)$ and previous updates. (C) A Recurrent Inference Machine is unrolled in time where a hidden memory state s is added. During training, estimates at each time step are subject to an error signal from the ground truth signal x to perform backpropagation [8].

As visualised in figure 5, the RIM consists of three main components: the likelihood driven by the forward model $p(y|x)$, the parametric prior for the intended solution $p_{\theta(x)}$, and the update function γ. In each iteration, the model collects the current estimate of x_t, computes the gradient of the combined likelihood and parametric prior, and then updates the estimate of x. To simplify this approach, the prior and the update functions have been combined into a single model with trainable parameters.

This simplifies the optimisation task because formulating a parametric prior can be extremely challenging due to the broad variability in MR images. The RIM's recursive approach allows the model to evaluate the gradient of the parametric prior rather than the parametric prior itself. Furthermore, the likelihood derived from the forward model is treated as an input to the RIM, which means no additional trainable parameters are required to describe this relationship.

The recursive function can now be rewritten into a function containing the combined trainable parameters (8). Here $\nabla_{y|x_t}$ is a simplified notation of $\nabla logp(y|x)(x_t)$, ϕ the combined trainable parameter based on the parametric prior parameter θ, and learning rate γ. Equation (9) visualises the input of the update function. This shows that the RIM is solely optimised on the basis of the gradients of the parametric prior. Furthermore, $g\phi(\nabla_{y|x_t}, x_t)$ explicitly defines the likelihood as an input of the RIM, reflecting the extrinsic information (forward model) that is injected into the model. The deterministic formulation of the gradient of the log-likelihood function, $\nabla_{y|x_t}$, is given by (10).

$$x_{t+1} = x_t + g_\phi(\nabla_{y|x_t}, x_t) \tag{8}$$

$$g\phi(\nabla_{y|x_t}, x_t) = \gamma_t(\nabla_{y|x_t} + \nabla_{x_t}) \tag{9}$$

$$\nabla_{y|x_t} := \frac{1}{n^2} A^*(A(x_t) - y) \tag{10}$$

Lastly, the RIM also contains latent memory, a typical characteristic of recurrent networks. This allows for memory storage for tracking progression and curvature, and determining a stopping criterion during training to prevent overfitting. The recursive function is adapted to (11) incorporating the hidden states of

the memory layers. Here g_ϕ^* is the update function for h, the hidden states, and g_ϕ is the update function of x, the predictions. Also, both initial statements have been added, where A^* is the adjoint operator.

$$\begin{aligned} h_0 &= 0, & x_0 &= A^*(y), \\ h_{t+1} &= g_\phi^*(\nabla_{y|x_t}, x_t, h_t), & x_{t+1} &= x_t + g_\phi(\nabla_{y|x_t}, x_t, h_{t+1}), \end{aligned} \tag{11}$$

Due to the nature of storing memory, RNN's perform well on out-of-distribution data [30]. However, compared to the deep end-to-end schemes, such as E2E-VN [6], the RIM performs slightly less due to the limited number of iterations required to avoid gradient instability. Therefore, a cascade-like architecture was designed to increase the number of iterations without vanishing or exploding gradients.

Karkalousos et al. proposed a method that combines multiple RIMs in a cascade formation [27]. This makes building a deep RNN network possible, increasing the model's number of trainable parameters. In this way, the RIM will act as a regulator updating the prediction with the update function shown in (12). Here, x_{k+1} is the reconstructed scan of the next cascade, and x_k is the intermediate reconstruction. It is essential to mention that the RIM commonly uses a gated recurrent unit (GRU), which is vulnerable to exploding or vanishing gradients [31]. However, the RIM used within the CIRIM is based on independently recurrent neural networks (IndRNN). The IndRNN solves gradient problems by making the neurons independent and constraining the recurrent weights. This helps to maintain a stable gradient flow over long sequences, making it easier to train very deep RNNs [32].

$$x_{k+1} = x_k + \lambda R_{RIM_k}(x_k) \tag{12}$$

This adaptation has made the CIRIM competitive with other deep learning-driven unrolled networks. Compared to other state-of-the-art unrolled networks, CIRIM is the only one that explicitly defines the forward model as an input and learns exclusively from the gradient of the parametric prior.

2.3 Multi-task learning

MRI reconstruction primarily focuses on retrieving the actual image from the measured data. However, post-processing tasks such as segmentation, classification, and pathological detection are crucial for medical decision-making within the diagnostic pathway. Therefore, it would be valuable to integrate post-reconstruction optimisation tasks with the inverse MRI reconstruction problem.

Multi-task learning focuses on using shared representations between tasks to enhance the performance of each individual task [2]. Therefore, it is crucial to identify whether the tasks have conflicting needs, which could result in destructive interference (negative transfer), which means the improvement of one task damages the performance of the other task [33].

Adler et al. proved that task adaptive reconstruction, which integrates post-processing steps in the inverse problem optimisation, improves the performance of the model [9]. Furthermore, they showed that classification and segmentation are suitable additional tasks due to their Bayesian optimisation approach, like the inverse reconstruction problem [9]. Figure 6 shows the workflow of clinical MRI, in which the recovered model parameter undergoes several downstream processing steps before being integrated into decision-making processes. A task-adaptive model combines the last three steps into one parametric model.

$$\begin{aligned} \hat{R} &: y \to x \\ \hat{T} &: x \to s \\ \hat{\mathcal{T}} \circ \hat{\mathcal{R}} &: y \to x \end{aligned} \tag{13}$$

In the case of segmentation, this can be stated as in (13). Here $\hat{R} : y \to x$ is the explicit forward operator of a trainable neural network for the reconstruction task, $\hat{T} : x \to s$ is the explicit forward operator of a trainable neural network for the segmentation and $\hat{\mathcal{T}} \circ \hat{\mathcal{R}} : y \to s$ is the combined forward operator of the reconstruction and the segmentation task.

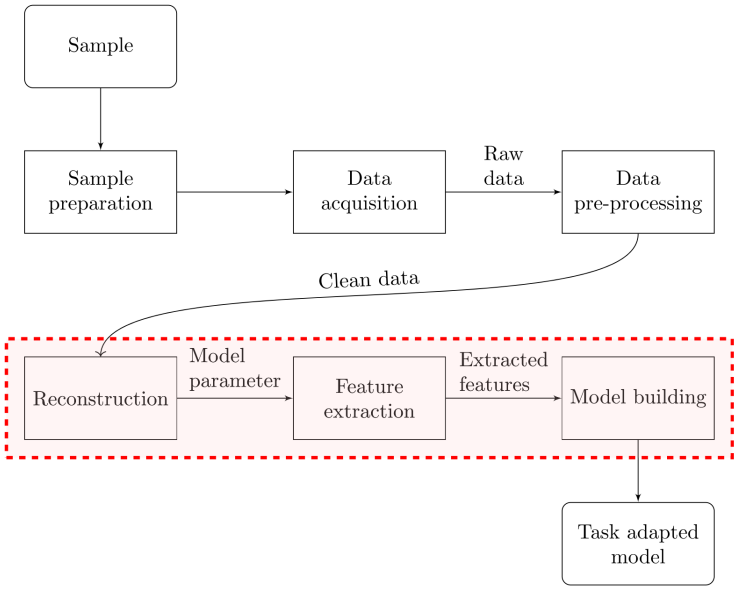


Figure 6: Typical workflow involving an inverse problem. The second row represents the data acquisition, where raw data are acquired and pre-processed, resulting in cleaned data. In the third row, the cleaned data are used as input to a reconstruction step that recovers the model parameter, which is then post-processed to extract features used as input for model building. The outcome is a task-adapted model that can be used for decision-making. The dotted part outlines steps that are unified by task-adapted reconstruction [9].

When combining both forward operators, different optimisation approaches can be applied. Three optimisation approaches can be applied: Sequential optimisation based on separate loss functions, end-to-end optimisation based solely on the adapted task loss function, and joint optimisation using a combined loss function with weighted factors. All these methods rely on the assumption that optimising one task benefits the other task. According to the results presented in Adler et al. research, joint optimisation results in better post-processing tasks (segmentation and classification) performance. Furthermore, alternative papers presenting joint MRI reconstruction segmentation networks show similar results [10], [11] [12].

However, one fundamental design choice has been made within these task adaptive networks. The segmentation task is identified as a post-processing step, reflecting only an output. Therefore, the segmentation task cannot be used as an informative operator for the reconstruction task. Utilising the segmentation output to enhance the reconstruction's performance could improve the overall performance. Multi-task architectures, where information sharing is applied between both networks, can be a valid solution to achieve this.

2.3.1 Multi-task networks

When developing multi-task architectures, it is vital to encourage information sharing that enables positive transfer while reducing the sharing of information that leads to negative transfer [34]. Therefore, the degree of shared information has to be fine-tuned to the specific needs of the combined optimisation task.

Hard parameter sharing is widely used in multi-task learning to improve efficiency and reduce overfitting by learning a common representation that generalises well across tasks. Hard parameter sharing involves the use of the same trainable layers for all tasks, while task-specific layers are used for each task's output. This approach reduces the risk of overfitting and enhances learning by leveraging commonalities between tasks [2].

However, due to the fundamental differences between tasks, hard parameter sharing may not always be appropriate, particularly when tasks have conflicting requirements. In such cases, soft parameter sharing

can be more effective, as they have their own trainable layers but share information through that encourage similarities between extracted features. This method allows for precisely detecting shared representation to enhance the overall performance. [35].

In the case of reconstruction and segmentation, the tasks are sequential, making it challenging to create a network where parameters are shared. However, in the case of cascade architectures, the output of one cascade can be connected to the next, ensuring information sharing between the two tasks [36]. Furthermore, this also leverages the ability to integrate attention modules, enhancing shared representation between tasks [37].

Karkalousos et al. suggested a Multi-Task Learning Reconstruction Segmentation (MTLRS) model [14]. This model uses the CIRIM as a backbone and incorporates an Attention U-Net within each cascade [38]. The optimisation strategies mentioned in Adler et al. were tested and evaluated [9]. Furthermore, an additional multi-task approach was added, which propagates each segmented intermediate prediction into the memory layers of the next cascade.

However, based on the results presented in the paper, no significant difference was found between the joint and the multi-task approaches. Since only one multi-task technique was implemented, and the evaluation approach was only based on two performance metrics, it is hard to argue whether the integrated approach benefited overall performance. Therefore, this research will implement different multi-task approaches and identify which is most beneficial for MRI reconstruction and segmentation within the MTLRS architecture.

2.4 Uncertainty Quantification

Using deep learning approaches in medical imaging has significantly improved speed and image quality [39]. However, deep learning models can be highly biased towards their training data or protocols, leading to misleading performance evaluation [40]. Within deep learning models for MRI reconstruction, Antun et al. have shown that small perturbations in the acquired k-space can result in significant artefacts, which could lead to wrong diagnosis [41]. As long as these perturbations affect current inverse solvers, an effective means to quantify the risk of failures is needed [42]. Uncertainty quantification is a proven framework to asses the malfunction of deep learning models predictions [43]. Predictive uncertainty exists out of two distinctive uncertainties: aleatoric and epistemic uncertainties.

Aleatoric uncertainty is an effect of the used data [43]. Provided data always includes bias and invalid measurements, ultimately affecting the trained models. This uncertainty is inherent to the data distribution, which makes it unabated. When assessing the model's uncertainty, the aleatoric uncertainty is always included.

Epistemic uncertainty refers to the lack of knowledge within the model. This type of uncertainty results from how the model interoperates and learns from the given data. There are two types of epistemic uncertainty: structural and parametric. Structural uncertainty focuses on the model's architecture, whereas parametric denotes the uncertainty in estimating the parameters under a specific specification [42].

3 Method

This research utilises the MTLRS model to evaluate different approaches of multi-task connections. The cascade structure of the reconstruction model (RIM) and the segmentation model (AttentionUnet) leverage the ability to propagate combined information into a following cascade. This research's key focus is on the connection of combined information between the segmentation and reconstruction modules. Several methods for combining and propagating information between these networks have been drawn from the literature and implemented in the MTLRS architecture. These will be thoroughly tested and evaluated with the help of the SKM-TEA dataset.

3.1 Dataset: SKM-TEA

SKM-TEA dataset is used in this research [15]. This dataset contains raw and processed DESS knee MRI scans. The dataset also includes soft tissue segmentations, pathological detection, and an analytical method to obtain T_2 magnetic tissue information from the knee. The dataset consists of two tracks: the DICOM track, which is only used for segmentation, detection and quantification, and the RAW Benchmark track, which is used for reconstruction and multi-task implementation.

The SKM-TEA dataset consists of 155 patient scans, roughly 25,000 slices. These patients represent sixteen pathological categories across joint effusion and meniscal, ligament, and cartilage lesions.

3.1.1 Provided datatypes

The raw MRI data consist of a target $3D$ reconstructed scan, a hybridized k-space (one dimension is already converted into image domain (x, k_y, k_z)), a sensitivity map estimated for the (k_y, k_z) plane, and a mask generating synthetic sparsely sampled k-space in the (k_y, k_z) plane. The k-space datatype contains eight or sixteen coil-specific k-space measurements; the number of coils varies between patients. Furthermore, each k-space object consists of two echoes, which can be used to estimate an analytical tissue-specific T_2 measurement. Figure 7 presents an overview of the data in the dataset for a specific patient.

3.1.2 Analytical T_2 estimation

The DESS sequence used for acquiring the SKM-TEA dataset has a physical property in which the transversal relaxation time T_2 can be determined based on the two measured echoes. Due to the relation between the two acquired echoes, T_2 estimation can be performed.

Sveinsson et al. propose an extended phase graph model to develop a linear approximation between the two echoes [44]. This linear relation can be used for accurate T_2 estimation. The equation described in (14) demonstrates this relation is based on acquisition parameters.

$$\frac{S_2}{S_1} = e^{-\frac{-2(TR-TE)}{T_2}} \frac{1 + e^{-\frac{TR}{T_1}}}{2} \tag{14}$$

Here, TR reflects the repetition time, TE the echo time, S_2/S_1 the fraction between the two echo signals and T_1 the longitudinal relaxation time. In the case of very long T_1, the second term becomes redundant.

The SKM-TEA dataset utilizes this linear operation to determine the T_2 value of specific soft tissue regions based on the two echoes, TR, TE, and the T_1 of the meniscus, which is derived from the literature [45].

3.2 Data preparation

The provided dataset is extensive and provides all the information needed to train a deep-learning model. However, the transversal (x,k_y,k_z) reconstruction approach limits the segmentation performance due to the limited number of slices containing multiple classes (Figure 8). Therefore, it would be advisable to reconstruct the raw data in the sagittal (k_x,k_y,z) plane. Unfortunately, due to the hybridised provided k-space, data transformation must be applied before the reconstruction over the sagittal plane can be performed.

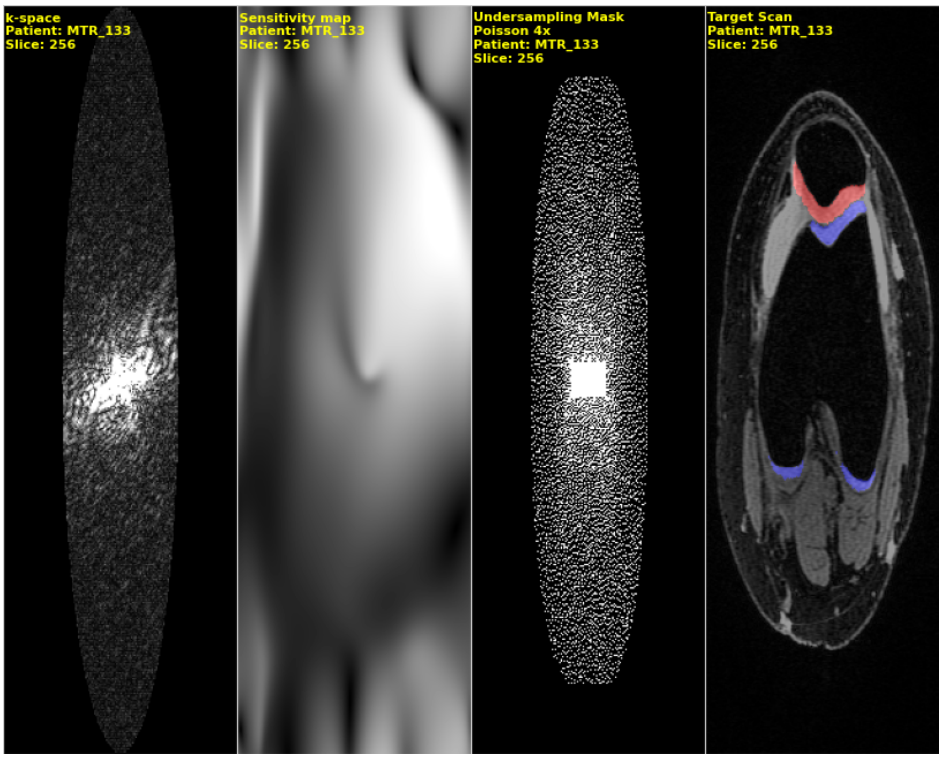


Figure 7: Overview of the provided dataset of a specific patient and slice. These images reflect the dataset over the transversal axes with a resolution of 512×160 pixels.

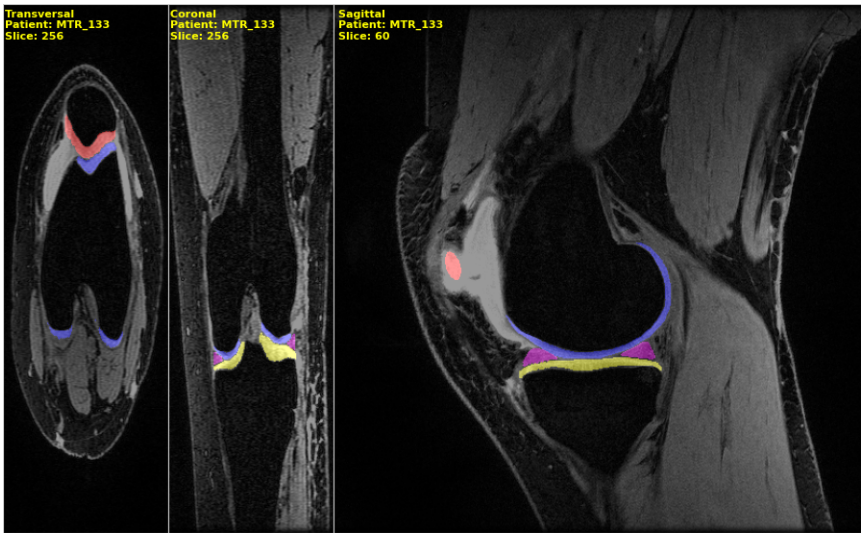
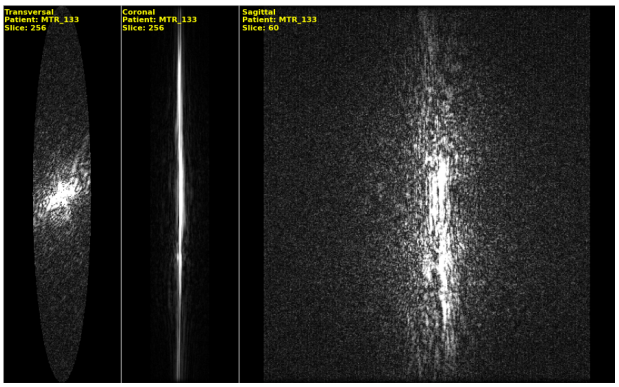


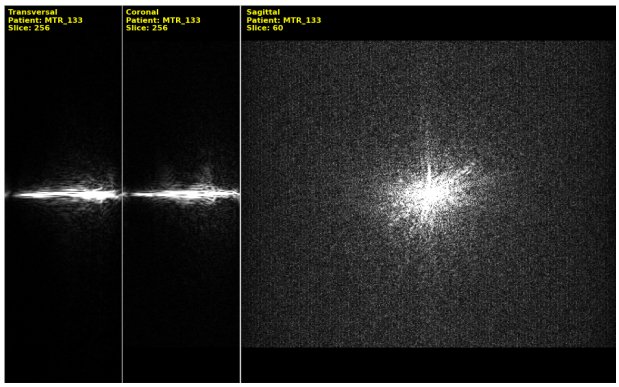
Figure 8: Target scan with segmentation labels in transversal, coronal and sagittal plane [15]. The Sagittal plane shows all the soft tissue classes within one slice. Furthermore, each class is represented by relatively big areas. Alternative orientations contain little or no slices with all the segmentation classes, and each class is represented only by small areas. The red label indicates the patellar cartilage, blue indicates the femoral cartilage, yellow is the tibial cartilage, and purple is the meniscus.

3.2.1 Transformation

The SKM-TEA dataset was acquired in a multi-coil setting with 2x1 parallel scanning (Accelerated MRI technique) [15]. A fully sampled k-space was synthesised from the undersampled k-space data. Afterwards, an iFFT was applied to generate the hybridised k-space alongside the readout dimension ($k_x \rightarrow x$). In this research, the readout dimension will be over the k_z axis. Therefore, a data transformation is needed. An FFT is performed to transform x to the frequency domain (k_x). Afterwards, the iFFT is applied to hybridise the k_z-axis into a spatial domain (z), resulting in a k-space which can be sagittal reconstructed.



(a) K-space before transformation where the k-space data can be described by (x, k_y, k_z). The transversal orientation is a visualization of k_y and k_z, the coronal of x and k_z, and sagittal k_y and x.



(b) K-space after transformation where the k-space data can be described by (k_x, k_y, z). The transversal orientation is a visualization of k_y and z, the coronal (2) of k_x and z, and sagittal k_y and k_x.

Figure 9: K-space orientation: (a) the original k-space orientation, (b) the k-space orientation after transformation. Here, k_i indicates if an axis is in the spatial-frequency domain, with i being a specific axis. Here, x reflects the axis running from the inferior to the superior side, y from the anterior to the posterior side, and z from the left to the right side.

3.2.2 Cropping and interpolation

The data have a size of $512 \times 512 \times 160$ voxels for each patient; when reconstructing over the z-axis, each image to reconstruct consists of a total of $512 \times 512 = 262,144$ pixels. This can be computationally heavy, especially when using a deep neural network with many trainable parameters. This research focuses on providing more insight into multi-task approaches within the MTLRS. Since many models will be trained, lowering the resolution increases the speed of inference and training. Furthermore, the results are saved as a $3D$ representation, and lowering the number of slices will significantly reduce the memory needed for storage. The k-space was cropped into a size of $256 \times 256 \times 80$ voxels by removing the outer lines, which leaves solely the inner center of the k-space. The cropping is applied to the unhybridized k-space (k_x, k_y, k_z).

Since the k-space consists of frequency information, removing the outer frequencies affects the data by lowering the resolution of the reconstructed scan. The segmentations and sensitivity maps do not represent frequency information. Therefore, cropping can not be applied. Interpolation using the nearest-neighbour method transforms the same data representation into a smaller size and, as a consequence, obtains bigger voxels. However, the sensitivity maps were acquired in the (y,z) plane, which results in imperfect estimation within the (x,y) plane. As visualised in figure 10, the sensitivity maps contain slice artefacts, the horizontal lines in the image. This was not a result of the interpolation since the same artefacts were visual in the high-resolution sensitivity maps. However, this is a result of the sensitivity maps being estimated in a different orientation.

The provided mask is only usable within the transversal plane due to the $2D$ undersampling formation. Therefore, the mask will be designed with the help of a designed masking function [14]. During training, a random $2D$ Gaussian mask with a centre-fraction of 0.02 and acceleration factor of 8 will be used to create a

sparsely sampled k-space. During inference, the Gaussian mask is identical for each patient to ensure valid comparison between different trained models. The mask ensures that each patient only contains 12.5% of the original fully sampled k-space. Figure 10 shows the transformed data and the synthesised mask used within this research.

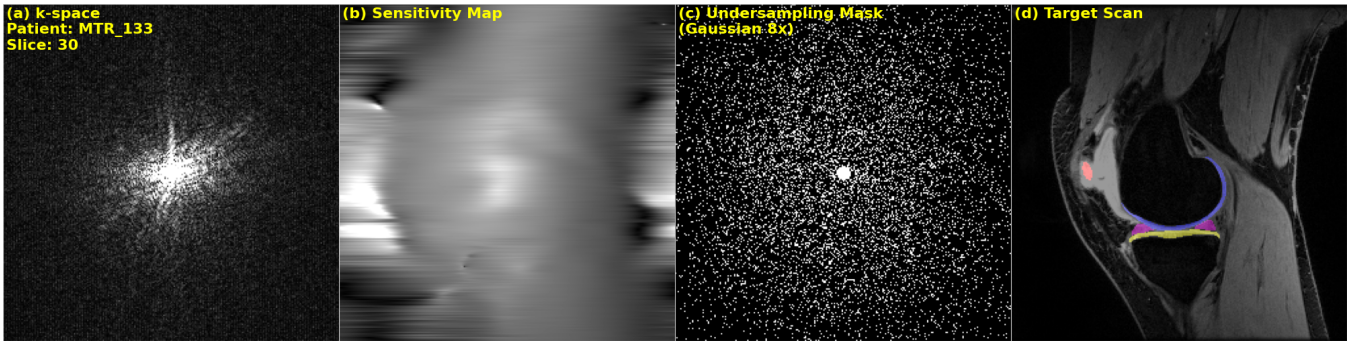


Figure 10: Data representation of transformed and resized data: (a) reflecting the cropped k-space, (b) the sensitivity map interpolated with the nearest neighbour, (c) the generated undersampling mask based on Gaussian distribution and (d) the reconstructed target scan (linear reconstruction as presented in Figure 11) with interpolated segmentation by nearest neighbour.

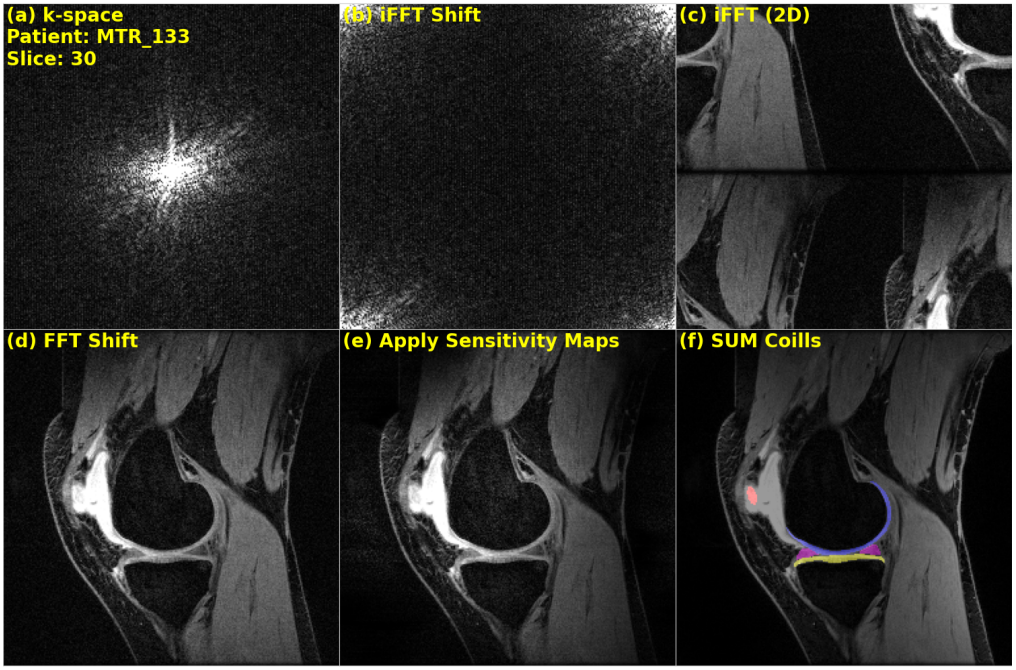


Figure 11: Reconstruction of a single slice from hybridized cropped k-space (k_x,k_y,z) to a 2D image. The process includes: (a) k-space data acquisition, (b) iFFT Shift for frequency positioning, (c) 2D iFFT, (d) FFT Shift, (e) application of sensitivity maps using Hermitian conjugated transformation, and (f) combining individual coil-specific images by summation.

After applying the transformations, the k-space can be reconstructed along the sagittal axis. The reconstruction of the hybridised k-space along the sagittal dimension can be performed using the discrete 2D iFFT and the interpolated sensitivity maps, which are implemented in the forward model of MTLRS (Figure 11).

3.3 MTLRS Architecture

The MTLRS network is a cascade architecture with multiple reconstruction and segmentation modules. Each cascade sequentially predicts an improved intermediate reconstruction and a segmentation. The architecture is similar to the CIRIM, in which each cascade acts as a regulator to update the predictions gradually [27].

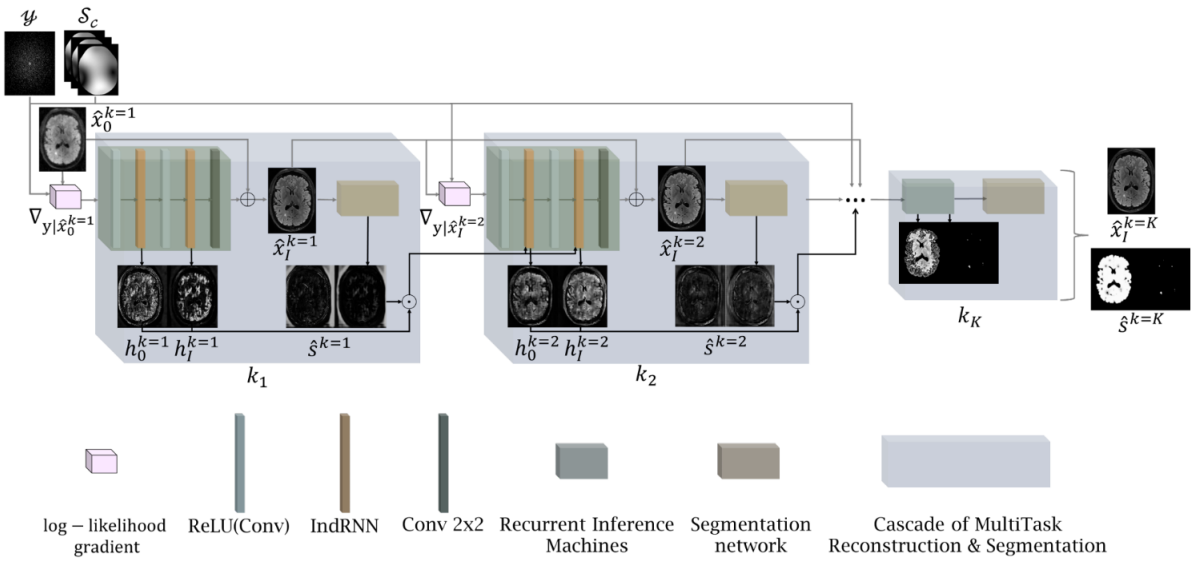


Figure 12: Schematic overview of the MultiTask Learning for accelerated-MRI Reconstruction and Segmentation (MTLRS) framework. MTLRS consists of K cascades (blue), a RIM reconstruction network (green) and an Attention Unet segmentation network (yellow). On each cascade, the network first performs a reconstruction $\hat{x}_I^k$, then a segmentation $\hat{s}^k$, which are integrated with the output of the hidden states (h_0^k and h_I^k). The hidden states are used to initialise the memory layers of the next cascade. After K cascades, the final reconstruction $\hat{x}_I^{k=K}$ and segmentation $\hat{s}^{k=K}$ are obtained [14].

Furthermore, the MTLRS incorporates a connection between the segmentation and the reconstruction modules, which includes the segmentation output within the update function of the reconstruction hidden states. Equation (15) shows how the segmentation output is incorporated within the RIM's hidden states. Here $\hat{s}^{k-1}$ is the segmentation output of a previous cascade, $\hat{x}_I^{k-1}$ the intermediate prediction of the RIM in the previous cascade, h^{k-1} is the output of the memory layer of the previous cascade (hidden states), and $\nabla_{y|\hat{x}_t}$ represents the gradient term of the forwards model log-likelihood. Since the segmentation is not available during the first cascade, this update function is only valid starting from the second cascade. The first cascade is based on the update function provided in (16).

$$\begin{aligned} h_0^{k\geq 2} &= \hat{x}_I^{k-1} * \hat{s}^{k-1} + h_I^{k-1} & \hat{x}_I^{k\geq 2} &= A^*(\hat{x}_I), \\ h_{t+1}^{k\geq 2} &= g_\phi^* \left(\nabla_{y|\hat{x}_t^k}, \hat{x}_t^k, h_t^k \right), & \hat{x}_{t+1}^{k\geq 2} &= \hat{x}_t^k + g_\phi(\nabla_{y|\hat{x}_t^k}, \hat{x}_t^k, h_{t+1}^k), \end{aligned} \tag{15}$$

$$\begin{aligned} h_0^{k=1} &= 0 & \hat{x}_I^{k=1} &= A^*(y), \\ h_{t+1}^{k=1} &= g_\phi^* \left(\nabla_{y|\hat{x}_t^k}, \hat{x}_I, h_t \right), & \hat{x}_{t+1}^{k=1} &= \hat{x}_t + g_\phi(\nabla_{y|\hat{x}_t}, \hat{x}_t, h_{t+1}), \end{aligned} \tag{16}$$

The operator $*$ that combines the segmentation with the intermediate predictions is not clearly stated in the MTLRS paper [14]. Also, the incorporation of this combined output within the hidden states (+ operator) is not clearly explained. More insight was gained from the provided codebase into these specific operations. From now on, this combined operator function will be formulated as the multi-task operator: $M(\hat{x}_I^{k-1}, \hat{s}^{k-1}, h_I^{k-1}) \rightarrow h_0^{k\geq 2}$. First, the multi-task operation will be explained as suggested in the MTLRS paper. Subsequently, alternative approaches based on techniques presented in previous research will be suggested.

3.3.1 MTLRS multi-task connection

As described above, the multi-task operator consists of two different operations: formulating the combined output based on the intermediate prediction and segmentation, and incorporating the combined output with the hidden states.

The combined output aims to create a variable with shape characteristics similar to the hidden states. Therefore, it is important to understand the shape of each variable:

Hidden states of the RIM (h_I^{k-1}): $[[B, V_1, H, W], [B, V_2, H, W]]$

Intermediate prediction ($\hat{x}_I^{k-1}$): *[B,H,W]*

Segmentation prediction (s^{k-1}): *[B,C,H,W]*

Here, B indicates the used batch size, V_i the specific output number of feature channels used in the convolution layer within the memory states, H and W the height and the width of the image and C the total number of segmentation classes.

As is clearly visible, the second dimension is unequal for all the variables, significantly increasing the difficulty of faithfully combining the three variables. Furthermore, the segmentation output does not reflect intensity-related information, but it reflects logits indicating the class importance of each pixel.

The combined output, $\hat{x}_I^{k-1} * s^{k-1}$, can be seen as a reconstruction-like feature map, in which the segmentation is used to indicate importance within the reconstructed image. As visualised in (17), the output is called the Synthetic Segmentation Features (SSF). The SSF represents a list containing the same number of tensors as in the hidden states with identical second dimensions, reflecting the number of feature channels.

$$\mathrm{SSF}_{\mathrm{logit}}^{k-1} = \left[\|^{\frac{v}{C}} \left(|\hat{x}_I^{k-1}| \cdot s^{k-1} \right) \middle| \, v \in V \right] \tag{17}$$

Here, $\|$ is the concatenation over the second dimension $\frac{v}{C}$ times, $\hat{x}_I^{k-1}$ reflects the complex intermediate prediction from the RIM module with an added channel dimension, s^{k-1} segmentation without the background class, v the number of channels for a specific convolution memory layer, C the total number of segmentation classes and V a list containing all the numbers of channels for each hidden state.

The propagation of the $\mathrm{SSF}_{\mathrm{logit}}^{k-1}$ with the hidden states by adding the tensors in the list (18)

$$h_0^{k\geq 2} = \{ h_{0,i}^{k\geq 2} \mid h_{0,i}^{k\geq 2} = h_{I,i}^{k-1} + \mathrm{SSF}_{\mathrm{logit},i}^{k-1},\ i = 1, 2, \ldots, \|V\| \} \tag{18}$$

This specific multi-task connection, **SUM LOGIT**, is identical to the proposed multi-task in the MTLRS papers. Alternative approaches for propagating segmentation information into the hidden states of the RIM are described below.

3.3.2 Alternative multi-task connections

Before explaining the alternative multi-task connections, it is essential to note that the segmentation approach in this research has been modified with regard to the original method used in the paper. Initially, the MTLRS is implemented as a multi-label classification problem, in which each class was treated as a separate binary classification task. This approach allows multiple classes to be assigned to the same pixel.

However, classes cannot coexist within the same pixel for the dataset used in this research, described in section 3.1. The segmentation scope is adjusted to a multi-class semantic segmentation approach. An additional class is introduced to represent non-target classes, altering the total number of predicted classes.

This change enables the use of softmax probabilities for class determination. In this revised approach, each class is represented by a probability, which indicates the likelihood that a pixel belongs to that class. This adjustment ensures that each pixel is assigned to only one class based on the highest probability.

The first alternative multi-task approach utilises this probabilistic approach. Instead of propagating on the basis of segmentation logits, the softmax is used to formulate an intermediate segmentation probability map. The softmax function helps to maintain numerical stability by normalising the segmentation logits. Without this normalisation, the logits can be large or small, and propagation into the hidden states can lead to numerical instability.

Furthermore, within this multi-task approach, all the non-background classes are combined to create a continuous synthetic segmentation feature map in which each channel reflects the same information.

$$\mathrm{SSF}_{\mathrm{softmax}}^{k-1} = \left[\|^{v} \left(|\hat{x}_I^{k-1}| \cdot \sum_{c=2}^{C} \mathrm{softmax}(s^{k-1})_c \right) \middle| \, v \in V \right] \tag{19}$$

Here, $\|$ is the concatenation over the channel dimension v times, $\hat{x}_I^{k-1}$ reflects the complex intermediate prediction from the RIM module, $\mathrm{softmax}(s^{k-1})_c$ is the segmentation probability of a specific class, L is the number of channels for a specific convolution memory layer, C is the number of segmentation classes and F indicates a list containing all the number of channels for each memory layer. Note that $\sum_{c=2}^{C}$ only combines the foreground classes. Therefore, the sum can never equal 1, so each pixel of the synthetic segmentation map is scaled based on the segmentation output.

The combined output $\mathrm{SSF}_{\mathrm{softmax}}^{k-1}$ is added to the hidden states similar to the original method as shown in (20). This can be seen as the second approach: **SUM SOFTMAX**

$$h_0^{k\geq 2} = \{h_{0,i}^{k\geq 2} \mid h_{0,i}^{k\geq 2} = h_{I,i}^{k-1} + \mathrm{SSF}_{\mathrm{softmax,i}}^{k-1},\, i = 1, 2, \ldots, \|V\|\} \tag{20}$$

Subsequently, two alternative approaches are proposed for propagating information into the hidden states. These methods are both attention networks, which use convolution layers with trainable parameters to enhance shared information of two variables.

The Task-Attention Module (**TAM**) is proposed to enhance the correlated information within similar task features [37]. It utilises identical shaped and formulated features to enhance the common information (Figure 13). First, a balance tensor is created to represent the shared information of the two sources. This is performed by concatenating both features along the channel dimension and using convolutional layers with trainable parameters to reduce the channel dimension into the original size. Afterwards, a sigmoid function is applied to scale the tensor. The balancing tensor generates a balanced output of the two features. The balancing unit can be described as in (21).

$$\begin{aligned} b_1 &= \sigma(\psi_1((h \parallel \mathrm{SSF}), \theta_1)) \\ b_2 &= \psi_2((b_1 \cdot h \parallel (1 - b_1) \cdot \mathrm{SSF}), \theta_2), \end{aligned} \tag{21}$$

Here b_1 reflects the balancing tensor, ψ_i is a convolutional layer with trainable parameter θ_i, σ is a sigmoid operation, $h \parallel SSF$ reflects the concatenation of reconstruction and synthetic segmentation features over the second dimension and b_2 is the balance output features.

The balanced output features are then fed into a series of convolution and deconvolution layers. After a sigmoid transformation, a matrix Z is formulated, which can be seen as an attention map. The attention map is then used to enhance the specific features of the reconstruction.

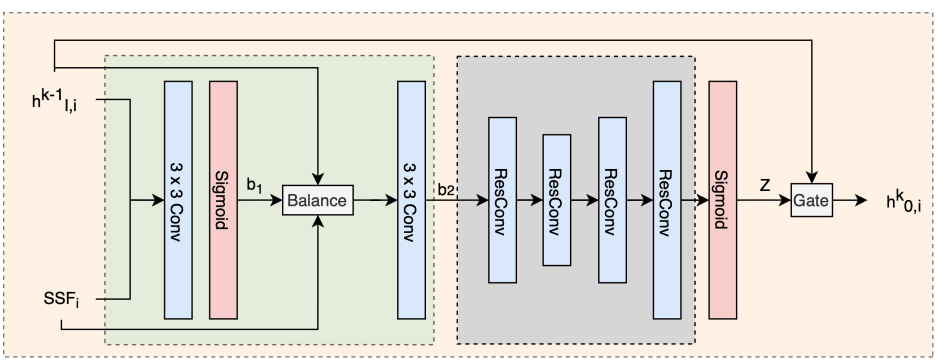
$$h_e = (1 + Z) \cdot h \tag{22}$$

The propagation of the enhanced reconstruction feature maps and the synthetic segmentation feature maps with the help of TAM is described in equation (25)

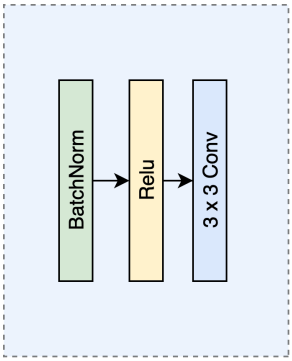
$$h_0^{k\geq 2} = \{h_{0,i}^{k\geq 2} \mid h_{0,i}^{k\geq 2} = TAM(h_{I,i}^{k-1}, \mathrm{SSF}_i^{k-1}),\ i = 1, 2, \ldots, \|V\|\} \tag{23}$$

Here, $h_{0,i}^{k\geq 2}$ is the memory layer list of a specific cascade, i is the index based on the length of V, the list containing all the filter numbers of the convolutional layers, SSF_i^{k-1}) are the synthetic segmentation features of a specific cascade and TAM reflects the task attention module.

This propagation approach has been applied for both SSFs, resulting in two distinctive approaches: **TAM LOGIT** and **TAM SOFTMAX**. A schematic overview of the TAM and the used convolutional layers is depicted in figure 13.



(a) The task attention module (TAM) with a balancing module (green) enhances the shared information within the segmentation and reconstruction feature channels. A residual block with convolutional and deconvolutional layers is used to obtain different spatial attention (grey). The attention tensor Z is used to enhance the original features of the hidden states with the attention map (Gate).



(b) **ResConv** module: This exists of a batch normalisation, ReLu activation and a convolution layer.

Figure 13: **TAM**: Task attention module.

The second propagation module takes a slightly different approach. The Spatially Adaptive Semantic Guidance (SASG) module does not use two sets of tensors with feature channels but a probability map of the segmentation to enhance the reconstruction features [46]. In this case, there is no need to build synthetic segmentation features (SSF). The segmentation logits are transformed into a probability map using a softmax layer, which is then propagated into the SASG module.

The module exists of 3×3 convolutional layers, LeakyReLU activation functions and instance normalisation. In the **SASG** module, a specific attention module is placed, which refines the reconstruction features based on the segmentation probability map. This module is based on equation 24

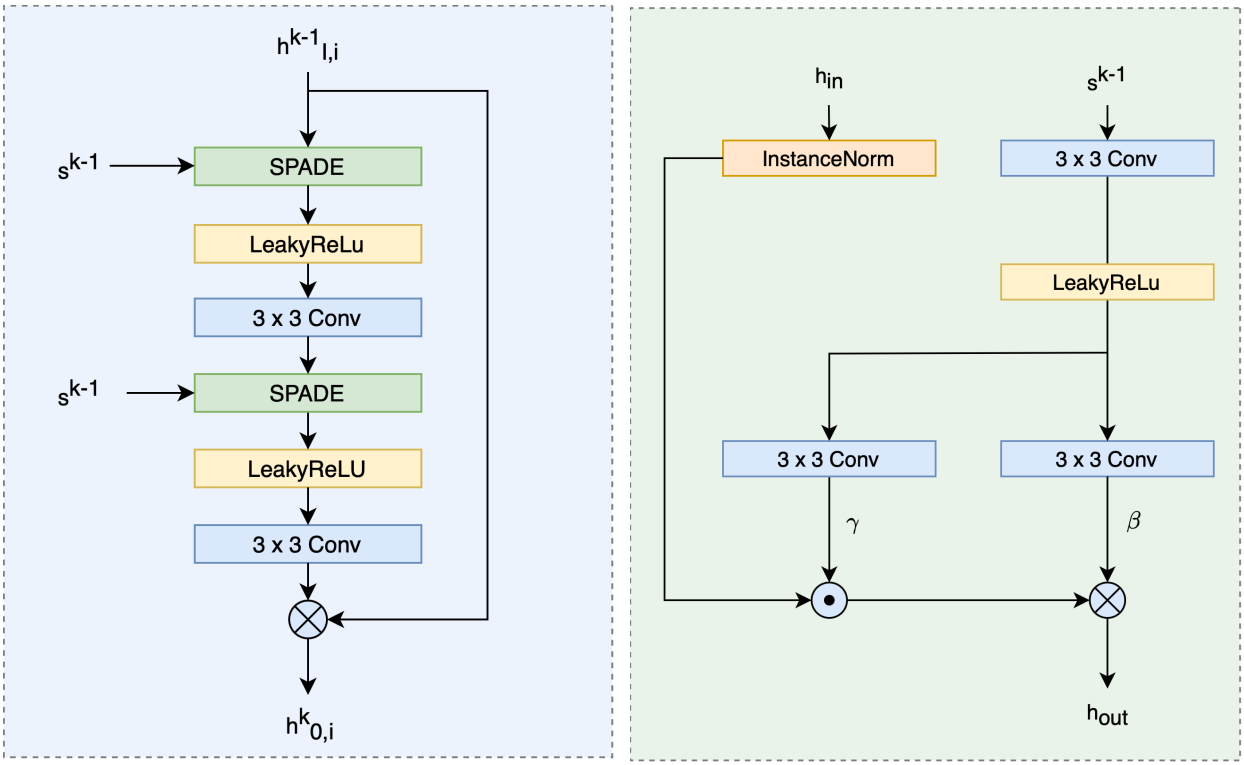
$$h_{out} = \gamma \odot \iota(h_{in}) + \beta \tag{24}$$

Here, γ is the scale factor, ι is the instance normalisation, h_{in} is the hidden state and β is the bias coefficient which is determined by multiple convolutional layers.

This attention module is used twice within the SASG module before a convolution and activation layer. The SASG's output is an enhanced representation of the reconstruction features. This module connects the reconstruction features of the memory layers and the segmentation output and propagates the output towards the next cascade.

$$h_0^{k\geq 2} = \{h_{0,i}^{k\geq 2} \mid h_{0,i}^{k\geq 2} = SASG(h_{I,i}^{k-1}, \mathrm{softmax}(\hat{s}^{k-1})),\ i = 1, 2, \ldots, \|V\|\} \tag{25}$$

An schematic overview of the **SASG** approach is depicted in figure 14.



(a) **SASG module:** Incorporated multiple convolution layers (blue), LeakyRelu activation (yellow) and SPADE module (green) to refine the hidden states features with the probability segmentation map.

(b) **SPADE module:** Incorporated multiple convolution layers (blue), LeakyRelu activation (yellow) and instance normalisation (green) to refine the hidden states features with the probability segmentation map.

Figure 14: **SASG** Spatially Adaptive Semantic Guidance module.

Both propagation modules, TAM and SASG, are initialised ones within the network. Therefore, the same trainable layers are used within each cascade. This ensures the focus lies on the inherent differences between the two module designs without the additional complexity of cascade-specific adaptation. It allows for a straightforward evaluation of each module's performance in a consistent and controlled environment.

3.4 Added principles to enhance segmentation output

Alongside the change in the segmentation scope, two additional strategies have been added to ensure consistency and confidence in the segmentation performance throughout the network. Since each cascade will propagate information based on the segmentation into the following cascade, it is essential to have a confident and consistent segmentation intermediate prediction.

3.4.1 Consistency within the segmentation task

Throughout the network process, it is apparent that the reconstruction performance gradually improves. However, for the segmentation case, it is valuable when the segmentation module can segment even non-perfect reconstructed scans. Otherwise, the false segmentation would affect the reconstruction of the next cascade. Furthermore, a more consistent segmentation output is valuable since this research aims to investigate various approaches of multi-task connections. Suppose the segmentation output significantly differs in each cascade. In that case it becomes harder to determine whether the multi-task approach affects the reconstruction performance or the segmentation output is the impacting factor. Therefore, segmentation consistency is applied to all the approaches. This ensures that each segmentation prediction is added to the segmentation output of the next cascade to maintain a consistent segmentation output. Mathematically,

this is described in equation (26)

$$S^k = \sum_{i=0}^{k} s^{k-i} \tag{26}$$

Here, S^k is the sum of the current and all previous segmentations, and s_{k-i} is the original segmentation of a specific cascade. This operation is implemented before the optional softmax operation to ensure valid probability maps if applied.

3.4.2 Calibration error and temperature scaling

Uncertainty can arise in classification problems such as object detection, segmentation, and class determination. One notable form of uncertainty is overconfidence, which occurs when the predicted class probabilities are optimised so that they do not accurately reflect the actual distribution of the data [47]. Overconfidence refers to the phenomenon in which predictive probabilities are pushed towards extreme values (0 or 1). A well-calibrated model should produce probabilities closer to 0.5 when it has difficulty identifying the correct class. This indicates that the segmentation model struggles to determine the pixel class accurately.

One method for overconfidence correction is temperature scaling (27). Temperature scaling uses a trainable parameter to scale the logits before performing a softmax [47]. This ensures that the predicted probabilities align with the data's distribution.

$$q_i = \frac{\exp(s_i/T)}{\sum_{c=1}^{C} \exp(s_c/T)} \tag{27}$$

For a specific class i, q_i is the scaled probability, s_i is the predicted segmentation logit, C is the total number of classes, and T is the parameter used to scale each class similarly.

Temperature scaling is applied to this research during the validation phase. At the end of each validation phase, starting at epoch four, the predicted segmentation logits of one patient are used to determine the additional parameter T. By changing the temperature variable, the cross-entropy loss over this specific patient is gradually reduced. The optimal value for T, which results in the lowest cross-entropy loss in this patient, is saved.

The temperature is then incorporated within the last layer of the segmentation model, and the logits are scaled before applying the softmax. During the initialisation of the training phase, this parameter is frozen again to ensure it is not changed during the training phase. The implementation is taken from the GitHub repository, as stated in Appendix G.

3.5 Loss functions and training strategy

Similar to the paper, the research utilises a two-fold reconstruction and segmentation loss. Here, both losses exist of a strategy to minimise the difference between the prediction and the target, and an optimisation strategy to maintain texture and shape. Unlike the paper, a categorical cross-entropy is used instead of a binary cross-entropy. In addition, the intermediate segmentations are integrated similarly to the intermediate reconstruction with a weighted loss. The weighted loss is scaled based on the sum of the weights instead of the total number of losses. This had to be changed since some training examples were affected by underflow (the total loss is smaller than the precision of the GPU).

$$L_{seg} = \beta L^{CE}(\hat{s}, s) + (1-\beta) L^{Dice}(\hat{s}, s) \tag{28}$$

$$L_{recon} = \beta L^{L1}(\hat{x}, x) + (1-\beta) L^{SSIM}(\hat{x}, x) \tag{29}$$

$$L_{joint} = \frac{1}{B} \sum_{b=1}^{B} (1-a) L_{recon}(\hat{x}_b, x_b) + a L_{seg}(\hat{s}_b, s_b) \tag{30}$$

The total loss can be described by (30), where the reconstruction loss is based on (29) and the segmentation loss on (28). The reconstruction loss consists of a L_1 loss, which minimises the difference between a predicted $\hat{x}$ and a target x scan and a structural similarity index measure (SSIM) loss, targeting the structural similarity between both scans. The segmentation loss consists of a categorical cross-entropy loss, which tries to minimise the differences between the probability of the predicted $\hat{s}$ and target s segmentation and a DICE loss, which focuses on creating similar shapes.

The MTLRS architecture is a recurrent structure in which each intermediate prediction is a gradually improved version of the previous one. Therefore, each intermediate loss is weighted with respect to the total number of recurrences. In the reconstruction case, this recurrent rate equals the number of iterations in the RIM module multiplied by the number of cascades in the network ($I \times K$). The cascade of RIMs can be seen as a regulator; therefore, the loss for each individual RIM is calculated first, and then the total loss based on the number of cascades is determined.

Originally the MTLRS paper suggests handling the L_1 loss as stated in (31); here K is the number of cascades, I is the number of iterations in the RIM, w_t reflects the calculated weights in the following way $w_t = 10^{-\frac{I-t}{I-1}}$, q is the total number of pixels in the scan, $\hat{x}_t^k$ is the predicted reconstructed scan within a specific cascade within a specific time step and x is the target scan. Based on this implementation, the total loss is significantly minimised, resulting in underflow for some specific slices.

$$L^{l1} = \frac{\sum_{k=1}^{K}(\frac{1}{qI}\sum_{t=1}^{I} w_t|\hat{x}_t^k - x|)}{K} \tag{31}$$

Therefore, we suggest dividing the summed loss by the sum of the weights instead of the number of iterations. An additional weight is added for each total cascade loss. This also prioritises the importance of the cascade's order. This approach is described in (32), with W_K and W_I reflecting the sum of the determined weights, which are calculated in the following order: $w_k = 10^{-\frac{K-k}{K-1}}$ and $w_t = 10^{-\frac{I-t}{I-1}}$. This weighted approach was also applied to the SSIM loss.

$$\begin{aligned} W_K &= \sum_{k=1}^{K} w_k \\ W_I &= \sum_{t=1}^{I} w_t \\ L^{l1} &= \frac{\sum_{k=1}^{K} w_k(\frac{1}{qW_I}\sum_{t=1}^{I} w_t|\hat{x}_t^k - x|)}{W_K} \end{aligned} \tag{32}$$

As mentioned above, intermediate segmentation is integrated with a weighted factor within the total segmentation loss. For each cascade, one segmentation is made. Therefore, the weighted categorical cross-entropy loss can be described in (33). Here, C stands for the total number of predicted classes, c is for a specific class and $\hat{s}_c^k$ for the predicted segmentation class probability.

$$L^{CCE} = \frac{\sum_{k=1}^{K} w_k(-\frac{1}{C}\sum_{c=1}^{C} s_c * \log(\hat{s}_c^k))}{W_K} \tag{33}$$

Including each intermediate prediction in the total loss stimulates the model to update the prediction gradually. This helps the model to maintain gradient flow and to become more generalised.

3.5.1 Optimiser

This research uses an Adam optimiser with a starting learning rate of $10-4$. An inverse square root schedule was used to gradually change the learning rate and prevent the model from overfilling or getting stuck at local minima.

3.6 Uncertainty Estimation Methods

Two different methods are implemented to estimate the epistemic uncertainty of each multi-task approach.

Deep Ensembles estimate the uncertainty based on different initialised neural network predictions. The individual-trained models are randomly initialised, meaning each parameter's weight is randomly chosen before the training phase. Since the determination of the parameter weights is stochastic, the final weights end up differently for each trained model. This results in variation between the predictions of each randomly initialised model. The standard deviation between these predictions can be used to determine the uncertainty within the specific approach. Lakshminarayanan, Pritzel, and Blundell argue that changing the weight initialisation, and training based on random data ordering are sufficient for estimating the epistemic uncertainty [48]. The deep ensemble approach is applied for each specific multi-task approach within this research. Furthermore, each approach's intermediate prediction is also saved; therefore, an uncertainty quantification can be determined for each cascade's intermediate reconstruction prediction $\hat{x}_I^k$.

Moreover, an alternative approach is suggested in this research for region-specific reconstruction uncertainty. The RIM generates multiple time step predictions $\hat{x}_t^k$ during training and inference. The standard deviation between all the time step predictions within one specific cascade indicates the behaviour and a possible epistemic uncertainty of the RIM. Suppose, the standard deviation within specific regions of the intermediate predicted scans is low; this would indicate the model is certain about this region since no updates are added to this specific region. On the contrary, a high standard deviation reflects a high uncertainty within this synthesis of this specific region. Lastly, we expect that the standard deviation within the last cascade would reflect the same epistemic uncertainty of the deep ensemble.

3.7 Experiments and validation

A total of six experiments are performed, varying in the propagation of the segmentation output in the hidden state. An overview is given in table 1.

Multi-task Approaches	Segmentation Output	SSF	Propagation Approach
JOINT	-	-	-
SUM LOGIT	logits	yes	+
SUM SOFTMAX	probability	yes	+
SASG	probability	no	SASG
TAM LOGIT	logits	yes	TAM
TAM SOFTMAX	probability	yes	TAM

Table 1: Overview of experiments showing the variation in the multi-task connection

The different approaches are assessed based on reconstruction, segmentation and T_2 evaluation metrics. Furthermore, uncertainty quantification is applied to investigate the epistemic uncertainty within the reconstructed scans.

3.7.1 Metric evaluation

The reconstructed scans are assessed based on various metrics, which utilise the target reconstruction to assess the quality of the predicted scan. The target scan is the reconstruction of a fully sampled k-space without using an undersampled mask. The applied metrics cannot process complex data, resulting in an assessment of the magnitude of both scans. Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Haar Wavelet-based Perceptual Similarity Index (HaarPSI) are used to quantify the performance of each approach. This is based on the $3D$ object of each patient.

The predicted segmentations are assessed based on segmentation metrics. Three metrics have been selected: Dice Similarity Coefficient (DICE), 95th percentile Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD). These metrics reflect different properties of the segmented area, and presenting all three provides an overview of the segmentation performance.

The previous metrics only quantify each task individually, which does not allow for quantifying the joint optimisation. Fortunately, this dataset can generate T_2 estimations based on the predicted reconstructed echoes and segmentation of the soft tissues. Evaluating the T_2 error between the target and the predictions can be used as a joint performance metric. The target and predicted T_2 will be determined for each patient, and the T_2 error for each soft tissue class will be assessed.

Uncertainty quantification of each multi-task approach is determined with the help of Deep Ensemble. All the multi-task approaches are trained five times with random initialised parameters, resulting in five models generating different predictions and metric scores. Based on these predictions, the mean and standard deviation metric score for each model is determined. This indicated the uncertainty of each multi-task approach based on the final prediction.

For each metric, a statistical analysis is performed to identify if a particular multi-task approach positively affects the performance of the MTLRS model. First, an ANOVA test is applied to determine whether the group of multi-task approaches significantly affects the metric's score. If this is valid, a Tukey Honest Significant Difference (HSD) test is used to identify which multi-task approach significantly improves the metric's score compared to the **JOINT** approach.

3.7.2 Intermediate reconstructions

The intermediate predicted reconstructions are determined for five different patients. These predictions are used to determine the quality of the intermediate reconstructions within the unrolled structure of the MTLRS

architecture. Furthermore, since each multi-task approach is initialised five times randomly, uncertainty quantification could also be applied (intermediate uncertainty).

Furthermore, the proposed uncertainty quantification method will be evaluated for these five patients. This method is not based on metrics but on uncertainty within the initialised pixels. Therefore, uncertainty within specific regions is visualised. The suggested approach will be visually compared to the deep ensemble to determine if the alternative approach can be used for region-specific uncertainty quantification.

4 Results

Figure 15 shows the reconstruction of an 8× Gaussian undersampling scan of a specific patient slice. The sparsely sampled k-space is reconstructed using linear reconstruction (zero-filled) and the six investigated reconstruction approaches. Furthermore, the metric scores of this particular slice have been determined and mentioned on the scan. Based on the figure 15, it can be determined that the reconstructions of all non-linear approaches are feasible and do not contain significant artefacts. The segmentation results are depicted in figure 16, showing a good segmentation performance of all the multi-task approaches. More visualisations have been added to Appendix A.1 and A.2.

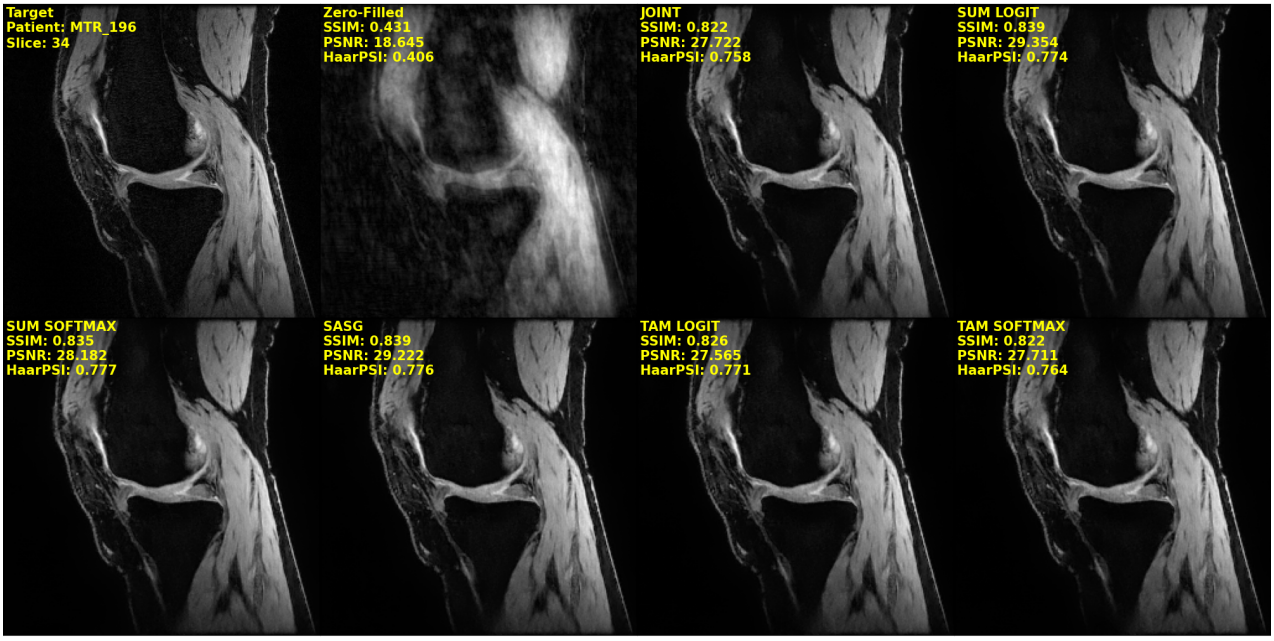


Figure 15: Visualisation of the first echo scan of a specific patient slice, starting from top-left showing the target scan, the zero-filled reconstruction based on the 8× Gaussian undersampling and then the various reconstructions of the different multi-task approaches. Furthermore, each approach's top left corner states the reconstruction's performance metrics.

4.1 Metrics performance of final predictions

All the scans within the test dataset are reconstructed and saved as a $3D$ scan. With the help of the suggested metrics, the metric scores are determined for each patient. First, the mean metric score of the whole test dataset is determined, and then the standard deviation between the random initialised models is calculated. This was done for the reconstruction, segmentation and T_2 evaluation metrics.

Table 2 contains the reconstruction metric, indicating that the **SUM LOGIT** and the **SASG** have an increased performance, compared to the **JOINT** approach, for all the reconstruction metrics. The mean increase is 0.017 for the SSIM, 0.8 for the PSNR and 0.1 for the HAARPSI. The **TAM LOGIT** and **TAM SOFTMAX** approaches show a decrease in metric performance, with a decline of 0.01 in SSIM, 0.02 in PSNR and 0.01 in HaarPSI. Furthermore, all three reconstruction metrics show similar behaviour across multi-task approaches, and the standard deviation within one model is not larger than the highest difference between the multi-task approaches.

Table 3, containing the segmentation metrics, shows that the **TAM LOGIT** approach contains the best mean segmentation score for all metrics (DICE: 0.797, HD95: 6.49, ASSD: 0.207). However, the standard deviation of all the metrics within one multi-task approach is higher than the difference between the multi-task approaches. For the estimated T_2 metrics, table 4 shows the mean estimated absolute T_2 for each target

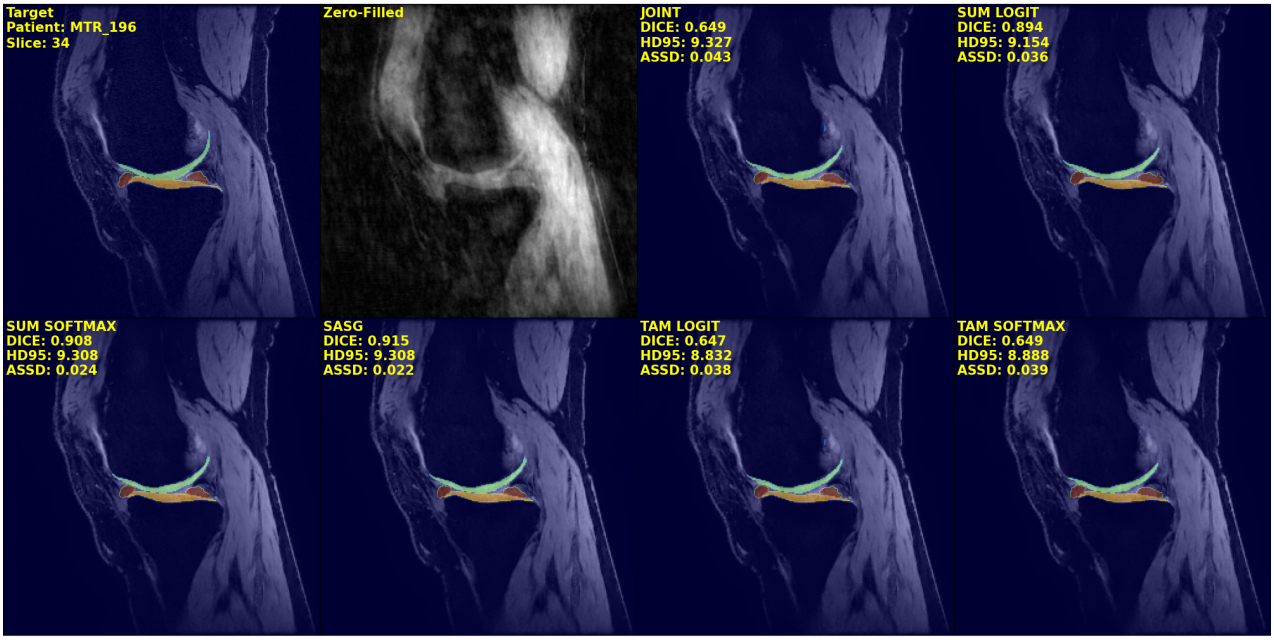


Figure 16: Visualisation of the first echo scan and the segmentation labels as an overlay, the target and the six multi-task approaches with three metrics to identify the segmentation performance. The light blue label indicates the patellar cartilage, the green the femoral cartilage, the orange the tibial cartilage, red the meniscus, and dark blue the background class. The segmentation performance metrics are stated in the top left corner of each individual image.

Multi-Task Approach	echo	SSIM	PSNR	HAARPSI
JOINT	1	0.869±0.003	29.344±0.29	0.770±0.007
	2	0.834±0.004	32.579±0.104	0.756±0.01
SUM LOGIT	1	0.878±0.005	29.616±0.287	0.777±0.009
	2	**0.852±0.007**	**33.385±0.572**	**0.795±0.015**
SUM SOFTMAX	1	0.871±0.01	28.96±0.858	0.773±0.017
	2	0.843±0.007	32.769±0.373	0.774±0.014
SASG	**1**	**0.886±0.002**	**30.158±0.116**	**0.780±0.007**
	2	0.852±0.006	33.228±0.285	0.787±0.01
TAM LOGIT	1	0.860±0.015	29.057±0.308	0.771±0.009
	2	0.822±0.021	32.359±0.354	0.756±0.023
TAM SOFTMAX	**1**	0.859±0.007	29.111±0.103	0.767±0.007
	2	0.817±0.025	32.198±0.527	0.731±0.037

Table 2: Reconstruction metrics per specific multi-task approach and echo. The mentioned standard deviation is determined based on the deep ensemble of each approach

Multi-Task Approach	DICE (↑)	HD95 (↓)	ASSD (↓)
JOINT	0.787±0.023	6.576±0.397	0.215±0.063
SUM LOGIT	0.783±0.007	6.69±0.125	0.251±0.031
SUM SOFTMAX	0.793±0.015	6.562±0.398	0.208±0.054
SASG	0.782±0.032	6.794±0.749	0.27±0.145
TAM LOGIT	**0.797±0.026**	**6.49±0.492**	**0.207±0.105**
TAM SOFTMAX	0.774±0.014	6.924±0.31	0.284±0.079

Table 3: Segmentation metrics per specific multi-task approach. The mentioned standard deviation is determined based on the Deep Ensemble of each multi-task approach.

Multi-Task Approach	MENISCUS	TIBIAL	PATELLAR	FEMORAL
JOINT	1.938±0.279	4.61±0.554	2.751±0.421	2.271±0.417
SUM LOGIT	**1.808±0.625**	**3.205±0.665**	**2.281±0.577**	**1.516±0.521**
SUM SOFTMAX	2.017±0.368	3.903±0.587	2.703±0.16	2.198±0.455
SASG	1.819±0.421	3.34±1.334	2.873±1.095	2.308±1.078
TAM LOGIT	1.815±0.46	4.918±0.616	2.556±0.497	2.989±1.048
TAM SOFTMAX	2.124±0.312	4.516±0.702	2.554±0.576	3.11±0.778

Table 4: Absolute error between target and predicted tissue-specific T_2 estimation (ms). The mentioned standard deviation is determined based on the Deep Ensemble of each multi-task approach. Typical cartilage T_2 values are 30-40ms, while meniscus T_2 values are 10-15ms) [15].

soft tissue. The table shows the error is the lowest for all soft tissues with the **SUM LOGIT** approach (MIN: 1.808, TC: 3.205, PC: 2.281, FC: 2.281). For some tissues, the standard deviation within one multi-task approach is higher than the difference between the multi-task approaches.

4.1.1 Significant Test

To indicate whether a multi-task approach significantly impacts upon the predictive result, an ANOVA and a Tukey Honest Significant Difference (HSD) Test (95%) were applied. The outcome of these tests can quantify the influence of the specific multi-task approach.

Table 5 shows the results of the ANOVA test based on the SSIM metric. It shows that the groups, such as the multi-task approach, patient, and echo, significantly influence the SSIM score. Furthermore, the combined group **multi-task approach:patient** is insignificant, whereas the other combined groups are significant. Since the multi-task approach is significant, it is feasible to use a Tukey HSD to determine the impact of each multi-task approach.

The Tukey HSD results based on the SSI metric are shown in table 6, indicating that multiple models are significantly different. The **SUM LOGIT** and the **SASG** are both significantly better, whereas **TAM LOGIT** and **TAM SOFTMAX** are significantly worse than the **JOINT** approach. Since we are interested in the performance compared to the **JOINT** approach, the pairwise test is performed against this model. For all the metrics used in this research, the ANOVA and Tukey HSD were performed and stated in appendix B.

	df	sum_sq	mean_sq	F	PR(>F)
multi-task approach	5	0.231	0.046	271.585	0
patient	35	1.039	0.030	174.369	0
echo	1	0.574	0.574	3370.240	0
multi-task approach:patient	175	0.026	0.000	0.871	0.879
multi-task approach:echo	5	0.0156	0.003	18.313	0
patient:echo	35	0.204	0.006	34.206	0
multi-task approach:patient:echo	175	0.014	0.000	0.455	1
Residual	1620	0.276	0.000		

Table 5: Results of the ANOVA test performed on SSIM reflect the variance source within this metric. Here, **df** is the degrees of freedom, reflecting the source of variation, **sum_sq** is the sum of squares for the specific source of variation, **mean_sq** is the mean of squares for the specific source of variation and **F** The F-value is a statistic that compares the variance between group means to the variance within groups. The **PR(>F)** indicates the probability of obtaining an F-statistic at least as extreme as the one calculated, assuming the null hypothesis is true. Values below 0.05 indicate that the observed differences among group means are unlikely to have occurred by chance. Therefore, the null hypothesis can be rejected and concluded that the factor has a significant effect on the dependent variable. **Note:** values stated as 0 are considered below 0.001.

multi-task approach 1	multi-task approach 2	mean_diff	p-adj	lower	upper	reject
JOINT	SUM LOGIT	0.015	0	0.008	0.022	TRUE
JOINT	SUM SOFTMAX	0.005	0.236	-0.002	0.012	FALSE
JOINT	SASG	0.017	0.000	0.010	0.024	TRUE
JOINT	TAM LOGIT	-0.011	0.000	-0.018	-0.004	TRUE
JOINT	TAM SOFTMAX	-0.008	0.014	-0.016	-0.001	TRUE

Table 6: Tukey Honest Significant Difference (HSD) test results based on the SSIM metric. This test is used to perform pairwise comparisons to determine whether the means of different groups are significantly different. Here, **multi-task approach 1 / multi-task approach 2** indicates which groups are being compared, **mean_diff** indicates the difference between the means of the two groups, **p-adj** is the adjusted p-value for multiple comparisons, **lower** and **upper** are the bounds of the 95% confidence interval for the mean difference, and **reject** indicates whether the null hypothesis of no difference between the group means is rejected (TRUE) or not (FALSE). **Note:** values stated as 0 are considered below 0.001.

4.2 Intermediate reconstruction performance

Five patients were selected and used to identify the predictive performance and deep ensemble uncertainty within the intermediate predictions. Figure 17 shows the SSIM score of the first echo of the five selected patients, where each individual box plot reflects the uncertainty range within each approach. This figure shows that the intermediate prediction of cascade two has the highest uncertainty for all approaches compared to the other cascades. Furthermore, it is evident that the **SASG** approach has the highest median and the lowest uncertainty range after the first cascade. Within appendix C, the intermediate predictions performance of the other two metrics show similar results.

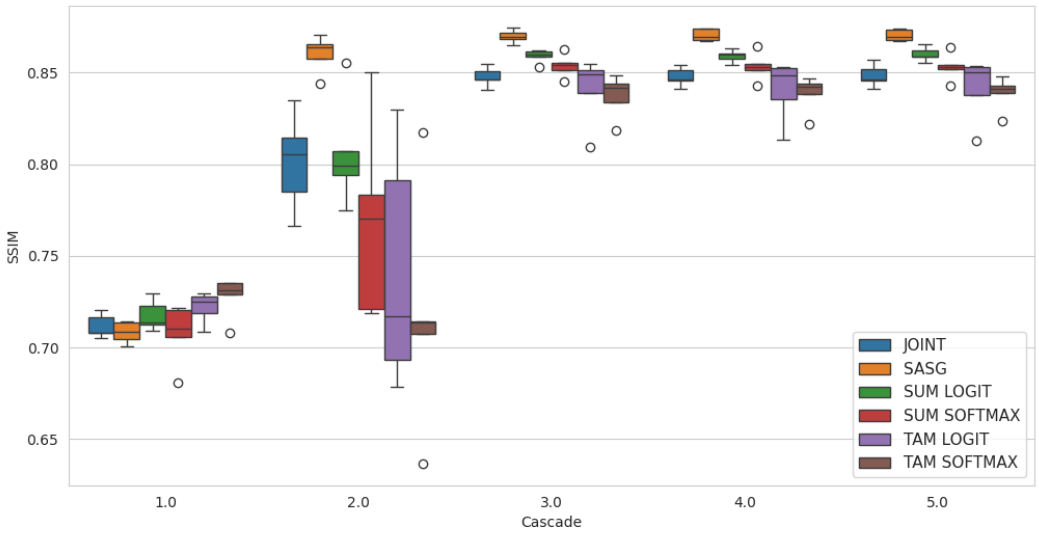


Figure 17: SSIM metric score of the intermediate reconstructions based on the first echo of five patients. Each multi-task approach was initialised five times, reflecting the measurement of predictive uncertainty.

Figure 18 visualises the intermediate uncertainty of a specific slice. This shows the change of uncertainty within specific regions over the intermediate predictions. It is visible that the first cascade for the most part is still uncertain about the background noise, whereas the second cascade shows a high uncertainty within the knee area. Within the second cascade, the knee's soft tissues and edges still show a higher uncertainty than other areas. Furthermore, the **SASG** approach has a lower overall uncertainty within the first and second cascade than the other approaches. However, for all the multi-task approaches, the uncertainty within the soft tissue in the last cascades is higher than the uncertainty depicted in the **JOINT** approach.

The proposed uncertainty quantification method, based on the standard deviation of the intermediate time steps, is assessed. This is done based on the visual comparison between figure 19 and figure 18. Both figures show that the first cascade is mostly focused on the background noise. However, besides that, the figures do not show similarities. Within the deep ensemble, the first two cascades show relatively high uncertainty, whereas the proposed method shows relatively high uncertainty within the first three cascades. Another notable thing is that both **TAM** approaches seem to have a lower uncertainty within the second cascade, but the uncertainty increases again in the third cascade (Figure 19). However, just like figure 18 shows, the **SASG** approach has a significantly lower uncertainty within a specific cascade than the alternative approaches. Lastly, the last cascade contains significantly lower uncertainties for all approaches within the proposed uncertainty method (Figure 19).

Based on the findings of these figures, the third and second cascades are visualised for multiple patients to identify whether these patterns are consistent in all patients. These figures can be found in Appendix D.

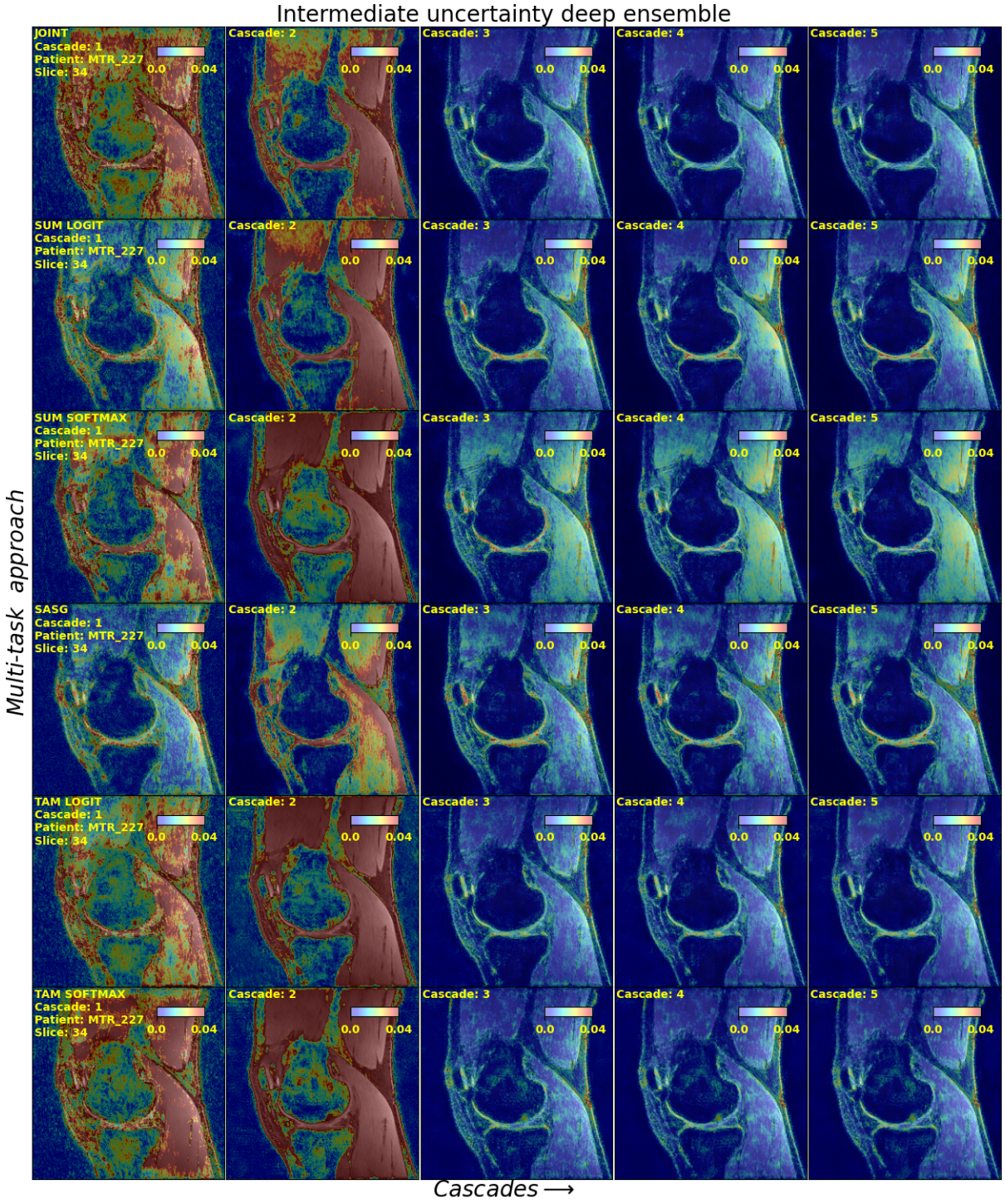


Figure 18: Visualisation of the intermediate uncertainty of the first echo. The uncertainty for each cascade has been determined based on the Deep Ensemble approach, in which each method was trained five times with random initialised parameters. Each row indicates a multi-task approach, and each column is a specific cascade within the MTLRS structure.

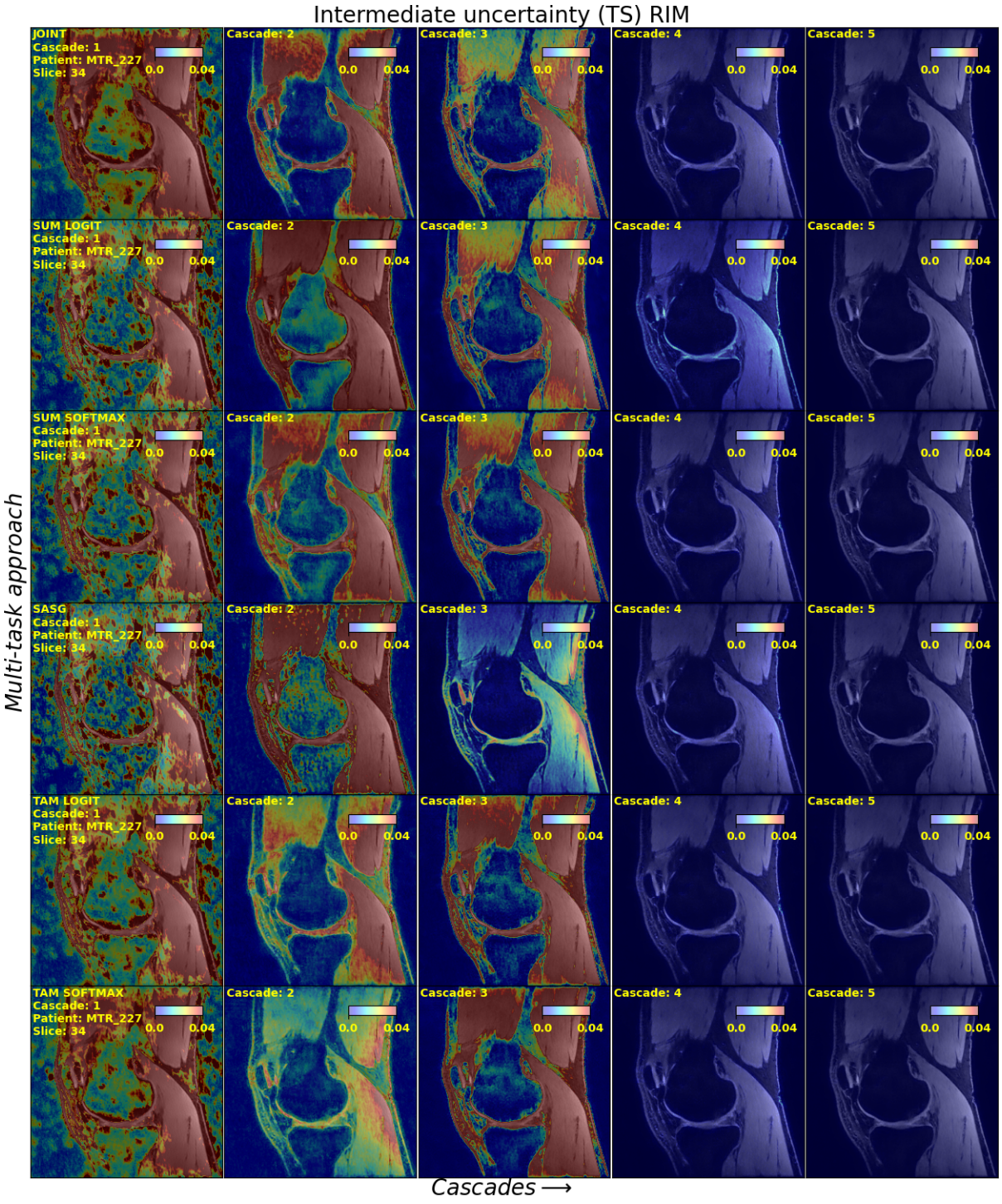


Figure 19: Visualisation of the proposed intermediate uncertainty of the first echo. The uncertainty for each cascade has been determined based on the intermediate time step, in which each time step prediction within a single RIM is used to determine the standard deviation. This shows whether the variations between the individual predictions within the RIM are similar or contain a high variation. Each row indicates a multi-task approach, and each column is a specific cascade within the MTLRS structure.

5 Discussion

The results of this study highlight the effectiveness of the **SASG** multi-task approach for the joint reconstruction and segmentation task in MRI. The reconstruction metrics and corresponding significance tests consistently demonstrate that the **SASG** model outperforms the **JOINT** approach in all selected reconstruction metrics (Table 6 and Appendix B.1). Furthermore, the **SASG** approach progressively improves the convergence within the cascade structure by improving the intermediate predictions over cascades while reducing uncertainty, compared to alternative approaches (Figure 17). These findings underscore the superiority of the **SASG** approach in enhancing MRI reconstruction quality.

In contrast, our results indicate no significant difference between the multi-task and **JOINT** approached when evaluating segmentation and T_2 metrics (Appendix B.2 and B.3). These results suggest that improved reconstruction does not directly translate into better segmentation outcomes and T_2 metrics in these segmented regions.

Due to the Bayesian approach used in the RIM, we expect the MTLRS model to show a decreasing trend in the deep ensemble (epistemic) uncertainty over the cascades. This hypothesis posits that an effective multi-task approach could accelerate this learning process, reducing intermediate uncertainty throughout the network. Our findings support this hypothesis, showing a noticeable decrease in intermediate uncertainty in the cascade predictions of the **SASG** approach (Figure 17). Furthermore, figure 18, reflecting the slice-specific intermediate uncertainty, shows an identical trend, supporting our hypothesis. Additionally, figure 30 in Appendix D indicates that the reduction in uncertainty for the **SASG** approach is not patient-dependent, but rather an inherent effect of the multi-task approach.

Moreover, based on the assumption that a better multi-task approach converges faster, it seems valid to assume that the deep ensemble uncertainty of the final prediction is lower than that of the joint approach. Contrary to our expectation, the deep ensemble uncertainty within the final reconstruction metrics, according to table 2, shows a slight increase. Furthermore, figure 18 shows a slight increase in uncertainty within the soft tissue regions. This can be related to the negative impact of the multi-task process on the update function of the last two RIM modules. The updates of the last two cascades are minimal, resulting in the multi-task connection having a high impact on the update function. Figure 33, within Appendix E, shows the standard deviation within the time step updates, showing the relatively high uncertainty on the soft tissue regions within the last step updates.

The proposed uncertainty epistemic quantification method, based on the standard deviation within the RIM-specific time step predictions, does not reflect the same behaviour as the deep ensemble uncertainty. The last cascade was expected to show behaviour similar to that of the deep ensemble approach; however, the standard deviation within the last cascade is much lower than that of the deep ensemble. Therefore, this approach can not be used for epistemic uncertainty quantification. However, it might be valuable for visualising the model's focus on specific regions in particular cascades in the unrolled network.

5.1 Relation to previous work

Within this work, all the approaches utilise joint optimisation for reconstruction and segmentation. Based on previous research by Karkalousos et al. and Adler et al., it is evident that joint optimisation leads to an improved performance of the segmentation task. Within this research, we found that the multi-task approach significantly affects the reconstruction and has no significant effect on the segmentation. Therefore, a combined reconstruction and segmentation model with joint optimisation and the **SASG** multi-task approach is assumed to significantly improve the overall performance.

Karkalousos et al. recommended a cascade configuration of five RIM modules [27]; however, our research findings indicate that a configuration of three modules might be sufficient for all approaches (Figure 17). This reduction in the number of cascades would significantly reduce the number of parameters within the MTLRS architecture and positively affect the inference and training time. Notable is that in this earlier work, a comparison was made over a large number of slices and subjects, introducing a large heterogeneity in the analysis and, therefore, small changes in SSIM over cascades [27]. However, within our research, the epistemic uncertainty was quantified with deep ensemble, showing that the metric scores of the last two

cascades do not exceed the uncertainty margins. Therefore, a setup with three cascades is recommended for the MTLRS model.

The SASG module was initially proposed by a paper describing a CS unrolled network for joint reconstruction segmentation [46]. Similarly to our implementation, this work incorporates the module for computing reconstruction feature maps and semantic segmentation probability maps. Therefore, it is valid to argue that the **SASG** module might be an effective multi-task approach for MRI reconstruction and segmentation in various unrolled networks.

5.2 Limitations

The nearest neighbour interpolation used to reduce the resolution of the segmentation labels has affected the results negatively. For some slices, the target labels do not reflect the actual cartilage within the scans (Appendix A.2). The nearest neighbour interpolation approach has diminished the visibility of thin, soft tissue label volumes due to excessive background labels.

Several extra changes focusing on consistency have been made across all approaches compared to the MTLRS paper. It is, therefore, difficult to argue whether these had a positive effect on the final performance metrics. The integration of the intermediate segmentations in determining the weighted segmentation loss ensured that each step of the segmentation process contributes to the overall optimisation. Furthermore, the segmentation consistency operator (Section 3.4.1) ensured a consistent intermediate segmentation (Figure 36 in Appendix F), which is beneficial for determining the best multi-task approach. However, this consistent segmentation throughout the network could also harm the overall performance. A strict approach limits the opportunity for the network to find alternative solutions, leading to even better-predicted performances.

Additionally, the TAM and SASG modules were initialised once, and reused over cascades. This approach helps to maintain the module's consistency across the different cascades. However, this could also have negatively affected the propagation of the features. Only one set of trainable parameters has been optimised to propagate all the hidden states. These hidden states can change throughout the cascade architecture, resulting in a mismatch between the determined weights and the cascade-specific hidden states.

5.3 Implication and future research

Performing uncertainty estimation within intermediate predictions generates insight into the intrinsic modelling of the unrolled network. It provides valuable insight into the effect of multi-task approaches, but could possibly also be used for other parametrisation choices, such as assessing the number of iterations needed within an unrolled network.

The inverse problem approach used in many supervised deep learning reconstruction techniques is not without assumptions. Block indicates some assumptions, such as tempered raw data, impossible reproducibility, and minimal performance assessment [49]. Most public datasets lack raw k-space measurements; instead, the k-space data are generated from the target images produced by the MRI scanner. These target images are not necessarily fully sampled because of a restriction in measurement time and a risk of motion artefacts. The shared images are significantly influenced by automated post-processing optimisation strategies, which do not accurately represent the originally acquired k-space lines. The SKM-TEA dataset adopted in this research was acquired with a modest accelerated method (2x1 Parallel Imaging), and the fully-sampled k-space was subsequently synthesised with Autocalibrating Reconstruction for Cartesian imaging (ARC) [15]. Since these interpolated k-space data, and not post-processed imaging data were used for training, the reported performance is expected to generalize well to the prospective imaging setting.

The reproducibility of deep learning models is a key focus for researchers due to the high complexity within the optimisation and validation approaches. Exactly reproducing presented results is challenging. Therefore, this research was performed on top of a repository focused on leveraging reproducibility within the MRI reconstruction community. The open-source repository can be found based on the link provided in Appendix G. This allows future methods to benchmark easily against the proposed approach.

Most papers solely utilise general computer vision metrics, being MSE, PSNR and SSIM to validate their presented results. These do not fully cover the radiologist's reflection of image quality. We therefore included HaarPSI as an additional metric, which was demonstrated to align better with radiologist readings [50].

For clinical application of combined reconstruction and segmentation in the knee, the **SASG** approach would be most beneficial. This approach results in the highest reconstruction performance without negatively affecting the segmentation. Before this approach can be implemented into real-world scenarios, it is advised to set up a clinical trial to compare this approach against the conventional techniques.

For future research, it would be valuable to assess the performance of the MTLRS with the SASG approach on the basis of the initial dataset and compare it with other multi-task unrolled networks. Moreover, an ablation study on loosening the segmentation consistency factors could be performed to indicate the influence of each separate method. Fewer restrictions can help the network find alternative optimisation solutions, improving overall performance. Lastly, another study on extending the multi-task approach by including pathological detection would be valuable to identify the impact of classification tasks on the multi-task approaches. The SKM-TEA dataset also consists of pathological labels and boundary boxes, providing the ideal basis for a multi-task reconstruction, segmentation and detection model.

6 Conclusion

This research utilises several validation methods to provide comprehensive insights into various multi-task approaches within the MTLRS architecture. The findings of this study demonstrate the effectiveness and stability of the **SASG** multi-task approach for the joint reconstruction and segmentation task in MRI. Specifically, the reconstruction metrics and associated significance tests consistently indicate that the **SASG** model significantly outperforms the **JOINT** approach across all selected reconstruction metrics. Additionally, the uncertainty quantification reveals a beneficial impact through the model's cascade structure, further underscoring the **SASG** approach's superiority in enhancing MRI reconstruction quality.

These insights suggest that the **SASG** approach is highly effective and robust across different metrics. It is a valuable candidate for broader implementation in unrolled MRI reconstruction and segmentation networks. Furthermore, the validation methodology utilised in this research offers a robust framework that could be adapted to gain deeper insights into unrolled networks for inverse problems.

This study contributes to a better understanding of the effect of multi-task approaches. It demonstrates the advantages of specific multi-task approaches holding significant potential for advancing the quality of MRI reconstructions, which could have profound implications for medical imaging and diagnostic accuracy. Furthermore, accelerating the diagnostic pathway will minimise the time needed and reduce the cost of imagery in medical healthcare.

7 Bibliography

[1] E. C. McNay et al. "Rising Use Of Diagnostic Medical Imaging In A Large Integrated Health System:" in: *Bone* 23.1 (2008), pp. 1–7. ISSN: 15378276. DOI: 10.1377/hlthaff.27.6.1491.Rising. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf.

[2] Rich Caruana. "Multitask Learning". In: *Machine Learning* 28.1 (1997), pp. 41–75. ISSN: 08856125. DOI: 10.1023/A:1007379606734. URL: http://link.springer.com/10.1023/A:1007379606734.

[3] Kerstin Hammernik and Florian Knoll. "Machine learning for image reconstruction". In: (2020). DOI: 10.1016/B978-0-12-816176-0.00007-7. URL: https://doi.org/10.1016/B978-0-12-816176-0.00007-7.

[4] Florian Knoll et al. "Deep Learning Methods for Parallel Magnetic Resonance Image Reconstruction". In: January (2019), pp. 128–140. URL: http://arxiv.org/abs/1904.01112.

[5] Patrick Putzky et al. "i-RIM applied to the fastMRI challenge". In: (2019). URL: http://arxiv.org/abs/1910.08952.

[6] Anuroop Sriram et al. *End-to-End Variational Networks for Accelerated MRI Reconstruction*. Tech. rep. URL: https://github.com/facebookresearch/fastMRI..

[7] Hemant Kumar Aggarwal, Merry P. Mani, and Mathews Jacob. "MoDL: Model Based Deep Learning Architecture for Inverse Problems". In: (Dec. 2017). DOI: 10.1109/TMI.2018.2865356. URL: http://arxiv.org/abs/1712.02862%20http://dx.doi.org/10.1109/TMI.2018.2865356.

[8] Patrick Putzky and Max Welling. "Recurrent Inference Machines for Solving Inverse Problems". In: Nips (2017). URL: http://arxiv.org/abs/1706.04008.

[9] Jonas Adler et al. "Task adapted reconstruction for inverse problems". In: *Inverse Problems* 38.7 (July 2022), p. 075006. ISSN: 0266-5611. DOI: 10.1088/1361-6420/ac28ec. URL: https://iopscience.iop.org/article/10.1088/1361-6420/ac28ec.

[10] Bin Sui et al. "Simultaneous image reconstruction and lesion segmentation in accelerated MRI using multitasking learning". In: *Medical Physics* 48.11 (Nov. 2021), pp. 7189–7198. ISSN: 24734209. DOI: 10.1002/mp.15213. URL: https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15213.

[11] Liyan Sun et al. "Joint CS-MRI Reconstruction and Segmentation with a Unified Deep Network". In: (May 2018). URL: http://arxiv.org/abs/1805.02165.

[12] Aniket Pramanik and Mathews Jacob. "Joint Calibrationless Reconstruction and Segmentation of Parallel MRI". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13803 LNCS.Xx (2023), pp. 437–453. ISSN: 16113349. DOI: 10.1007/978-3-031-25066-8{_}24.

[13] Yoeri E. Boink, Srirang Manohar, and Christoph Brune. "A Partially-Learned Algorithm for Joint Photo-acoustic Reconstruction and Segmentation". In: *IEEE Transactions on Medical Imaging* 39.1 (2020), pp. 129–139. ISSN: 1558254X. DOI: 10.1109/TMI.2019.2922026.

[14] Dimitrios Karkalousos et al. "MultiTask Learning for accelerated-MRI Reconstruction and Segmentation of Brain Lesions in Multiple Sclerosis". In: *Proceedings of Machine Learning Research* 078 (2023), pp. 1–15.

[15] Arjun Desai et al. "SKM-TEA: A Dataset for Accelerated MRI Reconstruction with Dense Image Labels for Quantitative Clinical Evaluation". In: *Joint Annual Meeting ISMRM-ESMRMB ISMRT 31st Annual Meeting* ML (2022). DOI: 10.58530/2022/0048.

[16] David J Griffiths and Darrell F Schroeter. *Introduction to Quantum Mechanics*. 3rd ed. Cambridge University Press, 2018.

[17] G. A. Wright. "Magnetic resonance imaging". In: *IEEE Signal Processing Magazine* 14.1 (1997), pp. 56–66. ISSN: 10535888. DOI: 10.1109/79.560324.

[18] American Institute of Physics. and American Association of Physicists in Medicine. *Medical physics*. American Association of Physicists in Medicine [by the American Institute of Physics].

[19] Allen D Elster and Jonathan H Burdette. *Questions & answers in magnetic resonance imaging*. English. 2nd ed. St. Louis SE - xiv, 333 p. : illustrations ; 23 cm: Mosby, 2001. ISBN: 0323011845; 9780323011846. DOI: LK-https://worldcat.org/title/488513360.

[20] Yang Song et al. "Solving Inverse Problems in Medical Imaging with Score-Based Generative Models". In: (Nov. 2021). URL: http://arxiv.org/abs/2111.08005.

[21] Jeffrey Fessler. "Model-based image reconstruction for MRI". In: *IEEE Signal Processing Magazine* 27.4 (2010), pp. 81–89. ISSN: 10535888. DOI: 10.1109/MSP.2010.936726.

[22] J. Hennig. *K-space sampling strategies*. July 1999. DOI: 10.1007/s003300050788. URL: http://link.springer.com/10.1007/s003300050788.

[23] Marc Herrmann et al. *Discrete Total Variation with Finite Elements and Applications to Imaging **. Tech. rep. URL: www.tu-chemnitz.de/herzog,.

[24] Michael S. Yao and Michael S. Hansen. *A path towards clinical adaptation of accelerated MRI*. Tech. rep. 2022, pp. 489–511. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10061571/pdf/nihms-1846161.pdf.

[25] Dong Liang et al. "Deep Magnetic Resonance Image Reconstruction: Inverse Problems Meet Neural Networks." In: *IEEE signal processing magazine* 37.1 (Jan. 2020), pp. 141–151. ISSN: 1053-5888. DOI: 10.1109/MSP.2019.2950557. URL: http://www.ncbi.nlm.nih.gov/pubmed/33746470%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7977031.

[26] Gushan Zeng et al. "A review on deep learning MRI reconstruction without fully sampled k-space." In: *BMC medical imaging* 21.1 (Dec. 2021), p. 195. ISSN: 1471-2342. DOI: 10.1186/s12880-021-00727-9. URL: http://www.ncbi.nlm.nih.gov/pubmed/34952572%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8710001.

[27] D. Karkalousos et al. "Assessment of data consistency through cascades of independently recurrent inference machines for fast and robust accelerated MRI reconstruction". In: *Physics in Medicine and Biology* 67.12 (June 2022), p. 124001. ISSN: 13616560. DOI: 10.1088/1361-6560/ac6cc2. URL: https://iopscience.iop.org/article/10.1088/1361-6560/ac6cc2.

[28] Masoumeh Dashti and Andrew M Stuart. "The bayesian approach to inverse problems". In: *Handbook of Uncertainty Quantification*. 2015, pp. 311–428. ISBN: 9783319123851. DOI: 10.1007/978-3-319-12385-1{_}7. URL: https://arxiv.org/pdf/1302.6989.

[29] Kerstin Hammernik et al. "Learning a variational network for reconstruction of accelerated MRI data". In: *Magnetic Resonance in Medicine* 79.6 (2017), pp. 3055–3071. ISSN: 15222594. DOI: 10.1002/mrm.26977. URL: https://arxiv.org/pdf/1704.00447.

[30] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training Recurrent Neural Networks". In: (Nov. 2012). URL: http://arxiv.org/abs/1211.5063.

[31] Kyunghyun Cho et al. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: (June 2014). URL: http://arxiv.org/abs/1406.1078.

[32] Shuai Li et al. "Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN". In: (Mar. 2018). URL: http://arxiv.org/abs/1803.04831.

[33] Hongxia Wang et al. "Adaptive Hard Parameter Sharing Method Based on Multi-Task Deep Learning". In: *Mathematics* 11.22 (Nov. 2023), p. 4639. ISSN: 22277390. DOI: 10.3390/math11224639. URL: https://www.mdpi.com/2227-7390/11/22/4639.

[34] Michael Crawshaw. "Multi-Task Learning with Deep Neural Networks: A Survey". In: (Sept. 2020). URL: http://arxiv.org/abs/2009.09796.

[35] Sebastian Ruder. "An Overview of Multi-Task Learning in Deep Neural Networks". In: (June 2017). URL: http://arxiv.org/abs/1706.05098.

[36] Jifeng Dai, Kaiming He, and Jian Sun. "Instance-Aware Semantic Segmentation via Multi-task Network Cascades". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2016-Decem. 2016, pp. 3150–3158. ISBN: 9781467388504. DOI: 10.1109/CVPR.2016.343. URL: https://arxiv.org/pdf/1512.04412.pdf.

[37] Zhenyu Zhang et al. *Joint Task-Recursive Learning for Semantic Segmentation and Depth Estimation.* Tech. rep. URL: https://openaccess.thecvf.com/content_ECCV_2018/papers/Zhenyu_Zhang_Joint_Task-Recursive_Learning_ECCV_2018_paper.pdf.

[38] Ozan Oktay et al. *Attention U-Net: Learning Where to Look for the Pancreas.* Tech. rep. URL: https://arxiv.org/pdf/1804.03999.

[39] Vineet Edupuganti et al. "Uncertainty Quantification in Deep MRI Reconstruction". In: (Jan. 2019). URL: http://arxiv.org/abs/1901.11228.

[40] Joseph Paul Cohen, Margaux Luck, and Sina Honari. "Distribution Matching Losses Can Hallucinate Features in Medical Image Translation". In: (May 2018). URL: http://arxiv.org/abs/1805.08841.

[41] Vegard Antun et al. "On instabilities of deep learning in image reconstruction and the potential costs of AI". In: *Proceedings of the National Academy of Sciences* 117.48 (Dec. 2020), pp. 30088–30095. ISSN: 0027-8424. DOI: 10.1073/pnas.1907377117. URL: https://pnas.org/doi/full/10.1073/pnas.1907377117.

[42] Riccardo Barbano et al. *Uncertainty quantification in medical image synthesis.* Tech. rep. 2022, pp. 601–641. DOI: 10.1016/B978-0-12-824349-7.00033-5. URL: https://discovery.ucl.ac.uk/id/eprint/10129844/1/BookChapter_Modelling_Uncertainty_General_Latex.pdf.

[43] Moloud Abdar et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges". In: *Information Fusion* 76 (Dec. 2021), pp. 243–297. ISSN: 15662535. DOI: 10.1016/j.inffus.2021.05.008. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001081.

[44] B. Sveinsson et al. "A simple analytic method for estimating T2 in the knee from DESS". In: *Magnetic Resonance Imaging* 38 (2017), pp. 63–70. ISSN: 18735894. DOI: 10.1016/j.mri.2016.12.018. URL: http://dx.doi.org/10.1016/j.mri.2016.12.018.

[45] Arjun D Desai et al. "Noise2Recon: Enabling Joint MRI Reconstruction and Denoising with Semi-Supervised and Self-Supervised Learning". In: (Sept. 2021). URL: http://arxiv.org/abs/2110.00075.

[46] Xiaodi Li and Yue Hu. "Cooperative-Net: An end-to-end multi-task interaction network for unified reconstruction and segmentation of MR image". In: (2024). DOI: 10.1016/j.cmpb.2024.108045. URL: https://doi.org/10.1016/j.cmpb.2024.108045.

[47] Chuan Guo et al. *On Calibration of Modern Neural Networks.* Tech. rep. 2017. URL: https://arxiv.org/pdf/1706.04599.pdf.

[48] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in Neural Information Processing Systems.* Vol. 2017-Decem. 2017, pp. 6403–6414. URL: https://arxiv.org/pdf/1612.01474.pdf.

[49] Kai Tobias Block. "Subtle pitfalls in the search for faster medical imaging". In: *Proceedings of the National Academy of Sciences* 119.17 (Apr. 2022), e2203040119. ISSN: 0027-8424. DOI: 10.1073/pnas.2203040119. URL: https://pnas.org/doi/full/10.1073/pnas.2203040119.

[50] Sergey Kastryulin et al. "Image Quality Assessment for Magnetic Resonance Imaging". In: *IEEE Access* 11 (2023), pp. 14154–14168. ISSN: 21693536. DOI: 10.1109/ACCESS.2023.3243466.

A Reconstruction and Segmentation results

A.1 Reconstitution results

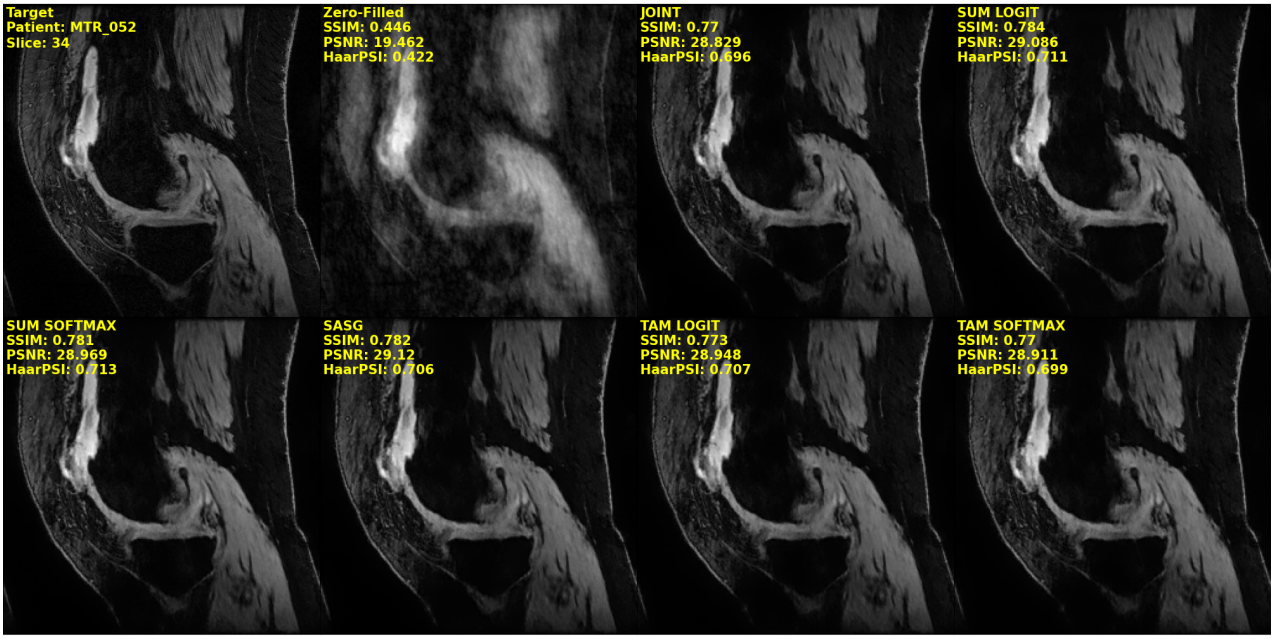


Figure 20: Visualisation of the first echo scan of a specific patient and slice, starting from top-left showing the target scan, the zero-filled reconstruction based on the 8× Gaussian undersampling and then the various reconstructions of the different multi-task approaches. Furthermore, each approach's top left corner states the reconstruction's performance metrics.

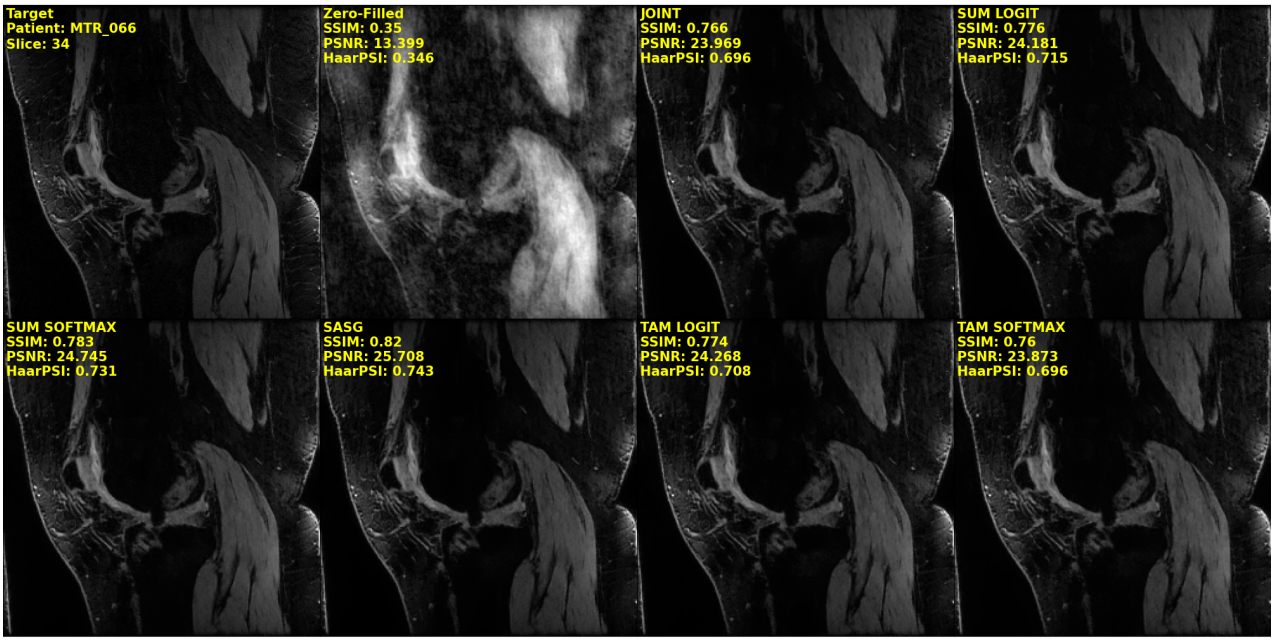


Figure 21: Visualisation of the first echo scan of a specific patient and slice, starting from top-left showing the target scan, the zero-filled reconstruction based on the 8× Gaussian undersampling and then the various reconstructions of the different multi-task approaches. Furthermore, each approach's top left corner states the reconstruction's performance metrics.

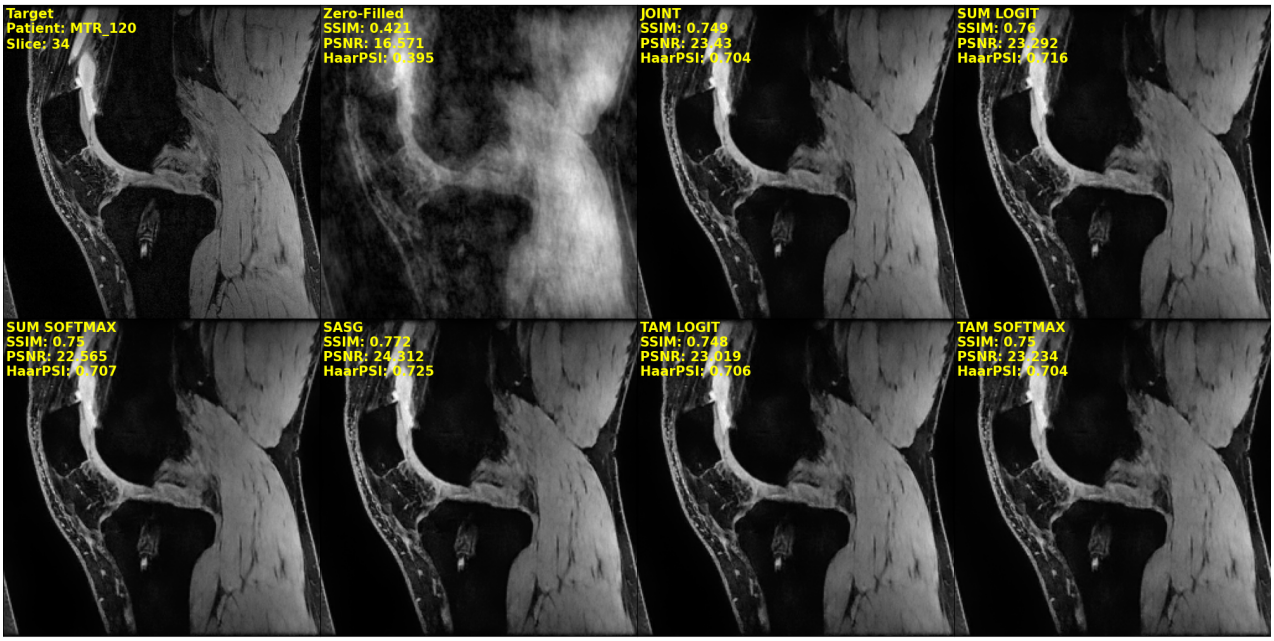


Figure 22: Visualisation of the first echo scan of a specific patient and slice, starting from top-left showing the target scan, the zero-filled reconstruction based on the 8× Gaussian undersampling and then the various reconstructions of the different multi-task approaches. Furthermore, each approach's top left corner states the reconstruction's performance metrics.

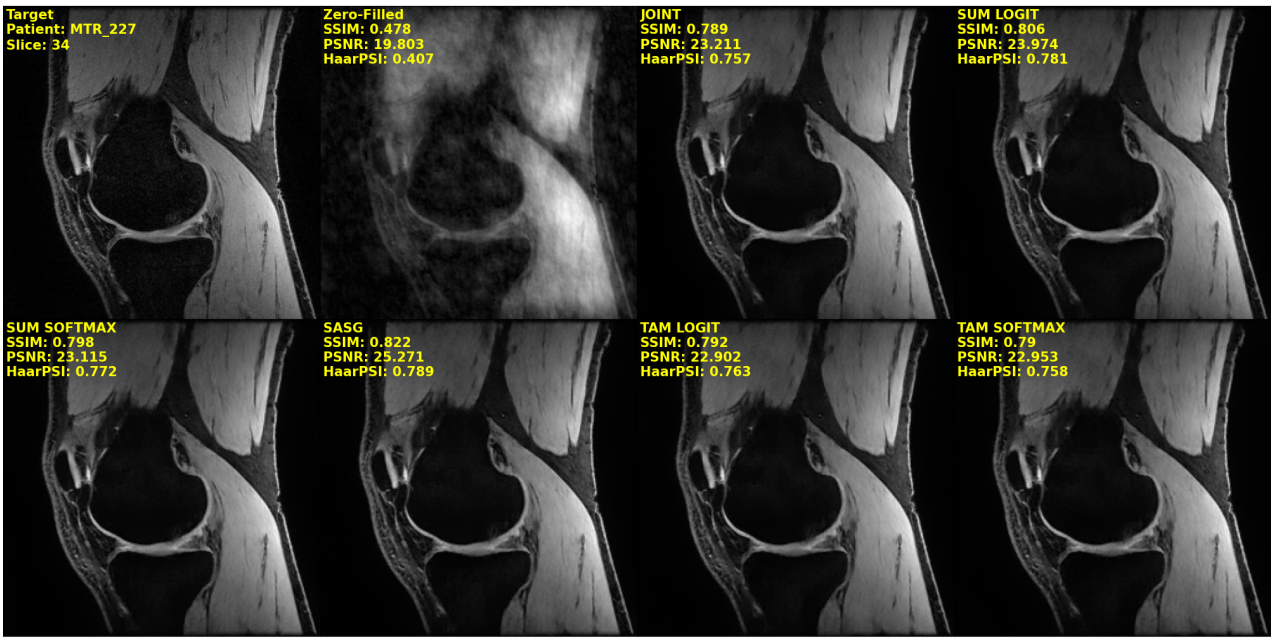


Figure 23: Visualisation of the first echo scan of a specific patient and slice, starting from top-left showing the target scan, the zero-filled reconstruction based on the 8× Gaussian undersampling and then the various reconstructions of the different multi-task approaches. Furthermore, each approach's top left corner states the reconstruction's performance metrics.

A.2 Segmentation results

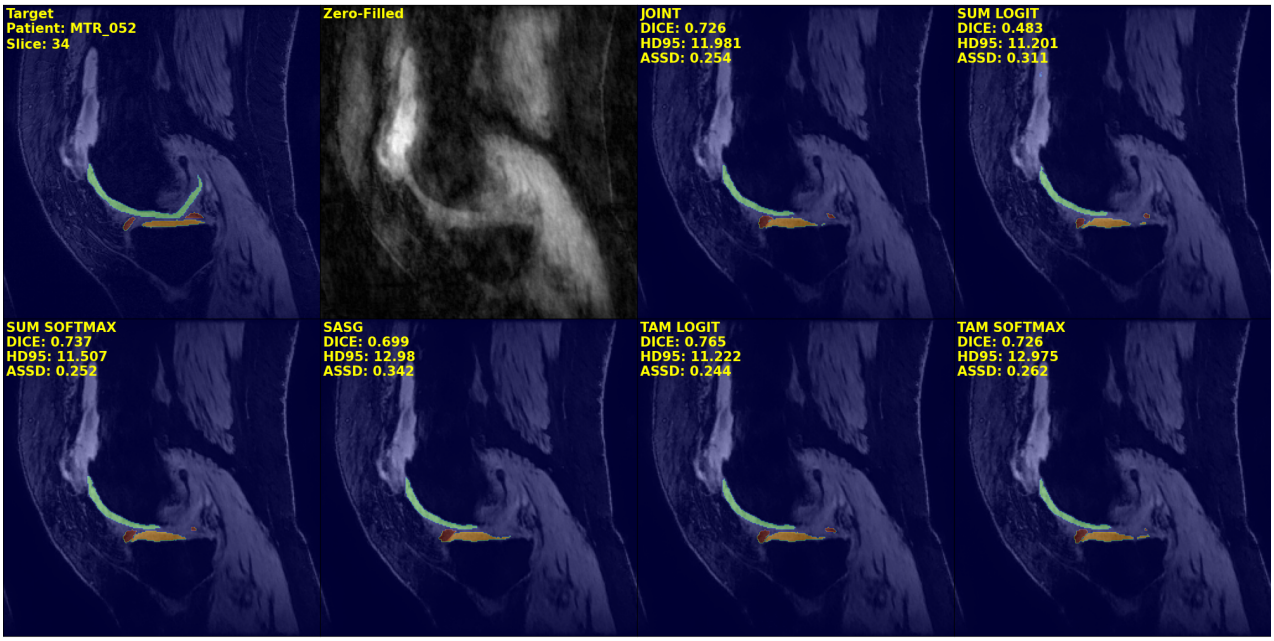


Figure 24: Visualisation of the first echo scan and the segmentation labels as an overlay, the target and the six multi-task approaches with three metrics to identify the segmentation performance. The light blue label indicates the patellar cartilage, the green the femoral cartilage, the orange the tibial cartilage, red the meniscus, and dark blue the background class. The segmentation performance metrics are stated in the top left corner of each individual image.

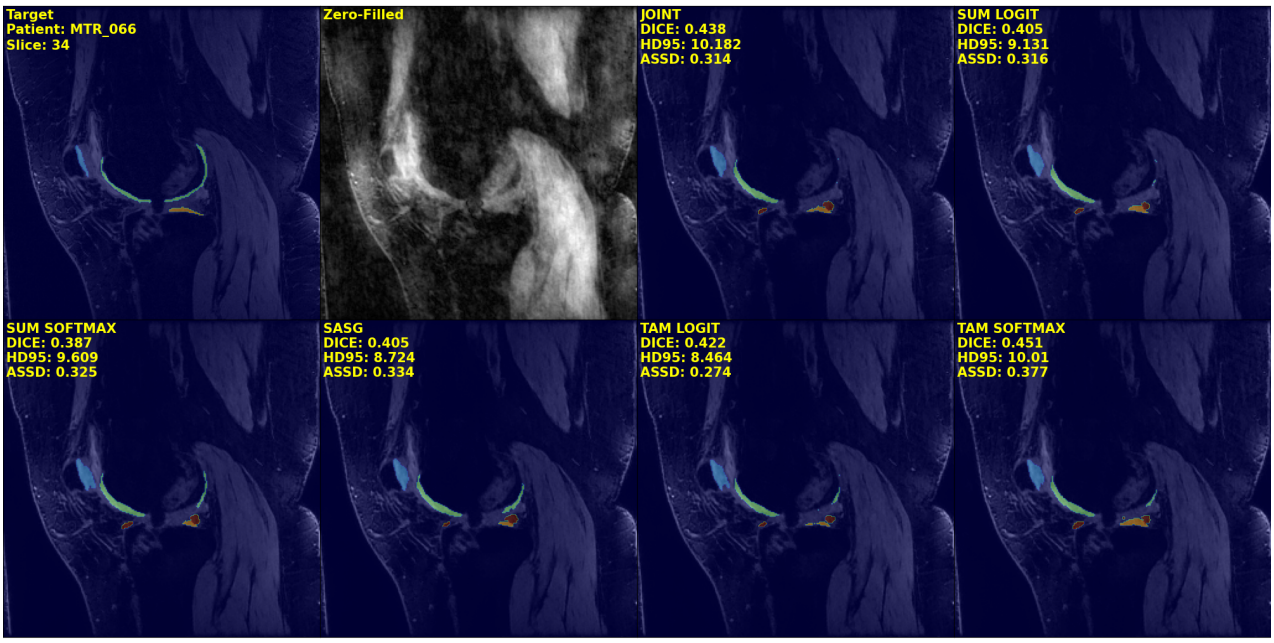


Figure 25: Visualisation of the first echo scan and the segmentation labels as an overlay, the target and the six multi-task approaches with three metrics to identify the segmentation performance. The light blue label indicates the patellar cartilage, the green the femoral cartilage, the orange the tibial cartilage, red the meniscus, and dark blue the background class. The segmentation performance metrics are stated in the top left corner of each individual image.

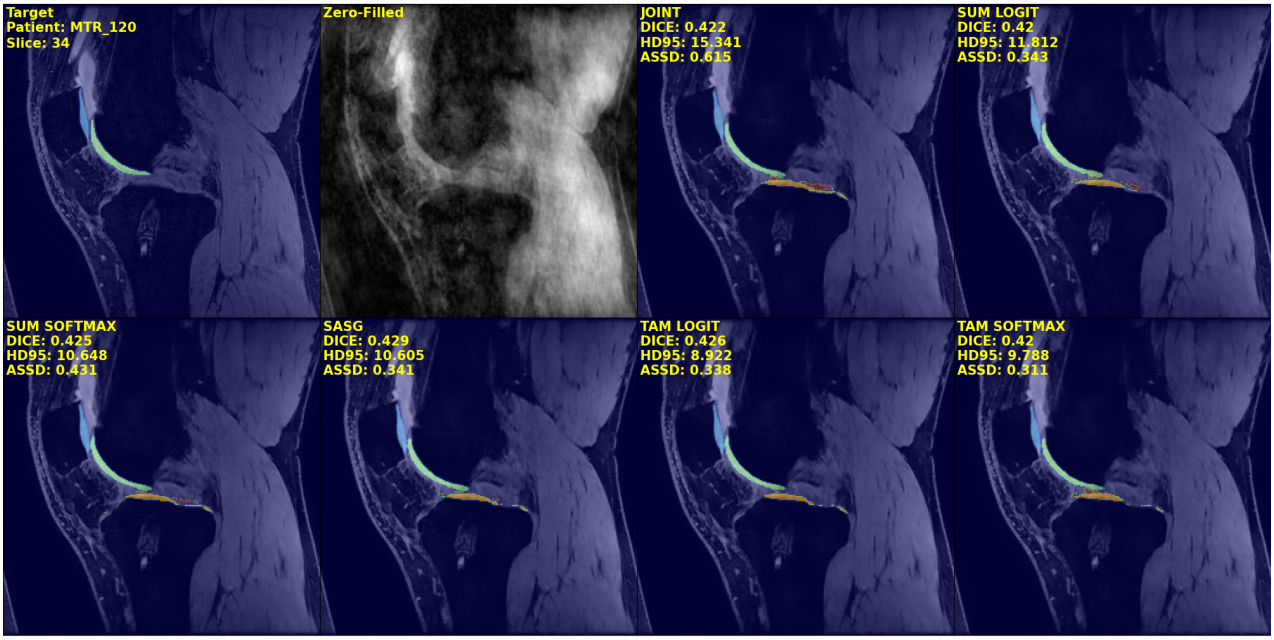


Figure 26: Visualisation of the first echo scan and the segmentation labels as an overlay, the target and the six multi-task approaches with three metrics to identify the segmentation performance. The light blue label indicates the patellar cartilage, the green the femoral cartilage, the orange the tibial cartilage, red the meniscus, and dark blue the background class. The segmentation performance metrics are stated in the top left corner of each individual image.

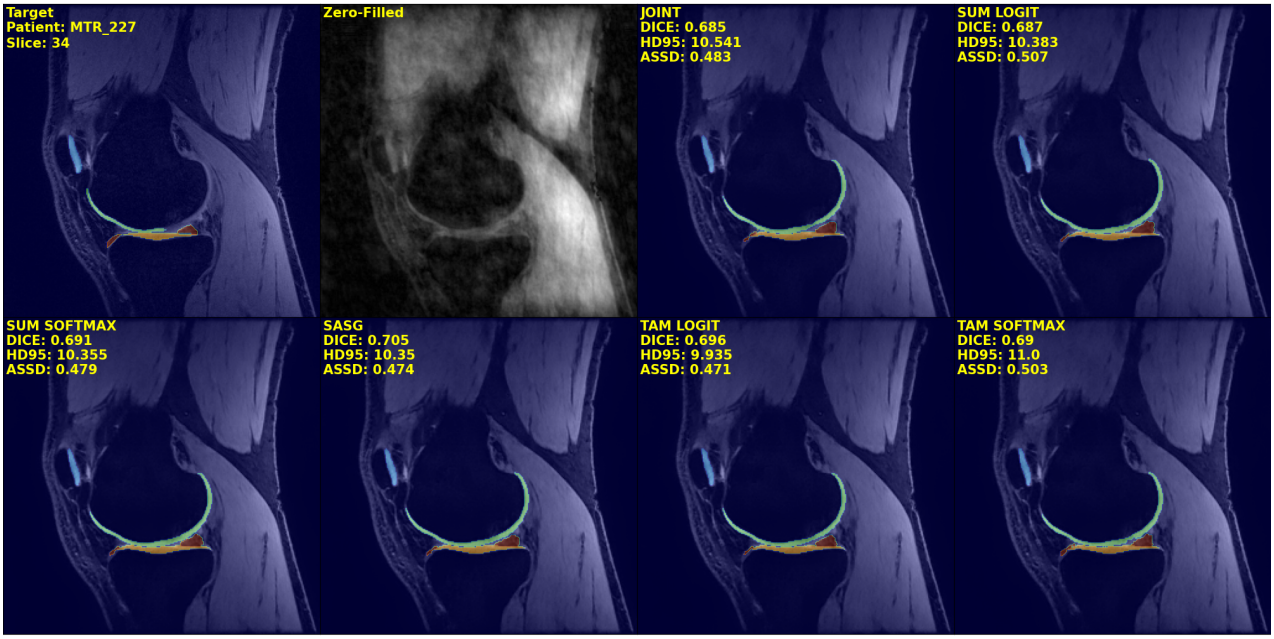


Figure 27: Visualisation of the first echo scan and the segmentation labels as an overlay, the target and the six multi-task approaches with three metrics to identify the segmentation performance. The light blue label indicates the patellar cartilage, the green the femoral cartilage, the orange the tibial cartilage, red the meniscus, and dark blue the background class. The segmentation performance metrics are stated in the top left corner of each individual image.

B Significant Tests

The results below indicate the significance of each multi-task approach based on various metrics.

B.1 Reconstruction Metrics

	df	**sum_sq**	**mean_sq**	**F**	**PR(>F)**
multi-task approach	5	0.252	0.050	180.500	0
patient	35	1.099	0.031	112.542	0
echo	1	0.018	0.018	63.779	0
multi-task approach:patient	175	0.053	0.000	1.089	0.213
multi-task approach:echo	5	0.141	0.028	101.275	0
patient:echo	35	0.581	0.017	59.538	0
multi-task approach:patient:echo	1175.000	0.031	0	0.643	1
Residual	1620	0.452	0		

Table 7: Results of the ANOVA test performed on HaarPSI reflect the variance source within this metric. Here, **df** is the degrees of freedom, reflecting the source of variation, **sum_sq** is the sum of squares for the specific source of variation, **mean_sq** is the mean of squares for the specific source of variation and **F** The F-value is a statistic that compares the variance between group means to the variance within groups. The **PR(>F)** indicates the probability of obtaining an F-statistic at least as extreme as the one calculated, assuming the null hypothesis is true. Values below 0.05 indicate that the observed differences among group means are unlikely to have occurred by chance. Therefore, the null hypothesis can be rejected and concluded that the factor has a significant effect on the dependent variable. **Note:** values stated as 0 are considered below 0.001.

multi-task approach 1	**multi-task approach 2**	**mean_diff**	**p-adj**	**lower**	**upper**	**reject**
JOINT	SUM LOGIT	0.022	0.000	0.014	0.029	TRUE
JOINT	SUM SOFTMAX	0.010	0.001	0.003	0.018	TRUE
JOINT	SASG	0.020	0.000	0.013	0.027	TRUE
JOINT	TAM LOGIT	0.000	1.000	-0.007	0.007	FALSE
JOINT	TAM SOFTMAX	-0.010	0.003	-0.018	-0.002	TRUE

Table 8: Results of the Tukey Honest Significant Difference (HSD) test based on the HaarPSI metric. This test is used to perform pairwise comparisons to determine if the means of different groups are significantly different. Where **multi-task approach 1 / multi-task approach 2** indicates which groups are being compared, **mean_diff** indicates the difference between the means of the two groups, **p-adj** is the adjusted p-value for multiple comparisons, **lower** and **upper** are the bounds of the 95% confidence interval for the mean difference, and **reject** indicates if the null hypothesis of no difference between the group means is rejected (TRUE) or not (FALSE). **Note:** values stated as 0 are considered below 0.001 .

	df	sum_sq	mean_sq	F	PR(>F)
multi-task approach	5	391.378	78.276	237.795	0
patient	35	4701.662	134.333	408.094	0
echo	1	5811.296	5811.296	17654.287	0
multi-task approach:patient	175	115.008	0.657	1.996	0
multi-task approach:echo	5	45.555	9.111	27.678	0
patient:echo	35	1213.715	34.678	105.348	0
multi-task approach:patient:echo	1175.000	58.642	0.335	1.018	0.425
Residual	1620	533.259	0.329		

Table 9: Results of the ANOVA test performed on PSNR reflect the variance source within this metric. Here, **df** is the degrees of freedom, reflecting the source of variation, **sum_sq** is the sum of squares for the specific source of variation, **mean_sq** is the mean of squares for the specific source of variation and **F** The F-value is a statistic that compares the variance between group means to the variance within groups. The **PR(>F)** indicates the probability of obtaining an F-statistic at least as extreme as the one calculated, assuming the null hypothesis is true. Values below 0.05 indicate that the observed differences among group means are unlikely to have occurred by chance. Therefore, the null hypothesis can be rejected and concluded that the factor has a significant effect on the dependent variable. **Note:** values stated as 0 are considered below 0.001.

multi-task approach 1	multi-task approach 2	mean_diff	p-adj	lower	upper	reject
JOINT	SUM LOGIT	0.329	0.504	-0.210	0.869	FALSE
JOINT	SUM SOFTMAX	-0.097	0.995	-0.622	0.428	FALSE
JOINT	SASG	0.731	0.001	0.206	1.256	TRUE
JOINT	TAM LOGIT	-0.253	0.742	-0.77	8 0.272	FALSE
JOINT	TAM SOFTMAX	-0.693	0.005	-1.250	-0.136	TRUE

Table 10: Results of the Tukey Honest Significant Difference (HSD) test based on the PSNR metric. This test is used to perform pairwise comparisons to determine if the means of different groups are significantly different. Where **multi-task approach 1 / multi-task approach 2** indicates which groups are being compared, **mean_diff** indicates the difference between the means of the two groups, **p-adj** is the adjusted p-value for multiple comparisons, **lower** and **upper** are the bounds of the 95% confidence interval for the mean difference, and **reject** indicates if the null hypothesis of no difference between the group means is rejected (TRUE) or not (FALSE). **Note:** values stated as 0 are considered below 0.001 .

.

B.2 Segmentation Metrics

	df	sum_sq	mean_sq	F	PR(>F
multi-task approach	5	0.058	0.012	21.835	0
patient	35	3.591	0.103	192.994	0
multi-task approach:patient	175	0.013	0.000	0.144	1
Residual	864	0.459	0.001		

Table 11: Results of the ANOVA test performed on DICE reflect the variance source within this metric. Here, **df** is the degrees of freedom, reflecting the source of variation, **sum_sq** is the sum of squares for the specific source of variation, **mean_sq** is the mean of squares for the specific source of variation and **F** The F-value is a statistic that compares the variance between group means to the variance within groups. The **PR(>F)** indicates the probability of obtaining an F-statistic at least as extreme as the one calculated, assuming the null hypothesis is true. Values below 0.05 indicate that the observed differences among group means are unlikely to have occurred by chance. Therefore, the null hypothesis can be rejected and concluded that the factor has a significant effect on the dependent variable. **Note:** values stated as 0 are considered below 0.001.

multi-task approach 1	multi-task approach 2	mean_diff	p-adj	lower	upper	reject
JOINT	SUM LOGIT	-0.003	0.995	-0.022	0.015	False
JOINT	SUM SOFTMAX	0.006	0.940	-0.013	0.025	False
JOINT	SASG	-0.005	0.971	-0.024	0.014	False
JOINT	TAM LOGIT	0.010	0.640	-0.009	0.029	False
JOINT	TAM SOFTMAX	-0.012	0.395	-0.031	0.006	False

Table 12: Results of the Tukey Honest Significant Difference (HSD) test based on the DICE metric. This test is used to perform pairwise comparisons to determine if the means of different groups are significantly different. Where **multi-task approach 1 / multi-task approach 2** indicates which groups are being compared, **mean_diff** indicates the difference between the means of the two groups, **p-adj** is the adjusted p-value for multiple comparisons, **lower** and **upper** are the bounds of the 95% confidence interval for the mean difference, and **reject** indicates if the null hypothesis of no difference between the group means is rejected (TRUE) or not (FALSE). **Note:** values stated as 0 are considered below 0.001 .

	df	sum_sq	mean_sq	F	PR(>F
multi-task approach	5	23.917	4.783	18.752	0
patient	35	472.797	13.508	52.956	0
multi-task approach:patient	175	8.413	0.048	0.188	1
Residual	864	220.397	0.255		

Table 13: Results of the ANOVA test performed on HD95 reflect the variance source within this metric. Here, **df** is the degrees of freedom, reflecting the source of variation, **sum_sq** is the sum of squares for the specific source of variation, **mean_sq** is the mean of squares for the specific source of variation and **F** The F-value is a statistic that compares the variance between group means to the variance within groups. The **PR(>F)** indicates the probability of obtaining an F-statistic at least as extreme as the one calculated, assuming the null hypothesis is true. Values below 0.05 indicate that the observed differences among group means are unlikely to have occurred by chance. Therefore, the null hypothesis can be rejected and concluded that the factor has a significant effect on the dependent variable. **Note:** values stated as 0 are considered below 0.001.

multi-task approach 1	multi-task approach 2	mean_diff	p-adj	lower	upper	reject
JOINT	SUM LOGIT	0.114	0.766	-0.130	0.357	False
JOINT	SUM SOFTMAX	-0.0142	1.000	-0.257	0.229	False
JOINT	SASG	0.218	0.108	-0.025	0.462	False
JOINT	TAM LOGIT	-0.086	0.917	-0.329	0.158	False
JOINT	TAM SOFTMAX	0.348	0.001	0.104	0.591	True

Table 14: Results of the Tukey Honest Significant Difference (HSD) test based on the HD95 metric. This test is used to perform pairwise comparisons to determine if the means of different groups are significantly different. Where **multi-task approach 1 / multi-task approach 2** indicates which groups are being compared, **mean_diff** indicates the difference between the means of the two groups, **p-adj** is the adjusted p-value for multiple comparisons, **lower** and **upper** are the bounds of the 95% confidence interval for the mean difference, and **reject** indicates if the null hypothesis of no difference between the group means is rejected (TRUE) or not (FALSE). **Note:** values stated as 0 are considered below 0.001 .

	df	sum_sq	mean_sq	F	PR(>F
multi-task approach	5	1.024	0.205	20.702	0
patient	35	7.801	0.223	22.523	0
multi-task approach:patient	175	0.420	0.002	0.243	1
Residual	864	8.550	0.010		

Table 15: Results of the ANOVA test performed on ASSD reflect the variance source within this metric. Here, **df** is the degrees of freedom, reflecting the source of variation, **sum_sq** is the sum of squares for the specific source of variation, **mean_sq** is the mean of squares for the specific source of variation and **F** The F-value is a statistic that compares the variance between group means to the variance within groups. The **PR(>F)** indicates the probability of obtaining an F-statistic at least as extreme as the one calculated, assuming the null hypothesis is true. Values below 0.05 indicate that the observed differences among group means are unlikely to have occurred by chance. Therefore, the null hypothesis can be rejected and concluded that the factor has a significant effect on the dependent variable. **Note:** values stated as 0 are considered below 0.001.

multi-task approach 1	**multi-task approach 2**	**mean_diff**	**p-adj**	**lower**	**upper**	**reject**
JOINT	SUM LOGIT	0.037	0.063	-0.001	0.074	False
JOINT	SUM SOFTMAX	-0.007	0.996	-0.044	0.031	False
JOINT	SASG	0.055	0.001	0.017	0.093	True
JOINT	TAM LOGIT	-0.007	0.994	-0.045	0.030	False
JOINT	TAM SOFTMAX	0.070	0.000	0.032	0.107	True

Table 16: Results of the Tukey Honest Significant Difference (HSD) test based on the ASSD metric. This test is used to perform pairwise comparisons to determine if the means of different groups are significantly different. Where **multi-task approach 1 / multi-task approach 2** indicates which groups are being compared, **mean_diff** indicates the difference between the means of the two groups, **p-adj** is the adjusted p-value for multiple comparisons, **lower** and **upper** are the bounds of the 95% confidence interval for the mean difference, and **reject** indicates if the null hypothesis of no difference between the group means is rejected (TRUE) or not (FALSE). **Note:** values stated as 0 are considered below 0.001 .

.

B.3 T_2 Estimation Metric

	df	sum_sq	mean_sq	F	PR(>F
multi-task approach	5	13.631	2.726	3.797	0.002
patient	35	1138.435	32.527	45.303	0
multi-task approach:patient	175	168.484	0.963	1.341	0.005
Residual	864	620.338	0.718		

Table 17: Results of the ANOVA test performed on meniscus T_2 error reflect the variance source within this metric. Here, **df** is the degrees of freedom, reflecting the source of variation, **sum_sq** is the sum of squares for the specific source of variation, **mean_sq** is the mean of squares for the specific source of variation and **F** The F-value is a statistic that compares the variance between group means to the variance within groups. The **PR(>F)** indicates the probability of obtaining an F-statistic at least as extreme as the one calculated, assuming the null hypothesis is true. Values below 0.05 indicate that the observed differences among group means are unlikely to have occurred by chance. Therefore, the null hypothesis can be rejected and concluded that the factor has a significant effect on the dependent variable. **Note:** values stated as 0 are considered below 0.001.

multi-task approach 1	multi-task approach 2	mean_diff	p-adj	lower	upper	reject
JOINT	SUM LOGIT	-0.130	0.942	-0.533	0.273	False
JOINT	SUM SOFTMAX	0.011	1.000	-0.392	0.414	False
JOINT	SASG	-0.119	0.960	-0.522	0.284	False
JOINT	TAM LOGIT	-0.123	0.953	-0.526	0.280	False
JOINT	TAM SOFTMAX	0.186	0.775	-0.217	0.589	False

Table 18: Results of the Tukey Honest Significant Difference (HSD) test based on the meniscus T_2 error metric. This test is used to perform pairwise comparisons to determine if the means of different groups are significantly different. Where **multi-task approach 1 / multi-task approach 2** indicates which groups are being compared, **mean_diff** indicates the difference between the means of the two groups, **p-adj** is the adjusted p-value for multiple comparisons, **lower** and **upper** are the bounds of the 95% confidence interval for the mean difference, and **reject** indicates if the null hypothesis of no difference between the group means is rejected (TRUE) or not (FALSE). **Note:** values stated as 0 are considered below 0.001 .

	df	sum_sq	mean_sq	F	PR(>F
multi-task approach	5	37.206	7.441	4.840	0
patient	35	1882.276	53.779	34.978	0
multi-task approach:patient	175	271.950	1.554	1.011	0.53
Residual	864	1328.415	1.538		

Table 19: Results of the ANOVA test performed on patellar T_2 error reflect the variance source within this metric. Here, **df** is the degrees of freedom, reflecting the source of variation, **sum_sq** is the sum of squares for the specific source of variation, **mean_sq** is the mean of squares for the specific source of variation and **F** The F-value is a statistic that compares the variance between group means to the variance within groups. The **PR(>F)** indicates the probability of obtaining an F-statistic at least as extreme as the one calculated, assuming the null hypothesis is true. Values below 0.05 indicate that the observed differences among group means are unlikely to have occurred by chance. Therefore, the null hypothesis can be rejected and concluded that the factor has a significant effect on the dependent variable. **Note:** values stated as 0 are considered below 0.001.

multi-task approach 1	multi-task approach 2	mean_diff	p-adj	lower	upper	reject
JOINT	SUM LOGIT	-0.470	0.132	-1.012	0.072	False
JOINT	SUM SOFTMAX	-0.214	0.869	-0.756	0.328	False
JOINT	SASG	0.122	0.988	-0.420	0.664	False
JOINT	TAM LOGIT	-0.196	0.908	-0.737	0.346	False
JOINT	TAM SOFTMAX	-0.197	0.905	-0.739	0.345	False

Table 20: Results of the Tukey Honest Significant Difference (HSD) test based on the patellar T_2 error metric. This test is used to perform pairwise comparisons to determine if the means of different groups are significantly different. Where **multi-task approach 1 / multi-task approach 2** indicates which groups are being compared, **mean_diff** indicates the difference between the means of the two groups, **p-adj** is the adjusted p-value for multiple comparisons, **lower** and **upper** are the bounds of the 95% confidence interval for the mean difference, and **reject** indicates if the null hypothesis of no difference between the group means is rejected (TRUE) or not (FALSE). **Note:** values stated as 0 are considered below 0.001 .

	df	sum_sq	mean_sq	F	PR(>F
multi-task approach	5	476.808	95.362	79.806	0
patient	35	2035.742	58.164	48.676	0
multi-task approach:patient	175	260.425	1.488	1.245	0.26
Residual	864	1032.410	1.195		

Table 21: Results of the ANOVA test performed on tibial T_2 error reflect the variance source within this metric. Here, **df** is the degrees of freedom, reflecting the source of variation, **sum_sq** is the sum of squares for the specific source of variation, **mean_sq** is the mean of squares for the specific source of variation and **F** The F-value is a statistic that compares the variance between group means to the variance within groups. The **PR(>F)** indicates the probability of obtaining an F-statistic at least as extreme as the one calculated, assuming the null hypothesis is true. Values below 0.05 indicate that the observed differences among group means are unlikely to have occurred by chance. Therefore, the null hypothesis can be rejected and concluded that the factor has a significant effect on the dependent variable. **Note:** values stated as 0 are considered below 0.001.

multi-task approach 1	**multi-task approach 2**	**mean_diff**	**p-adj**	**lower**	**upper**	**reject**
JOINT	SUM LOGIT	-1.405	0.000	-1.935	-0.875	True
JOINT	SUM SOFTMAX	-0.942	0.000	-1.472	-0.412	True
JOINT	SASG	-1.270	0.000	-1.800	-0.741	True
JOINT	TAM LOGIT	0.308	0.559	-0.222	0.838	False
JOINT	TAM SOFTMAX	-0.094	0.996	-0.624	0.435	False

Table 22: Results of the Tukey Honest Significant Difference (HSD) test based on the tibial T_2 error metric. This test is used to perform pairwise comparisons to determine if the means of different groups are significantly different. Where **multi-task approach 1 / multi-task approach 2** indicates which groups are being compared, **mean_diff** indicates the difference between the means of the two groups, **p-adj** is the adjusted p-value for multiple comparisons, **lower** and **upper** are the bounds of the 95% confidence interval for the mean difference, and **reject** indicates if the null hypothesis of no difference between the group means is rejected (TRUE) or not (FALSE). **Note:** values stated as 0 are considered below 0.001 .

	df	sum_sq	mean_sq	F	PR(>F
multi-task approach	5	324.276	64.855	63.104	0
patient	35	1779.032	50.829	49.457	0
multi-task approach:patient	175	188.396	1.077	1.047	0.336
Residual	864	887.983	1.028		

Table 23: Results of the ANOVA test performed on femoral T_2 error reflect the variance source within this metric. Here, **df** is the degrees of freedom, reflecting the source of variation, **sum_sq** is the sum of squares for the specific source of variation, **mean_sq** is the mean of squares for the specific source of variation and **F** The F-value is a statistic that compares the variance between group means to the variance within groups. The **PR(>F)** indicates the probability of obtaining an F-statistic at least as extreme as the one calculated, assuming the null hypothesis is true. Values below 0.05 indicate that the observed differences among group means are unlikely to have occurred by chance. Therefore, the null hypothesis can be rejected and concluded that the factor has a significant effect on the dependent variable. **Note:** values stated as 0 are considered below 0.001.

multi-task approach 1	multi-task approach 2	mean_diff	p-adj	lower	upper	reject
JOINT	SUM LOGIT	-0.755	0.000	-1.246	-0.265	True
JOINT	SUM SOFTMAX	-0.259	0.659	-0.750	0.232	False
JOINT	SASG	0.036	1.000	-0.455	0.527	False
JOINT	TAM LOGIT	0.717	0.001	0.226	1.208	True
JOINT	TAM SOFTMAX	0.839	0.000	0.348	1.330	True

Table 24: Results of the Tukey Honest Significant Difference (HSD) test based on the femoral T_2 error metric. This test is used to perform pairwise comparisons to determine if the means of different groups are significantly different. Where **multi-task approach 1 / multi-task approach 2** indicates which groups are being compared, **mean_diff** indicates the difference between the means of the two groups, **p-adj** is the adjusted p-value for multiple comparisons, **lower** and **upper** are the bounds of the 95% confidence interval for the mean difference, and **reject** indicates if the null hypothesis of no difference between the group means is rejected (TRUE) or not (FALSE). **Note:** values stated as 0 are considered below 0.001 .

C Intermediate reconstruction

This appendix includes all the figures relating to the metric evaluation of the intermediate predictions and uncertainty.

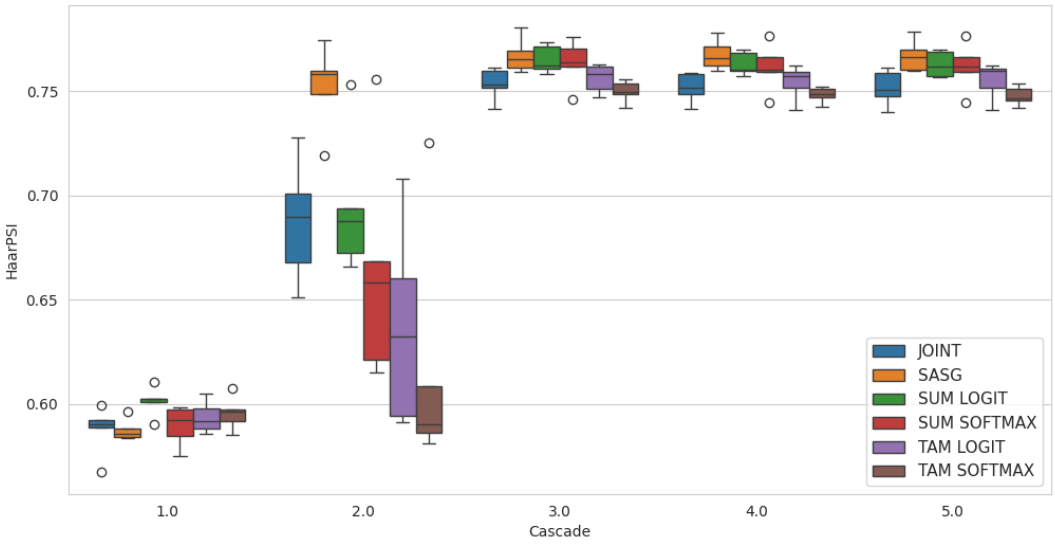


Figure 28: HaarPSI metric score of the intermediate reconstructions based on the first echo of five patients. Each multi-task approach was initialised five times, reflecting the measurement of predictive uncertainty.

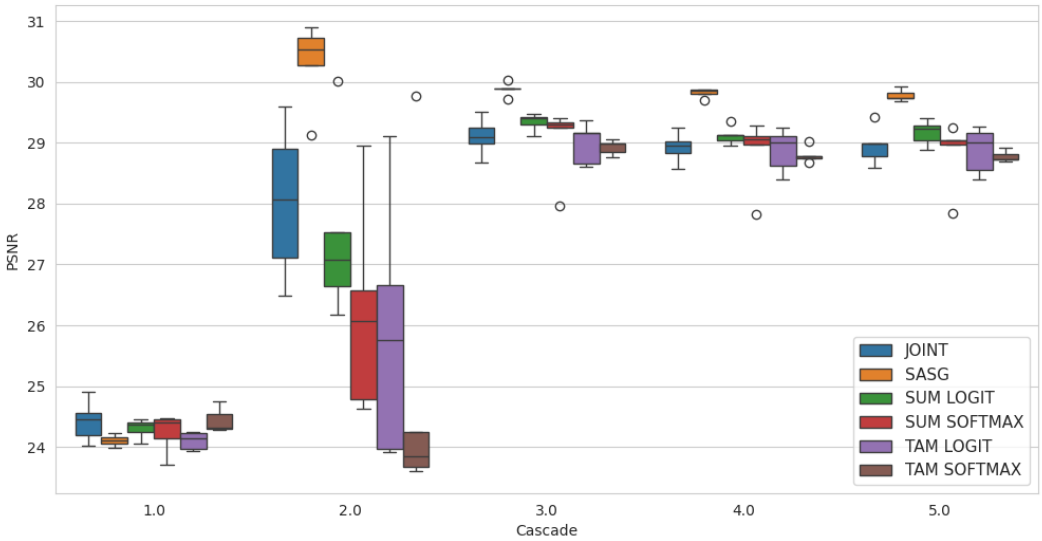


Figure 29: PSNR metric score of the intermediate reconstructions based on the first echo of five patients. Each multi-task approach was initialised five times, reflecting the measurement of predictive uncertainty.

D Intermediate reconstruction uncertainty

This appendix includes all the uncertainty figures based on the intermediate and final predictions.

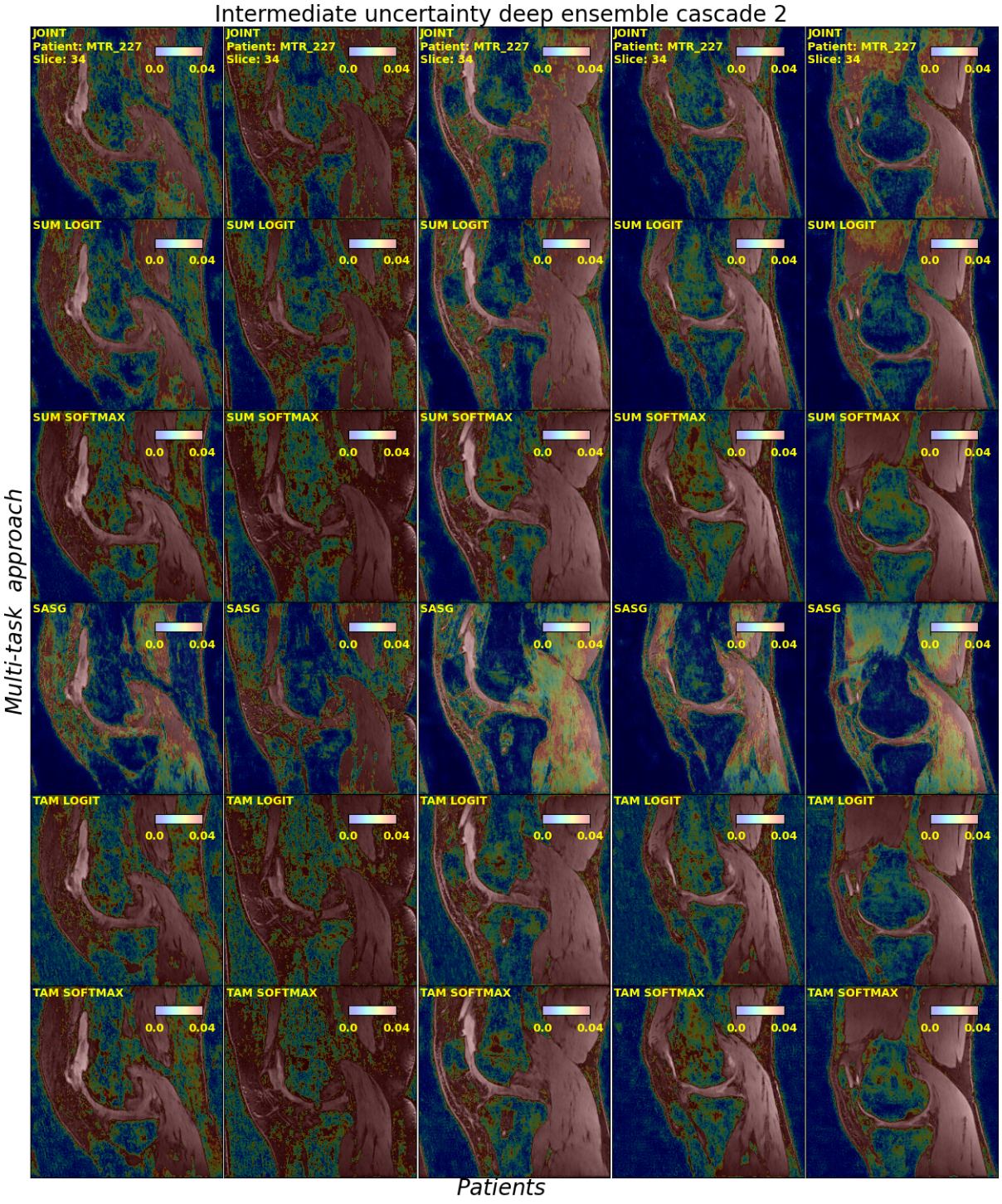


Figure 30: Visualisation of the intermediate uncertainty of the first echo in cascade 2. The uncertainty for each cascade has been determined based on the Deep Ensemble approach, in which each method was trained five times with random initialised parameters. Each row indicates a multi-task approach, and each column is a specific specific patient slice from the test dataset.

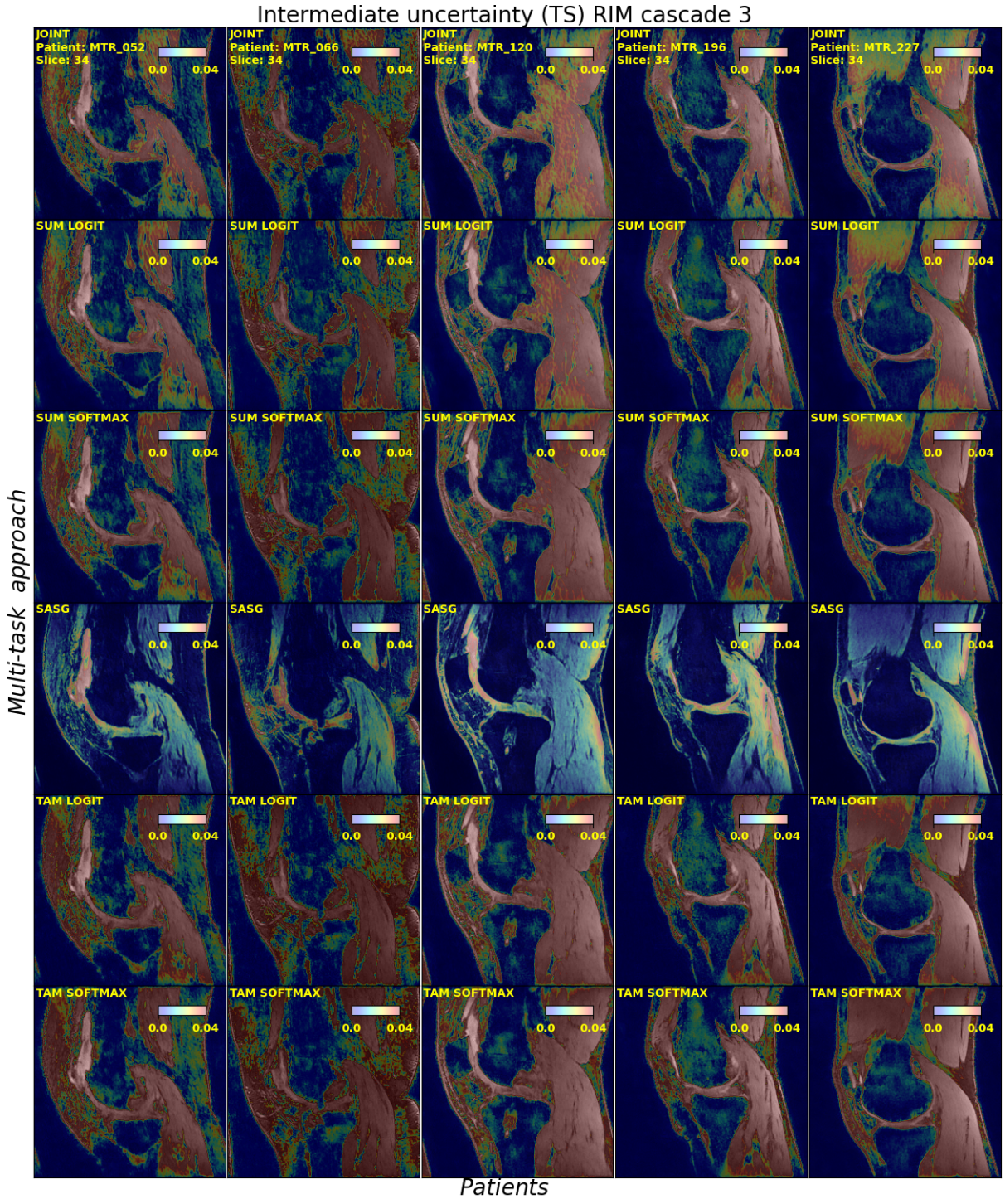


Figure 31: Visualisation of the proposed intermediate uncertainty of the first echo in cascade 3. The uncertainty for each cascade has been determined based on the intermediate time step, in which each time step prediction within a single RIM is used to determine the standard deviation. This shows whether the variations between the individual update predictions within the RIM are similar or contain a high variation. Each row indicates a multi-task approach, and each column is a specific patient slice from the test dataset

E Intermediate update uncertainty

This appendix includes uncertainty figures based on the intermediate and final update predictions.

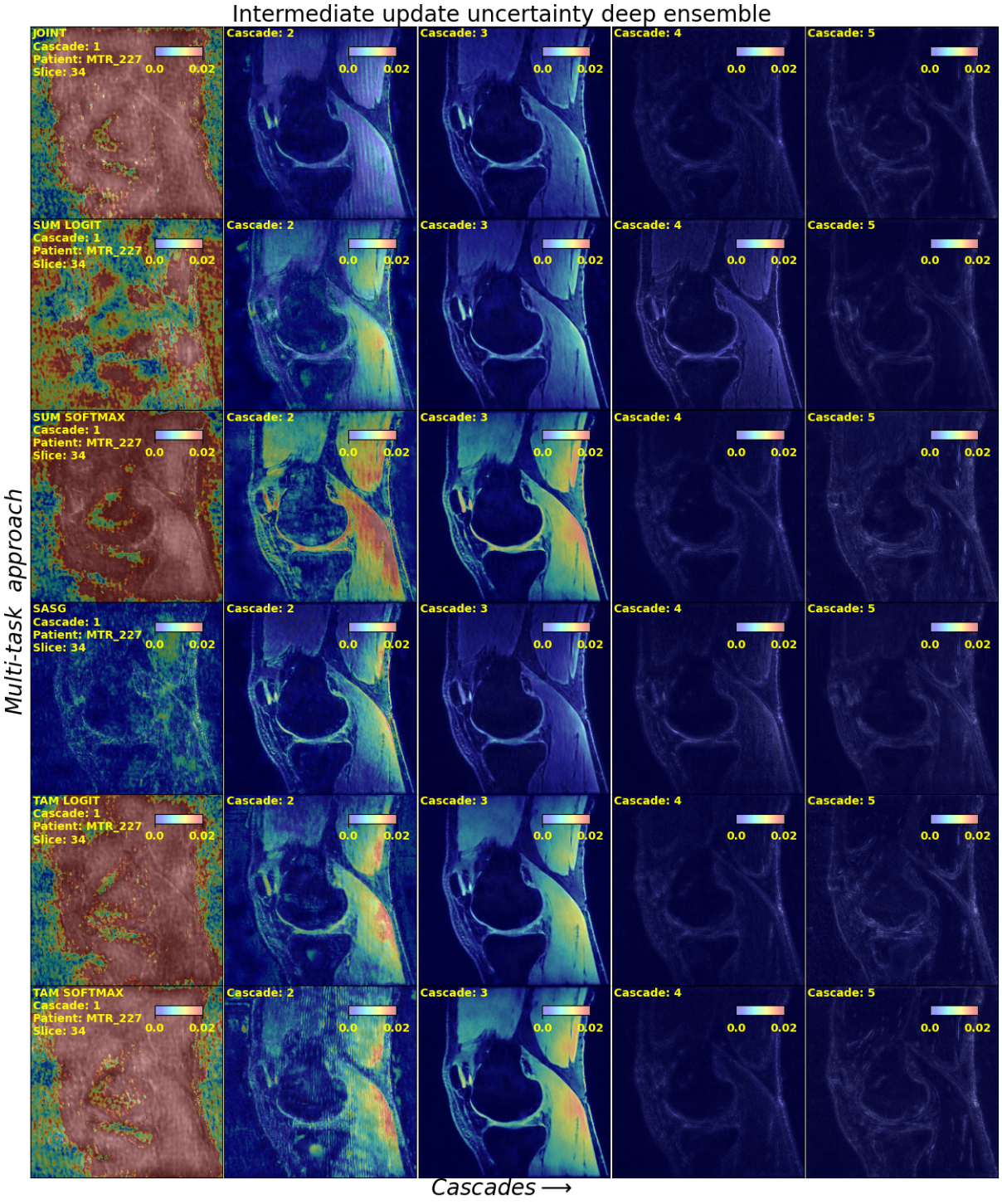


Figure 32: Visualisation of the intermediate uncertainty of the first echo update gradient. The uncertainty for each cascade has been determined based on the Deep Ensemble approach, in which each method was trained five times with random initialised parameters. Each row indicates a multi-task approach, and each column is a specific cascade within the MTLRS structure.

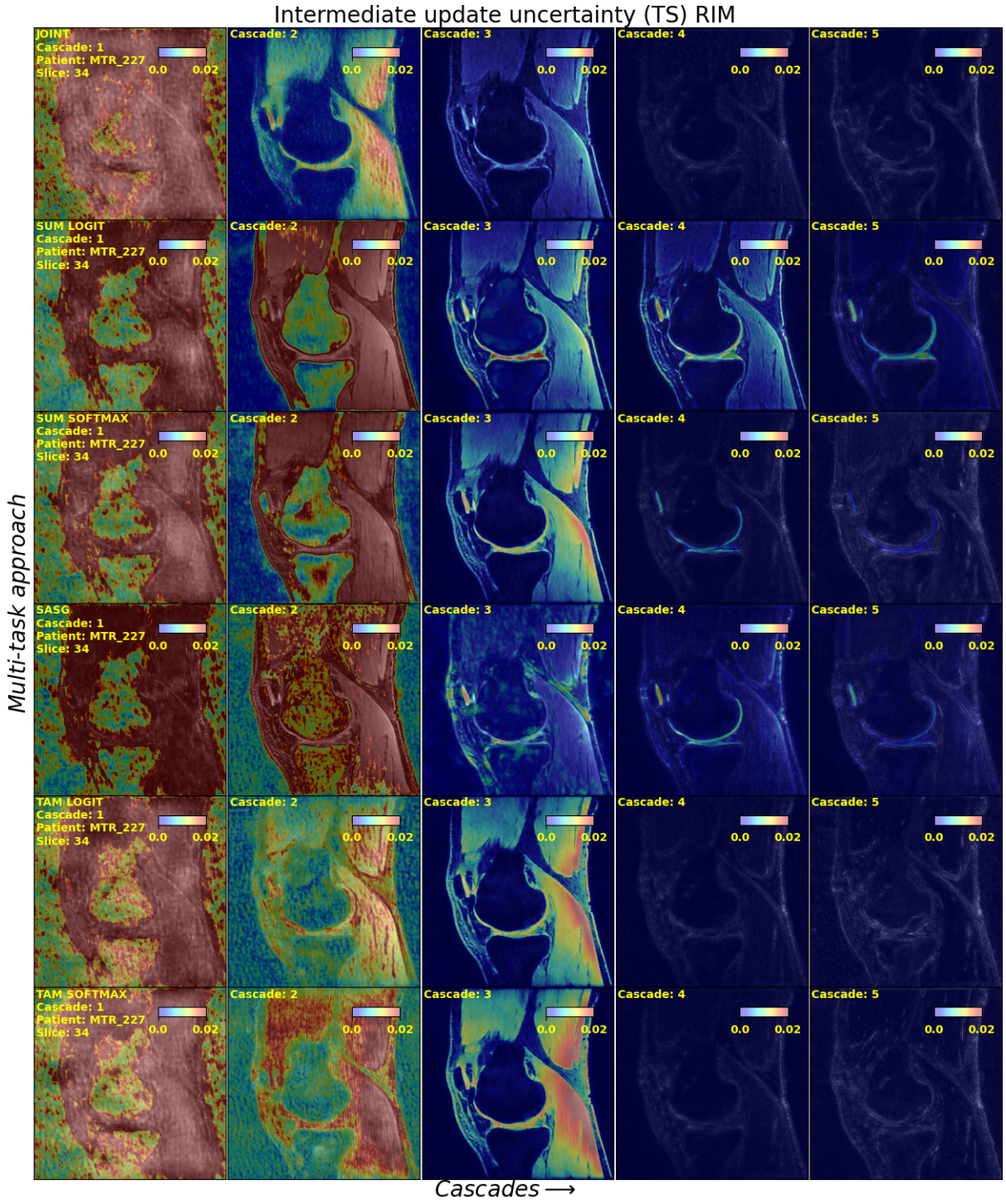


Figure 33: Visualisation of the proposed intermediate uncertainty of the first echo gradient update. The uncertainty for each cascade has been determined based on the intermediate time step update, in which each time step update prediction within a single RIM is used to determine the standard deviation. This shows whether the variations between the individual update predictions within the RIM are similar or contain a high variation. Each row indicates a multi-task approach, and each column is a specific cascade within the MTLRS structure.

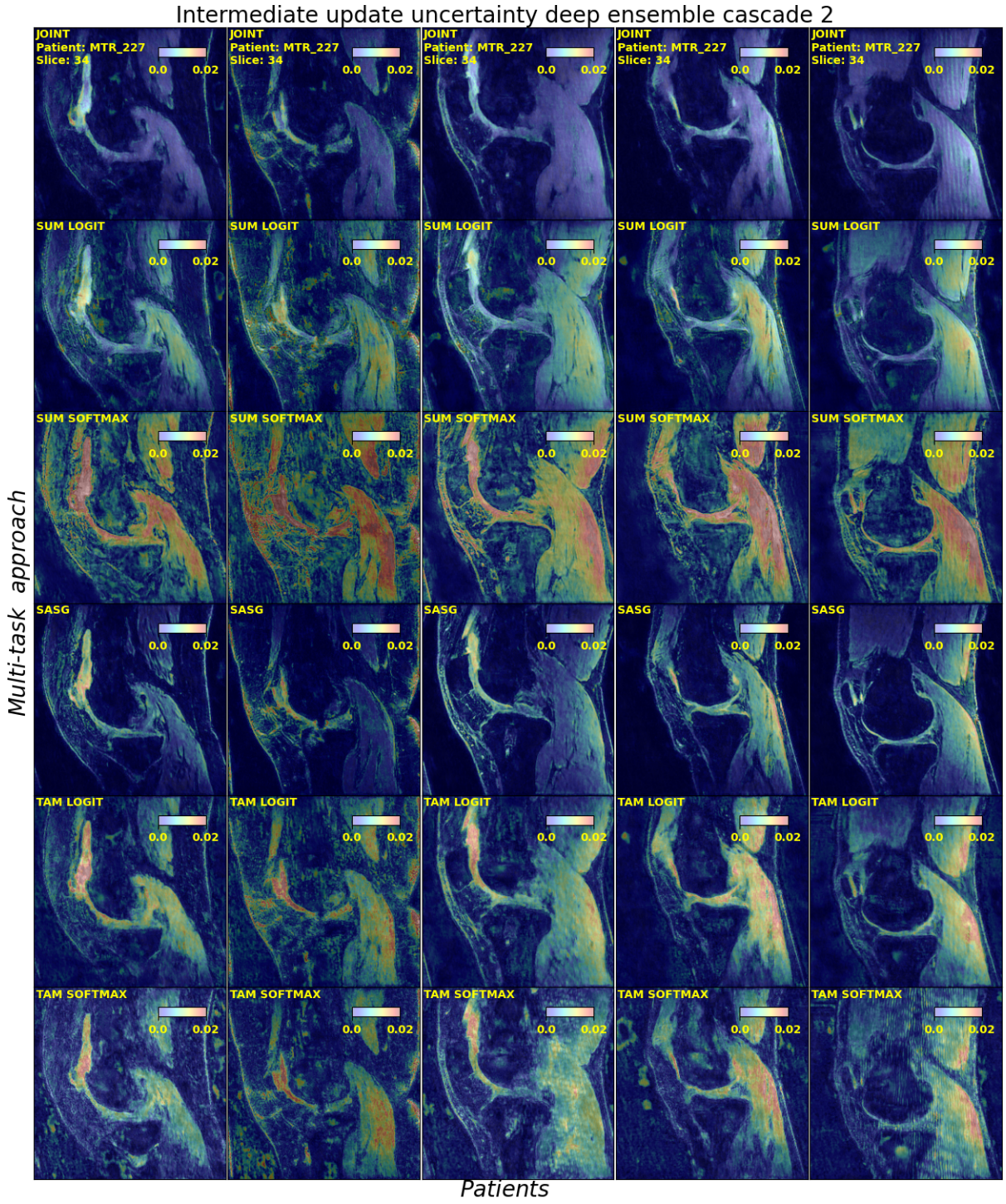


Figure 34: Visualisation of the intermediate uncertainty of the first echo gradient update in cascade 2. The uncertainty for each cascade has been determined based on the Deep Ensemble approach, in which each method was trained five times with random initialised parameters. Each row indicates a multi-task approach, and each column is a specific specific patient slice from the test dataset.

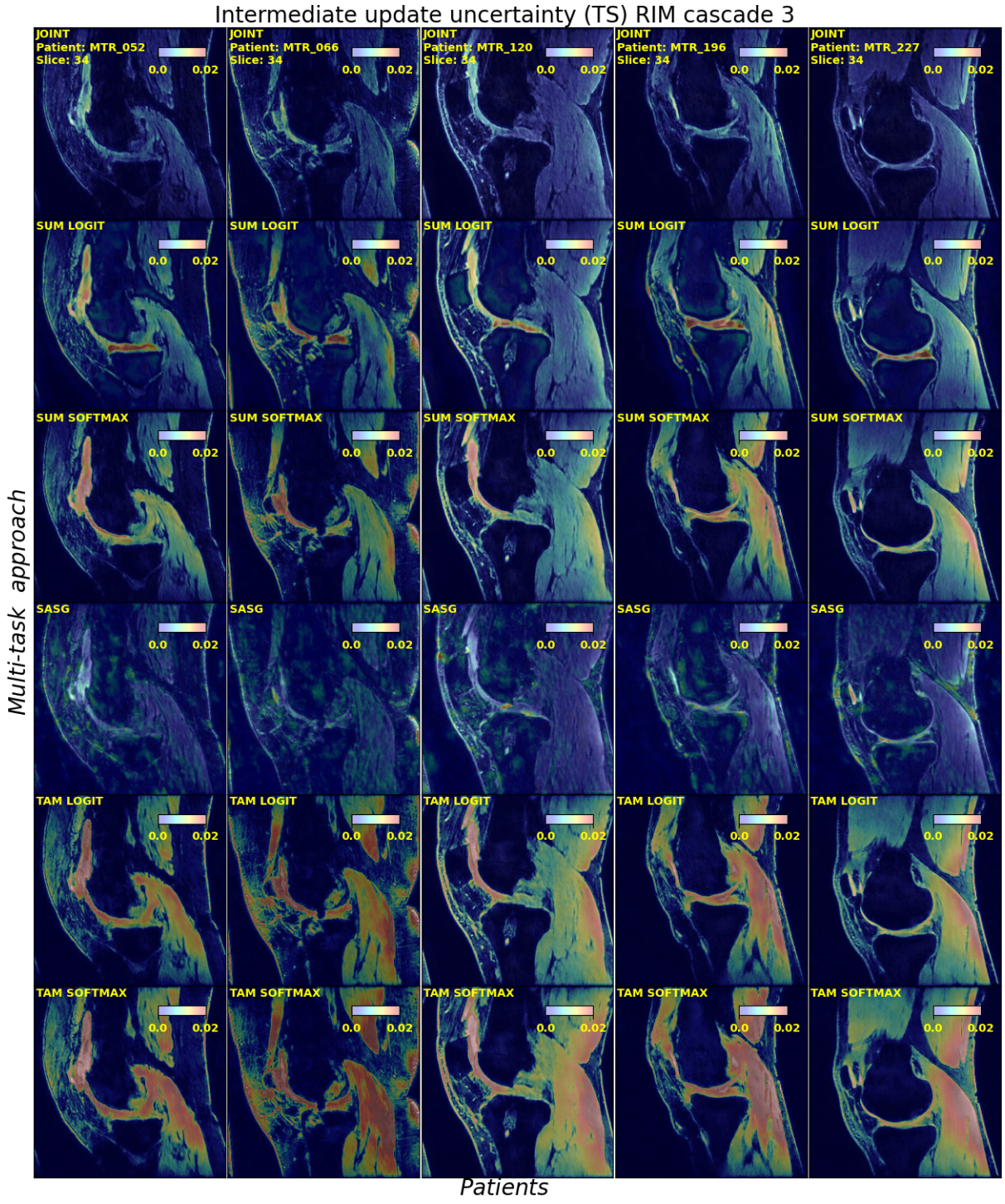


Figure 35: Visualisation of the proposed intermediate uncertainty of the first echo gradient update in cascade 3. The uncertainty for each cascade has been determined based on the intermediate time step update, in which each time step update prediction within a single RIM is used to determine the standard deviation. This shows whether the variations between the individual update predictions within the RIM are similar or contain a high variation. Each row indicates a multi-task approach, and each column reflects a specific patient slice from the test dataset

F Intermediate segmentation

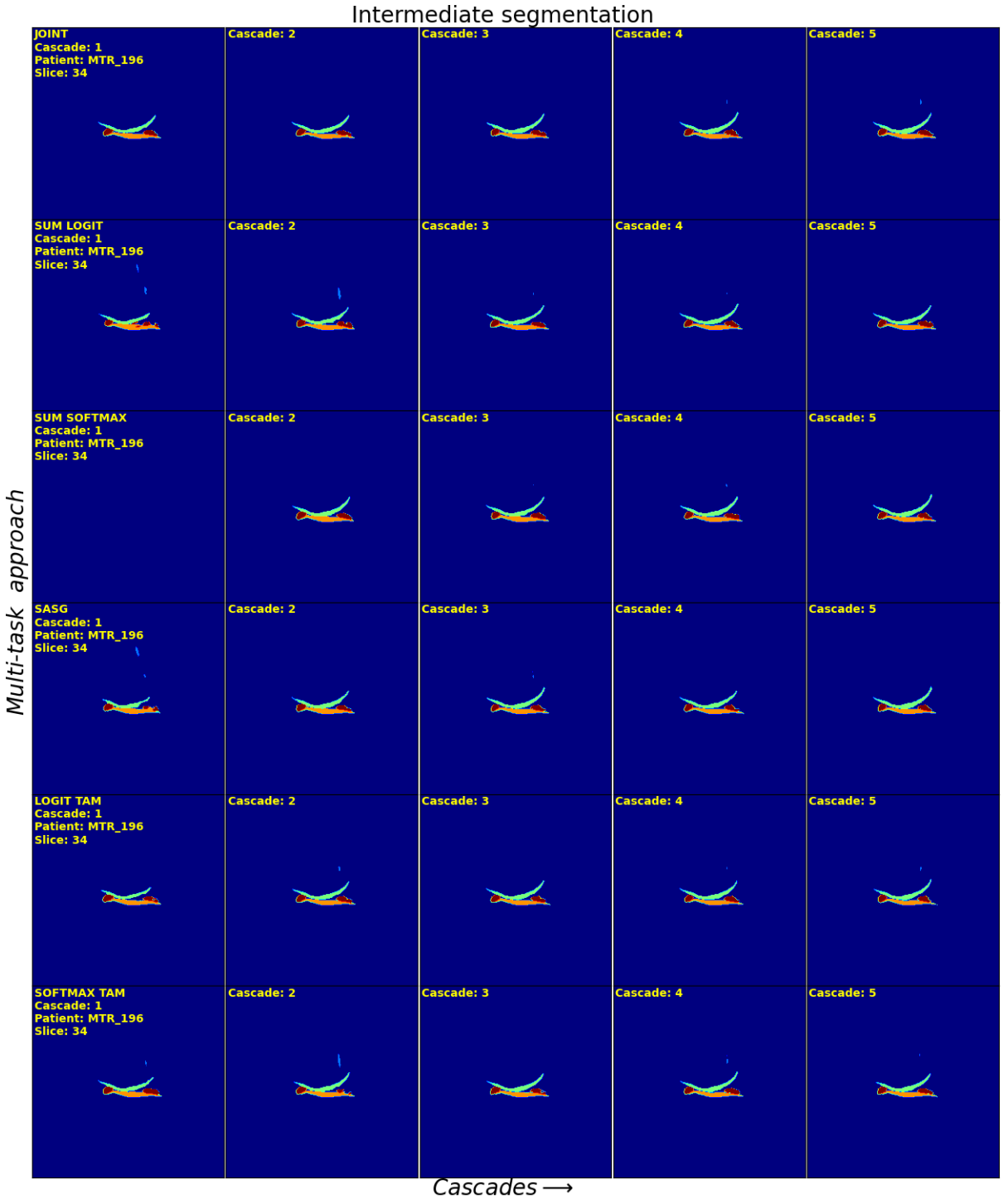


Figure 36: Visualisation of the intermediate segmentation of the. Each row indicates a multi-task approach, and each column is a specific cascade within the MTLRS structure.

G GitHub repositories

GitHub repository for temperature scaling: `https://github.com/gpleiss/temperature_scaling/blob/master/README.md`

GitHub repository Atommic: `https://github.com/wdika/atommic`

GitHub repository Atommic (forked version) including adaptions made in this research: `https://github.com/TimPaquaij/atommic`

GitHub repository PIQ (Image2Image metrics): `https://github.com/photosynthesis-team/piq`

H AI Statement

During this research, the author used ChatGPT 4.0 and Grammarly Pro to correct his generated text in terms of grammar, spelling, and fluency. ChatGPT was also used to assist in coding in Python, bash, and Cuda but not to generate codes and text from scratch. After using these tools, the authors reviewed and edited the content as needed. Furthermore, some code implementations were taken from GitHub; if so, these implementations contain the URL of the source of the data and the paper's author (If available). The author takes full responsibility for the content of the presented work.

I Acknowledgement

For this research, Atommic, an open-source repository for magnetic resonance imaging on GitHub was used. On top of this repository, the suggested approaches have been added and integrated within the workflow to maintain reproducibility. Furthermore, malfunctions and errors found during this research have been addressed with the authors and incorporated within the repository. Special thanks to Dimitris Karkalousos for providing and guiding us through the workflow of the repository.