Deep Learning-Based Contrast Transformation of High Resolution 3D Gradient Echo Images Trained Using Low Resolution 2D Turbo Spin Echo Images

Ryan Pollitt[1]

***Abstract*—Acquiring multiple clinically relevant MR image contrasts from a single scan is an emerging trend in MR imaging to reduce the total scan time of an exam. A deep learning-based approach that takes this idea one step further is BoneMRI, which generates synthetic CT images from MR images. These images are generated using a 3D RF-spoiled T1-weighted multiple Gradient-Echo sequence (GE), but the sequence is often combined with a 2D T1-weighted Turbo Spin Echo (TSE), because the latter provides better T1 contrast from a clinical perspective. To reduce the total scan time, we investigate a deep learning-based approach to generate synthetic TSE images from the GE images. We propose a training approach, called the High-to-low approach, to keep the high through-plane resolution of the GE images, while still applying the contrast transformation using only the lower resolution TSE as target data. Additionally, we implement a network architecture, called HighResNet, and redesign it for the synthesis of TSE images from GE images. The proposed approach and network do not necessarily need to be used together and both were validated against a more often used approach and network, respectively. Experiments using scans of the cervical spine showed that the High-to-low approach was capable of keeping the higher through-plane resolution of the GE images, while also achieving significantly higher image similarity to the lower resolution ground-truth TSE after downsampling. The experiments also demonstrated that HighResNet synthesized high resolution synthetic TSE images with fewer artifacts than the often used U-Net. The results show that a neural network is capable of learning an MR contrast transformation between a higher resolution input image and lower resolution output image without sacrificing performance or image resolution.**

***Index Terms*—Convolutional neural networks, deep learning, image synthesis, Magnetic Resonance Imaging**

I. INTRODUCTION

MAGNETIC Resonance Imaging (MRI) is a versatile imaging modality, which allows for the visualization of both structural and functional properties of tissues. Although many sequences exist with different applications and visualization abilities, a trade-off often has to be made between the amount of information gathered, e.g. by employing multiple different sequences, and the amount of time spent scanning. Shorter sequences, and by extension shorter total scan time, were traditionally achieved by sacrificing some Signal to Noise Ratio (SNR) and/or resolution. A new trend is emerging, however, where instead of shortening the duration of individual scans, the total number of sequences is reduced with sequences and/or processing techniques that are capable of generating multiple contrasts.

Underlying all of the standard MRI contrasts are physical properties of the imaged tissues, like the longitudinal relaxation time T_1, transverse relaxation time T_2, and proton density ρ. These properties, among others, largely determine the signal intensities and image contrasts. In light of using one sequence to generate many contrasts one could argue that this is best achieved using quantitative MRI (qMRI), which measures these physical properties. The necessary contrasts could then in principle be synthesized using signal equations.

Two examples of qMRI are Magnetic Resonance Fingerprinting (MRF) [1] and Magnetic Resonance Spin TomogrAphy in Time-domain (MR-STAT) [2]. MRF uses a pseudorandomized sequence, varying e.g. the flip angle and repetition time, which creates a distinct signal for materials with different physical properties. This signal can then be matched to a precomputed dictionary of possible combinations of physical properties with their simulated signal evolution [1]. In contrast to MRF, MR-STAT is directly applied to the k-space data, performing both the localization and parameter estimation simultaneously. MR-STAT promises fast acquisitions with scan times on the order of seconds per 2D slice. Most of the time required for acquiring the final parameter estimation in MR-STAT is spent on the computation of the parameters, but the philosophy behind MR-STAT is that this compute time is much less expensive than scan time [2]. Another qMRI method is Quantification of Relaxation times And Proton density by Multiecho Acquisition of a Saturation-recovery using Turbo spin-Echo Readout (QRAPMASTER), which is used commercially in the creation of Synthetic MRI (SyntheticMR, Linköping, Sweden). The synthesis consists of the quantitative measures from a roughly 6 minute scan, which are directly used to generate multiple synthetic contrasts with 1 minute of post-

[1] Student of the Master's programme Medical Imaging at Utrecht University

processing time [3], [4].

The qMRI methods rely on physics-driven synthesis, but data-driven approaches using deep learning have also seen a lot of research interest recently, which is another way of reducing the amount of sequences (and thus total scan time) needed to generate multiple contrasts. Examples include the translation from T1-weighted images (T1w) to T2-weighted images (T2w), T1w to T2 Fluid-Attenuated Inversion Recovery (T2-FLAIR), T2w to T2-FLAIR, Proton Density weighted (PDw) to T1w and other permutations of these contrast combinations [5]–[9].

The synthesis of contrasts need not be limited to MRI-to-MRI contrast pairs. Another active area of research is the synthesis of Computed Tomography (CT) images from MRI [10]–[12]. In addition to making more efficient use of scan time, these methods have the potential to render the use of a CT scan redundant in some cases. These methods would significantly decrease the radiation burden on the patient by facilitating the use of radiation-free workflows and would also simplify workflows for which both MRI and CT need to be acquired for soft- and hard tissues [13].

One might assume that mapping from the quantitative MRI maps to CT would perform better, because it is more robust to which type of scanner is used [14] and because it would mean mapping from one quantitative measurement of the tissues to another, instead of mapping from a qualitative to a quantitative one. Because non-quantitative MRI sequences use the physical properties of the tissues to generate contrast, however, these can be tuned to create more salient features to distinguish important tissue types for synthetic CT (sCT) generation. Better contrast between e.g. air, soft tissue and bone might help deep learning-based approaches in detecting relevant features.

Although qMRI methods do allow for the synthesis of additional MRI contrasts, which could then be used for sCT generation, they have one big drawback, which is their relatively large slice thickness. This slice thickness is often required for SNR considerations for accurate parameter estimation, ranging from 3 to 5 mm in the previously mentioned MRF, MR-STAT and QRAPMASTER/Synthetic MRI [1]–[4]. This is far above the common submillimetre resolution of CT scans [15], which makes CT's so attractive, and would in turn limit the usefulness of the sCT created from qMRI-derived images.

Our research will focus specifically on BoneMRI (MRIGuidance, Utrecht, The Netherlands), which is a commercialized deep learning-based implementation of MRI-based CT synthesis. It makes use of a 3D RF-spoiled T1-weighted multiple Gradient-Echo sequence, with an almost in phase- (aIP) and almost opposed phase (aOP) component [11]. These will be referred to simply as (aIP/aOP) GE for the rest of the paper.

The BoneMRI sequence is often performed in combination with a 2D T1-weighted Turbo Spin Echo (referred to as TSE henceforth) sequence, because this is often the preferred T1-weighted contrast in the clinic. Ideally, the TSE could somehow be substituted for the sake of total scan time efficiency.

A natural extension of the BoneMRI workflow to achieve this goal would be to not only synthesize CT images, but improved T1w contrast like in the TSE images as well. The mapping from one T1-weighted contrast image to another is conceptually simpler than mapping from MRI contrast to Hounsfield Units, which was shown to be possible.

Therefore, we investigate different methods of generating TSE images from the BoneMRI GE images with a dataset of cervical spine scans. To the best of our knowledge no research has been performed on generating TSE images from GE images, which is expected, because one could simply substitute a GE for a TSE in most cases if the latter were preferred.

The acquired TSE images have thicker sagittal slices than the GE images. The thicker slices means that although the TSE has a generally preferable contrast, it does come at a cost in the amount of spatial through-plane information contained in the images. Ideally, the extra information contained in the GE images would be kept, while applying the contrast transformation from GE to TSE, resulting in a synthetic TSE with a higher resolution than the ground-truth TSE, equal to the resolution of the input images. This higher resolution synthetic TSE could allow radiologists to investigate certain structures like the neuroforamina much better with the option for multiplanar reformation. The challenge here is that we only have the thicker slice ground-truth TSE images for training, so a new approach is needed to learn the contrast transformation while keeping the higher through-plane resolution of the GE image. In this paper we developed a method to achieve exactly this, which we call the High-to-low approach.

The main contributions of this paper are two-fold:

- We redesign an existing Convolutional Neural Network architecture, called HighResNet to fit the TSE synthesis problem at hand.
- We introduce a new training strategy (the High-to-low approach), which makes more optimal use of the information included in the input images and allows for the reconstruction of synthetic TSE images at a four times higher resolution than the ground-truth TSE images.

The remainder of the paper is structured as follows: the network architecture and training strategy described in the contributions will be discussed in section II with the results provided in section III and ending with the discussion and conclusion in section IV and V, respectively.

II. Methods

This section covers the data and data (pre)processing, the network architectures, the training procedures for the Low-to-low approach, the High-to-low approach, and the evaluation metrics. In contrast to the High-to-low approach, the Low-to-low approach does not keep the higher through-plane resolution of the GE images and instead resamples the GE images to the TSE resolution. The Low-to-low approach mimics the more generally used approach in image synthesis, where both the input- and target images have the same resolution.

II. A. Data

The cervical spine data used in this study were acquired on a 1.5T scanner (Ingenia; Philips Healthcare, Best, Netherlands) and consisted of GE and TSE images from 25 patients over the age of 50 with cervical radiculopathy. A 3D RF-spoiled T1-weighted multiple gradient-echo sequence was used for the BoneMRI images, for which the first echo was acquired almost opposed phase and the second almost in phase. The GE images were acquired at a voxel size[2] of 0.744 mm × 0.744 mm × 1.8 mm with a SENSE factor of 1.3 and reconstructed at 0.744 mm × 0.744 mm × 0.9 mm. The 2D T1-weighted TSE images were acquired in a sagittal orientation with a saturation slab applied to most of the anatomy anterior to the trachea. The TSE images were acquired and reconstructed with a voxel size of 0.488 mm × 0.488 mm × 3 mm with a slice gap of 0.3 mm and an echo train length of nine. The rest of the acquisition parameters are summarized for both scans in Table I.

TABLE I
ACQUISITION PARAMETERS FOR THE GE AND TSE IMAGES

Acquisition parameter	GE	TSE
T_R	7 ms	506 ms
T_E	2.1 ms / 4.2 ms	7 ms
Flip angle	10°	90°
FOV	250 × 250 × 90 mm^3	250 × 250 × 36 mm^3
Bandwidth	542 Hz/pixel	322 Hz/pixel
Acquisition time	3 min 53 sec	2 min 3 sec

In addition to the GE and TSE images, the dataset also contained bone masks, which were created outside of this research from CT images from the same patients and manually refined and subsequently registered to the GE images.

II. B. Data pre-processing

All 25 image pairs were visually inspected, leading to the exclusion of one patient's data, which contained motion artifacts. Body masks were automatically created based on the TSE images for all remaining patients by thresholding and binary filling. The interface between the saturation slab and the anatomy posed a challenge for this method, however, so the saturation slab was removed from the mask manually. Additionally, because the scans were mostly focussed on the spine, most of the image volume anterior to the trachea was manually cropped out.

The GE images were registered to the TSE images for both the Low-to-low and High-to-low approach, so the target data would remain intact and devoid of resampling artifacts. To allow the registration algorithm to take the back/neck and table/air interface into account, the aforementioned body masks were dilated on the posterior side.

Registration was performed using Elastix [16], consisting of a translation and a multi-resolution nonrigid B-spline transform with MI as the similarity metric. The latter was regularized by a rigidity penalty, which enforced the bones and a small area around it to stay rigid using the bone masks. The penalty partially prevents the registration from applying nonrigid transformations locally on the bones, because these are only able to move rigidly in the body. Because the bone masks were based on CT scans they did not completely cover the same anatomy as the MRI scans due to a difference in patient position and FOV. The registrations of six patients visibly failed near these missing parts of the bone mask with bones bending in physically impossible ways. Therefore, these patients' bone masks were manually edited to include all of the skull/vertebrae, where necessary. Note that these were of a lower quality than the existing bone masks and were therefore only used for registration and not during evaluation.

All registration parameters were qualitatively tuned by visually assessing the registration quality in three representative patients. During this process one additional patient was excluded, because they lacked a bone mask, which was needed for proper registration, leaving 23 patients.

After registration, the GE images were resampled to the TSE's resolution of 0.488 mm × 0.488 mm × 3.3 mm for the Low-to-low approach and to a resolution of 0.488 mm × 0.488 mm × 0.825 mm for the High-to-low approach.

To sample useful voxels, the aforementioned body masks were also used for training. They were dilated on the posterior side to include the body/background interface with a kernel size equal to the patch size. The dilation ensures that a sampled patch contains a maximum of ~50% background, so the network is not trained on purely background.

Because the GE images were registered to the TSE images, the GE would sometimes get pushed inwards near the edges, leaving "empty" voxels, which were automatically set to a value of -1, allowing them to be easily detected and masked out as well.

Finally, a threshold was applied to the GE images to set any negative values from the 3rd order B-spline interpolation to 0 and both the GE and TSE images were normalized by dividing by the 99th percentile intensity out of all values inside the sampling mask.

II. C. Networks

1) U-Net

We used two distinct network architectures to perform the GE to TSE translation. We used a standard implementation of the well-known U-Net as a baseline [17], with one small addition: a configurable Batch Normalization (BN) layer after each 3×3×3 convolution. U-Net was chosen, because it was used successfully for medical image synthesis by itself or inside a Generative Adversarial Network (GAN) as the generator [9], [12], [18], [19].

[2] Any sequence of numbers relating to the three spatial dimensions in this paper will be written in the order: superior-inferior, anterior-posterior, left-right

2) HighResNet

Our proposed network was a network derived from HighResNet as described in [20], which is a residual network. The basic idea behind residual networks that makes them attractive for the GE to TSE mapping is that they have identity mappings built into them. This built-in identity mapping is something other types of networks do not have and struggle to learn [21]. This property should allow the network to propagate information contained in the GE images more easily, which could be beneficial for training, because the (aIP) GE images and TSE images are already relatively similar.

What separates HighResNet from most other residual networks is that it does not perform any downsampling on the input or feature maps. Instead, it uses dilated convolutions to capture information at increasingly larger scales. The potential benefit of the dilated convolutions is that the feature maps remain at the resolution of the input images, which can prevent blurring from downsampling operations, like the max pooling used in U-Net [10].

The architecture of HighResNet as used in this paper is shown in Fig. 1, which has some modifications compared to the one from [20]. Firstly, the original implementation contained a 3×3×3 convolution before all of the residual blocks, which would not allow for the information contained in the GE to flow through the network unhindered, so this layer was removed. Secondly, if the number of channels increases after a 3×3×3 convolution before the element-wise addition, a mechanism is needed to increase the number of channels of the data that flows through the residual connection. In the original network the increase in channels was achieved by zero-padding along the channel dimension with the extra amount of channels necessary. The zero-padding was replaced by a 1×1×1 convolution (indicated by the dark blue arrows), because this so-called projection shortcut was found to perform slightly better in [21]. Finally, whereas the original network used dilation factors of 1 (no dilation), 2 and 4 along every dimension, we implemented dilations that would result in the most isotropic kernels possible. Note that the Low-to-low training procedure and the High-to-low approach used input data with different resolutions, resulting in dilation factors of (4, 4, 1) and (4, 4, 2), respectively.

Both networks were implemented in Python 3.7 using PyTorch 1.5.0/1.10.

II. D. Training for the Low-to-low approach

To train the networks using the Low-to-low approach, 3D patches were sampled at random locations inside the sampling mask from the GE- and TSE images. Each batch contained patches sampled from as many different patients as possible, by cycling through the patients after sampling each patch.

The aIP- and aOP GE patches were concatenated along the channel dimension to form the input, which were forwarded through the network to finally be compared to the TSE patches via a simple L1 loss, which computes the mean absolute error between the output- and the target data. The L1 loss is one of the most used loss functions in image synthesis and creates less blurry images than an L2 loss (mean squared error) [10].

The patches had a size of 56×56×8 voxels (27.3 mm × 27.3 mm × 26.4 mm), which was chosen with U-Net's max pooling layers in mind. These force all of the dimensions of the data to be divisible by $2^{\#\text{max pooling layers}}$, because each of these layers reduces the size by a factor two. Because the data only contained eleven voxels in the left-right direction, the maximum patch size was constrained to eight along this dimension. The number of voxels in the other two directions were chosen such that the patches were as isotropic as possible, while being divisible by $2^{\#\text{max pooling layers}}$ in all three dimensions.

1) Hyperparameter search

To find the best set of hyperparameters for HighResNet and U-Net a grid search was employed to search for the best learning rate, number of channels C (as in Fig. 1; directly influences the number of trainable parameters), and batch size. On top of training a regular U-Net, we also trained a U-Net with a Batch Normalization (BN) layer after each 3×3×3 convolution to more closely match HighResNet, which also employs BN layers.

The different batch size- and learning rate values were chosen based on preliminary tests with U-Net. The number of trainable parameters were chosen to match as closely as possible between U-Net and HighResNet and were constrained by HighResNet's larger memory footprint. The possible values for these hyperparameters during the grid search are summarized in Table II.

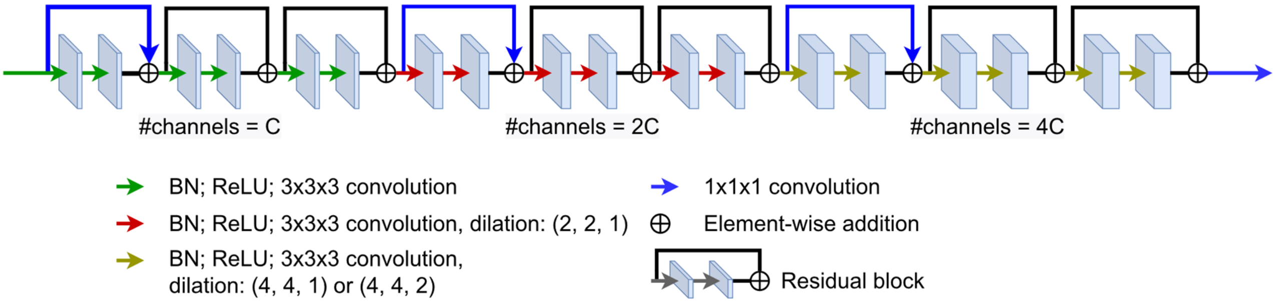

#channels = C
#channels = 2C
#channels = 4C
BN; ReLU; 3x3x3 convolution
BN; ReLU; 3x3x3 convolution, dilation: (2, 2, 1)
BN; ReLU; 3x3x3 convolution, dilation: (4, 4, 1) or (4, 4, 2)
1x1x1 convolution
Element-wise addition
Residual block

Fig. 1. Schematic overview of HighResNet's architecture. The cuboids denote feature maps, which double in number of channels every time the dilation factor changes. Note that although the projection shortcuts/residual connections connect to the arrows of the Batch Normalization (BN), Rectified Linear Unit (ReLU) and convolution, they do not act on the output of these, but on their input.

TABLE II
HYPERPARAMETER VALUES FOR THE GRID SEARCH

Model	Batch size	Learning rate	#Trainable Parameters (M)
HighResNet	16, 32, 64	1×10^{-3}, 5×10^{-3}, 1×10^{-2}	2.446, 2.998, 3.397
U-Net*			2.563, 3.050, 3.579

* Note that the BN layers contain additional trainable parameters, but these increase the total number of trainable parameters by a negligible amount

All three networks (HighResNet, U-Net, U-Net+BN) were trained with all 27 possible combinations of the selected hyperparameters with the Adam optimizer and a weight decay of 1×10^{-5}.

To constrain the amount of time needed for these experiments and to evaluate the different combinations at similar points in their training trajectories in terms of convergence, an early stopping criterium was used. The criterium was defined as follows: if the validation loss does not decrease with 0.1% with respect to the lowest validation loss measured so far for 80 epochs in a row, the training is stopped. To smooth out random variations in the validation loss a Gaussian kernel with a standard deviation of 10 was applied to the loss values before evaluating the early stopping criterium.

Normally, an epoch is defined as having trained the network on all of the training data once (and optionally validated on all the validation data), but because we randomly sample patches from the training data this definition is not applicable. Instead, an epoch was defined as 800 training patches and 160 validation patches, which corresponds roughly to the amount of voxels that fall inside the training- and validation sampling masks, respectively.

All 27 hyperparameter combinations were used to train each network once until convergence (according to the early stopping criterium) with the same set of 15 training patients and 3 validation patients. To compare the hyperparameter combinations more robustly, a five-fold cross-validation was performed with the top 10 hyperparameter combinations from the aforementioned 27 runs. The top 10 was chosen based on the median validation loss across 3200 validation patches, which is roughly 20 times the amount of voxels eligible for sampling in the validation set. Only the top 10 of these runs were cross-validated to save on runtime. The same early stopping criterium as before was used.

The median loss across 3200 validation patches was computed for each of these 5×10 runs. The mean of the five medians from the cross-validations was computed to determine which of the combinations achieved the lowest validation loss.

The resultant best hyperparameter combinations are shown in Table III. These were used to train all three networks five more times for 1500 epochs, which is roughly double the maximum amount of epochs needed to trigger the early stopping criterium out of all 150 cross-validation runs. Additionally, all but one of the 18 patients used for training/validation in the previous experiments were used for training. One patient's data, which contained motion artifacts was excluded, with the remaining 5 patients used as the test set.

TABLE III
BEST HYPERPARAMETER COMBINATIONS PER NETWORK

	Learning rate	Batch size	#Trainable parameters
HighResNet	0.001	16	2.446 M
U-Net	0.0001	16	3.579 M
U-Net+BN	0.0001	16	2.563 M

Another difference compared to the other experiments for the final runs is that reflection padding was used instead of zero padding, because it was found to work better in the High-to-low approach. The reflection padding also resulted in both visual and quantitative improvements for the Low-to-low approach. We did not repeat the earlier experiments with reflection padding because of time limitations, but we do not expect much of an effect from the padding on which hyperparameter combination works best.

II. E. High-to-low approach

The High-to-low approach makes use of the fact that the GE images contain a lot of information that is partially lost when downsampling to the TSE resolution (0.744 mm × 0.744 mm × 0.9 mm → 0.488 mm × 0.488 mm × 3.3 mm) as was done in the Low-to-low approach. Instead, the GE images were resampled to an intermediate resolution of 0.488 mm × 0.488 mm × 0.825 mm, foregoing the 3.67× downsampling in the left-right direction and therefore maintaining more of the information contained in the GE. Because the target data has only been acquired with thicker sagittal slices, a different strategy is necessary to be able to train a network on these input- and target data. An overview of the approach is shown in Fig. 2. The High-to-low approach trains a network to use the extra information contained in the GE images, while also synthesising the TSE images at a higher resolution than the target data.

To achieve the higher resolution, the network operates fully at the resolution of the GE images, resulting in a high resolution synthetic TSE (sTSE_{HR}) as the output of the network, which is subsequently downsampled by a factor of four in the left-right direction by using a strided convolution with a 3D downsampling kernel of size 1×1×4. The downsampled/low resolution synthetic TSE (sTSE_{LR}) is subsequently compared to the target TSE via an L1 loss like in the Low-to-low approach. Additionally, the sTSE_{HR} image is optionally compared to the aIP GE image with a structural consistency loss that keeps some of the structure present in the aIP GE.

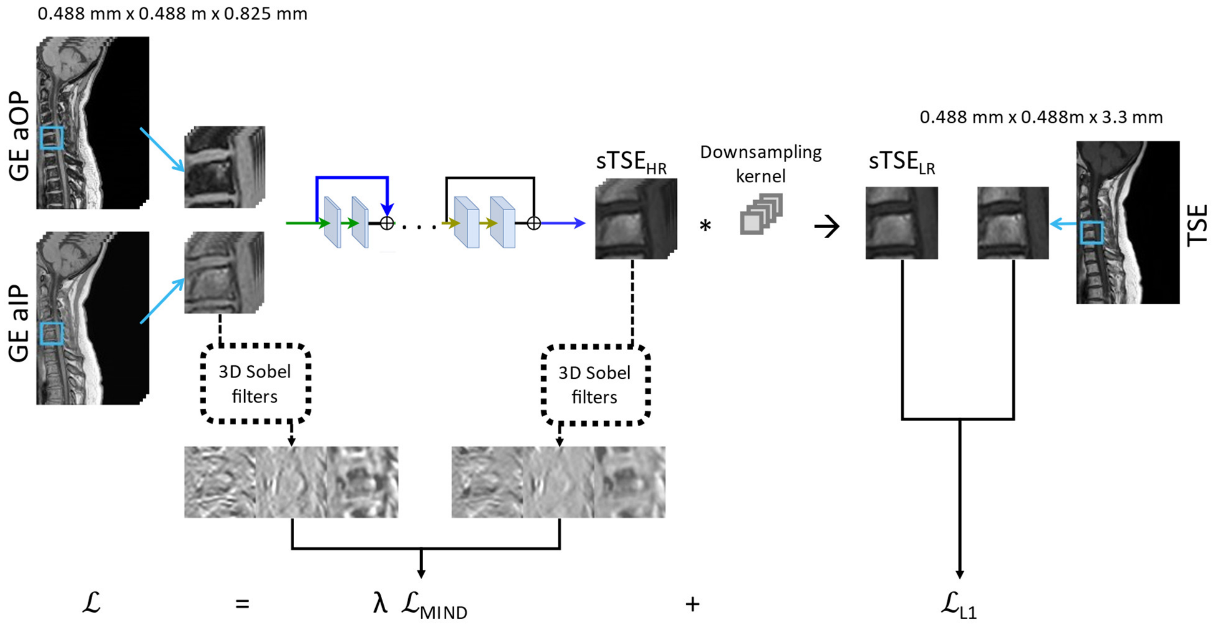


Fig. 2. Overview of the High-to-low approach during training. Patches are extracted from both GE images, which both have a four times smaller voxel size in the left-right direction than the TSE, indicated by the four sagittal slices (note that all operations are performed in 3D). The network operates at the resolution of the GE images throughout, resulting in a high resolution synthetic TSE (sTSE_{HR}), which is subsequently downsampled with a factor of four in the left-right direction. The high resolution output of the network is optionally compared to the aIP GE patch via the $\mathcal{L}_{\text{MIND}}$ term, which is applied to the 3D Sobel filtered patches. The low resolution output is compared to the target TSE image with an L1 loss.

1) Downsampling kernel

The downsampling kernel is an important part of the High-to-low approach. Due to the way the data was resampled, each sagittal TSE slice has four corresponding GE slices. Because there is a slice profile over the actual spatial extent of each TSE slice, however, it is more similar to some of those four GE slices than others. Using the similarity metrics Structural Similarity Index Measure (SSIM) and Mutual Information (MI), we found that the two left-most of these slices were always more similar to the TSE, followed by the two neighbouring slices on either side (extending into a different set of four GE slices on the left side).

Therefore, the downsampling kernel would ideally weight the most similar slices the most when downsampling. In Python this could correspond to a kernel with the weights of e.g. [1/2, 1/2, 0, 0]. But when a kernel of this form is used, the network learns to recognize the fact that two of the slices are irrelevant to the loss and it starts to exhibit slice-dependent behaviour such that every two sagittal slices the sTSE_{HR} looks very different. The learning of spatial locations is a known phenomenon in deep learning and can even happen with fully convolutional neural networks like HighResNet, although it is much more likely to happen with networks that include e.g. max pooling and transposed convolutions like U-Net [22].

To circumvent the slice-dependent behaviour it is also not possible to use a kernel that takes all slices into account even if all of the weights are the same, e.g. [1/4, 1/4, 1/4, 1/4]. From experiments we noticed that the networks were able to learn that the two left-most slices were more similar to the TSE image, which lead the network to create images where these slice pairs in the sTSE_{HR} became slightly brighter than the others.

To make it impossible for the network to learn which slice is which, we shifted the GE patch—and by extension the sTSE_{HR} patch—with a random offset of a few voxels to the left or right for each iteration and moved the kernel weights with it as shown in Fig. 3. The top right shows the average weighting per image slice, where the two slices that are weighted by ⅓ represent the slices that are most similar to the TSE.

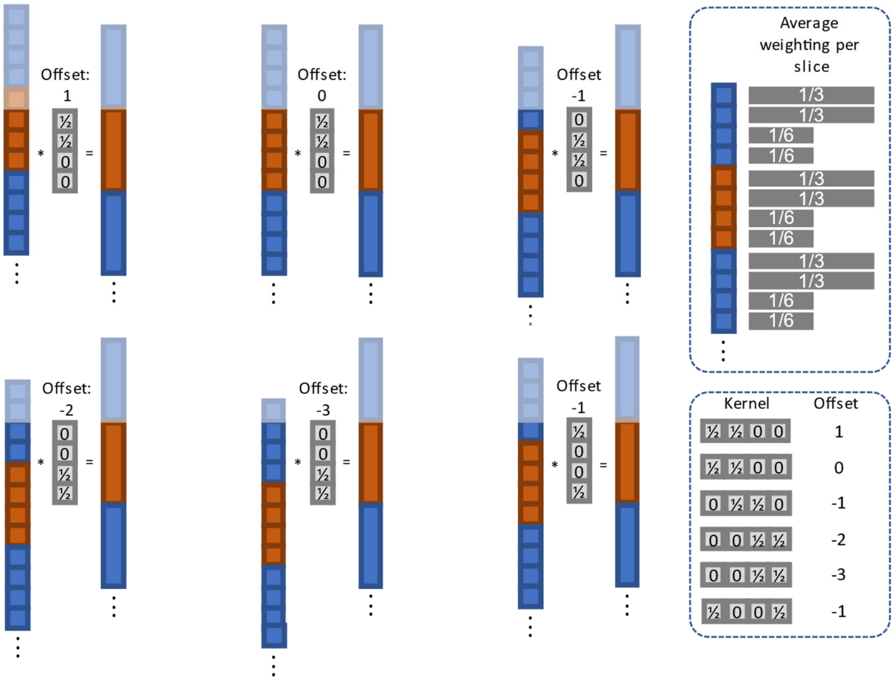


Fig. 3. The random offset applied to the sTSE$_{HR}$ patch (squares) in relation to the sTSE$_{LR}$ patch (rectangles) to make it impossible for the network to learn which spatial location contains which sagittal slice. The average weighting per slice shows how much each of the slices is weighted on average from the image's perspective.

Fig. 3 shows all of the possible kernel/offset combinations on the bottom right, which have the important properties that the average weighting of each location is 1/4 from the network's perspective and that each location is included and excluded equally across multiple iterations. From the image's perspective, however, the two slices that are more similar to the TSE are weighted twice as much as the other two, as evident from the average slice weighting on the top right of Fig. 3.

Note that on average the orange voxel in the sTSE$_{LR}$ is computed by weighting the two top-most orange voxels in the sTSE$_{HR}$ by a factor of 1/3 and the orange and blue voxel above and below those two with a factor of 1/6. This is because the location of the aforementioned blue voxel in the GE is more similar to the orange location in the TSE than the lowest of the four orange voxel locations. Therefore, with an offset of -1 this average weighting could be described as a kernel with weights [1/6, 1/3, 1/3, 1/6], but for an offset of 0 the blue voxel on top would not be taken into account with a kernel of size 1×1×4.

Finally, another type of data augmentation that was applied during the High-to-low approach was flipping of the patches along the left-right direction every other iteration.

2) *Structural consistency loss*

The structural consistency loss term forces the network to keep some of the detail present in the aIP GE and was implemented using the Modality Independent Neighbourhood Descriptor (MIND) and Sobel filters. Note that this loss term is a more general concept and can also be applied to the images themselves and/or using a different metric like the MI or SSIM.

The Modality Independent Neighbourhood Descriptor (MIND) loss was based on [23], in which it was used to guide the optimization of multi-modal deformable image registration. The MIND loss enforces the edges of the sTSE$_{HR}$ to correlate locally to those of the aIP GE image. We hypothesized that this could be beneficial visually, because a lot of detail is lost when only an L1 loss is used, creating smooth-looking images.

To compute the loss a MIND map of both tensors is created, which serves as a modality independent intermediate representation of the images, allowing their differences to be minimized without affecting the underlying contrast of the sTSE$_{HR}$. The MIND map has a size B×6C×D×H×W (batch size, channels, depth, height, width), where B×C×D×H×W are the original tensor dimensions of e.g. the Sobel features (C=3). It has six times as many channels, each of which tell us something about how similar a small 3×3×3 area around each of those voxel's six neighbours are to its own surrounding area, quantified by a weighted square of differences between all the voxel values. For a more in-depth formulation of the MIND map, see [23].

With the MIND maps of both tensors determined, the MIND loss is defined as the MSE between both MIND maps:

$$\mathcal{L}_{MIND} = \frac{1}{|\mathcal{R}|}\sum_{r\epsilon\mathcal{R}} |MIND(GE_{aIP,Sobel}) - MIND(sTSE_{HR,Sobel})|^2, \quad (1)$$

where $GE_{aIP,Sobel}$ and $sTSE_{HR,Sobel}$ are the tensors containing a batch of patches from the aIP GE and sTSE$_{HR}$ images after application of the Sobel filters, respectively, with $\mathcal{R}$ representing all of the voxels.

With the MIND loss in place the total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{L1} + \lambda\,\mathcal{L}_{MIND}, \quad (2)$$

where λ is set to 0 for experiments with only the L1 loss (referred to as L1-only) and set to 0.1 for all experiments using the MIND loss (referred to as MIND). A weighting of 0.2 was also used, but this deteriorated the metrics too much, so it was fixed at 0.1.

3) Training

Both HighResNet and U-Net were trained with the High-to-low approach. Training was performed on patches of size 56×56×32 (27.3 mm × 27.3 mm × 26.4 mm) voxels, which have the same spatial extent as in the Low-to-low approach, but with more voxels.

The learning rate and batch size that were found to work best for HighResNet and U-Net in the Low-to-low approach were transferred to the High-to-low approach, but the number of trainable parameters for HighResNet had to be reduced from its best setting (as found in the grid search) to 1.797 M to fit the model on the GPU's. U-Net's number of trainable parameters were scaled down with roughly the same factor to account for HighResNet's reduction in parameters. Additionally, the networks were only trained for 1000 epochs instead of 1500, because training at this higher resolution is much slower. Like in the final runs of the Low-to-low approach, each of the High-to-low runs was repeated five times with 17 patients in the training set and 5 patients in the test set.

4) Comparison to the Low-to-low approach

To compare the High-to-low approach to the Low-to-low approach, a few adjustments were made to the latter. The High-to-low approach makes use of the prior knowledge that some of the GE slices correspond more to the TSE image than others and effectively uses a convolutional downsampling kernel with weights [1/6, 1/3, 1/3, 1/6] (note that this kernel only works for an offset of -1) as demonstrated in Fig. 3.

Therefore, training was performed again, this time using the Low-to-low approach with GE images resampled to the TSE resolution using the weighted downsampling by convolution (after resampling to a resolution of 0.488 mm × 0.488 mm × 0.825 mm with 3rd order B-spline interpolation during registration). This resampling scheme resulted in significantly higher SSIM and MI between the GE and TSE as compared to resampling directly to the TSE resolution using a 3rd order B-spline interpolation.

Training was performed for 1000 epochs with the same parameter count as used in the High-to-low approach. Additionally, because the High-to-low approach used flipping along the left-right direction as extra data augmentation, this was also applied for the comparison.

5) Validation of the MIND loss

The High-to-low approach produces high resolution synthetic TSE images, for which no ground-truth image exists. To validate the use of the MIND loss and its effect on the sTSE_{HR}, we designed a separate experiment. Here, we also used the weighted GE downsampling scheme as described in the previous section.

Using this version of the GE data, the network operates at the resolution of the actual TSE data, which is 0.488 mm × 0.488 mm × 3.3 mm. The data is subsequently downsampled by a downsampling kernel that operates in the superior-inferior direction, resulting in a low resolution synthetic TSE with voxels of size 1.952 mm × 0.488 mm × 3.3. The new sTSE_{LR} image is subsequently compared to a downsampled version of the real TSE image, downsampled with average kernel weights of [1/6, 1/3, 1/3, 1/6] as before. By training in this way, we have a ground-truth image to which we can directly compare the sTSE_{HR} to see if the MIND loss assists in creating better images at the higher resolution. Note that only the downsampled TSE images are used during training in this approach, so the network is never trained on the actual TSE images in any way.

6) Training the downsampling kernel

During training the downsampling kernel weights are fixed and randomly chosen in correspondence with the offset as shown in Fig. 3. These weights might not be optimal for inference, because they are subject to certain constraints to prevent the network from showing slice-dependent behaviour. To find the most optimal downsampling kernel, the weights of the network were frozen after training and the downsampling kernel was trained with the same hyperparameters as the network, but without weight decay, flipping and random offsets, and with a lower learning rate of 0.0001. The offset was fixed at 0 such that no padding was required. The weights were initialized at [1/2, 1/2, 0, 0].

II. F. Evaluation

1) Inference

Patches were randomly sampled from the image volumes during training, but during inference a different sampling strategy was applied. Patches were sampled with a fixed stride that was smaller than the patch size, creating overlap between patches such that most voxel intensities were predicted multiple times. Each patch was weighted by an isotropic Gaussian kernel with sigma's heuristically set to 1/6th of the patch size (in voxels) in each dimension. This kernel decreases the number of potential artifacts near the edges of the patch resulting from padding, because these locations are weighted less than the centre of each patch.

The aforementioned weighted fusion is also applied in the High-to-low approach, but is slightly more intricate there. The sTSE_{LR} image is not created by downsampling the sTSE_{HR} after weighted fusion, but both the sTSE_{HR} and sTSE_{LR} images are constructed simultaneously by weighted fusion. In practice, this means that each sTSE_{HR} patch is predicted by the network and subsequently downsampled to be put into its corresponding location in the sTSE_{LR} image. Therefore, the stride for sampling the GE patches has to be equal to four in the left-right direction to keep the 4-to-1 correspondence between the sTSE_{HR} and sTSE_{LR} slices. During inference two different kernels were used: (1) a kernel with weights [1/2, 1/2, 0, 0] and (2) the trained downsampling kernel as described in the previous section.

The strides were originally set to a third of the patch size, resulting in a stride of (18, 18, 2) for the Low-to-low approach. But as mentioned before, the stride had to have a value of four in the left-right direction for the High-to-low approach, giving a stride of (18, 18, 4) for the sTSE_{HR}, which translates to a stride of (18, 18, 1) for the sTSE_{LR}. Therefore, in the comparison between the Low-to-low- and High-to-low approach, a stride of (18, 18, 1) was used for the Low-to-low approach.

2) Metrics

Three evaluation metrics were evaluated on the normalized data inside the body mask and the bone mask for all comparisons:

the Mean Absolute Error (MAE), Mutual Information (MI) and Structural Similarity Index Measure (SSIM). The MAE shows how well the networks have been able to learn the GE to TSE translation as formulated during training, because this is always a part of the loss function.

Because the MAE is (part of) the loss function, the other two metrics give a more training-independent evaluation of the performance. The MI gives a global measure of the nonlinear correlation between the sTSE and TSE, which is more forgiving to systematic under- or overestimations than the MAE. These systematic under- or overestimations could leave the overall contrast roughly the same, as long as they are small enough, so it is important to have a metric that allows for this. In all cases the MI is evaluated with 128 histogram bins. Finally, the SSIM looks at the local image structure and was originally developed to quantify perceived image quality; it is maximized when both input images are identical [24]. The original paper used a window of size 11×11 pixels, which was increased to a size of 21×21×3 voxels to take the 3D image structure into account, while keeping a roughly isotropic window size.

3) Statistical analysis

To assess differences between networks or approaches a paired Student's t-test was used to compare the means of all five patient metrics from two networks/approaches, indicating if the means over the patients and repetitions were significantly different. For all tests $P<0.05$ was considered as a statistically significant difference between the two means.

III. Results

III. A. Low-to-low approach

In the Low-to-low approach, the hyperparameters of HighResNet, U-Net and U-Net+BN were all tuned via a partially cross-validated grid search with the resulting best hyperparameter combinations shown in Table III. Using these hyperparameter combinations all three networks were trained five times resulting in the evaluation metrics in Fig. 4. These were evaluated inside the entire body mask and inside the bone mask (denoted by the bone subscript).

Fig. 4 clearly demonstrates that HighResNet significantly outperformed both U-Net variants on five of the six evaluation metrics. Additionally, it shows only a significant difference between the two U-Net variants on the MI_{bone} metric, indicating that the addition of the BN layer does not affect the results much.

A visual comparison between HighResNet and U-Net is shown in Fig. 5, with U-Net+BN left out, because of the similarity of its results to U-Net. Because each network was trained five times, the images were generated using the network from the run that achieved the median metric value most often. This is also the case for the rest of the qualitative results.

The two columns on the right show the Relative Error (RE) and Structural Dissimilarity Index Measure (dSSIM), which is simply 1 – SSIM. Notice that areas in the RE and dSSIM maps that light up in one network often also light up in the other network's output, indicating that most of these areas are difficult for both networks. An example of this is the area indicated by the arrow in the top right of the sTSEs, where both HighResNet and U-Net struggle. This area is actually hyperintense in the GE images, but dark in the TSE images. HighResNet correctly predicts the lower intensity, but overestimates the extent of this darker area. U-Net on the other hand fails to predict the lower intensity.

Another important aspect of the images is the area indicated by the arrow on the top left, which shows an artifact related to the image registration. As mentioned before, when the GE was pushed inward near the image edges, the "empty" space was filled with an image intensity of -1, which is subsequently set to 0 during pre-processing. This thresholding will sometimes leave a dark strip/area near the image edges, which is a feature the network does not recognize from the training data, resulting in artifacts.

A clear difference between the two networks is pointed out by the lowest arrow, where a thin fatty structure is almost absent on U-Net's output and much more pronounced in HighResNet's output. The arrow also points at a dark interface between the muscle and fat, which is also much more visible on HighResNet's output than on U-Net's.

Finally, the dSSIM maps show a slightly higher intensity in the centre of most of the upper vertebrae for both networks. This demonstrates the fact that the networks tend to create more smooth images, thereby reducing the detail and speckle present inside the vertebrae.

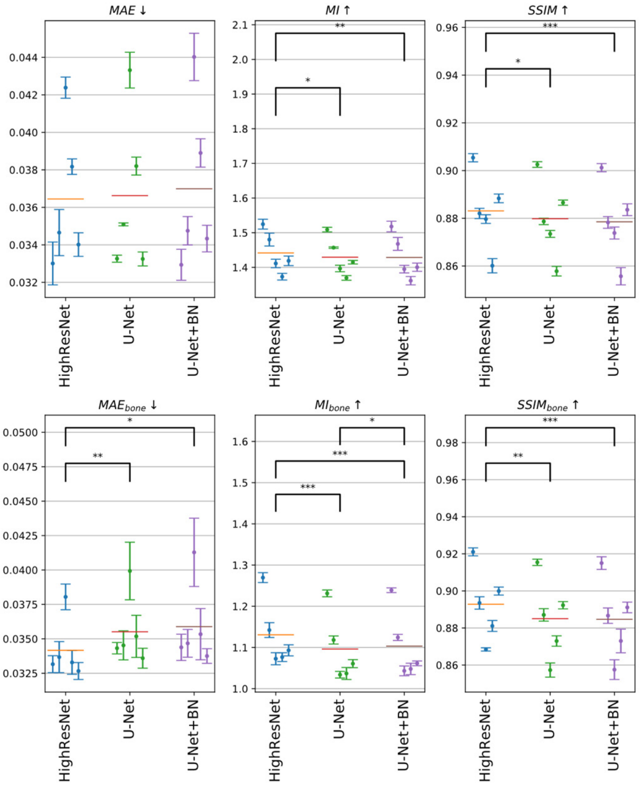


Fig. 4. Results of training all three networks five times using the Low-to-low approach. Each datapoint represents a patient with the standard deviation over five repetitions. Horizontal bars indicate the mean over all patients and repetitions. All metrics were evaluated inside the entire body mask (no subscript) and inside the bone mask (bone subscript).

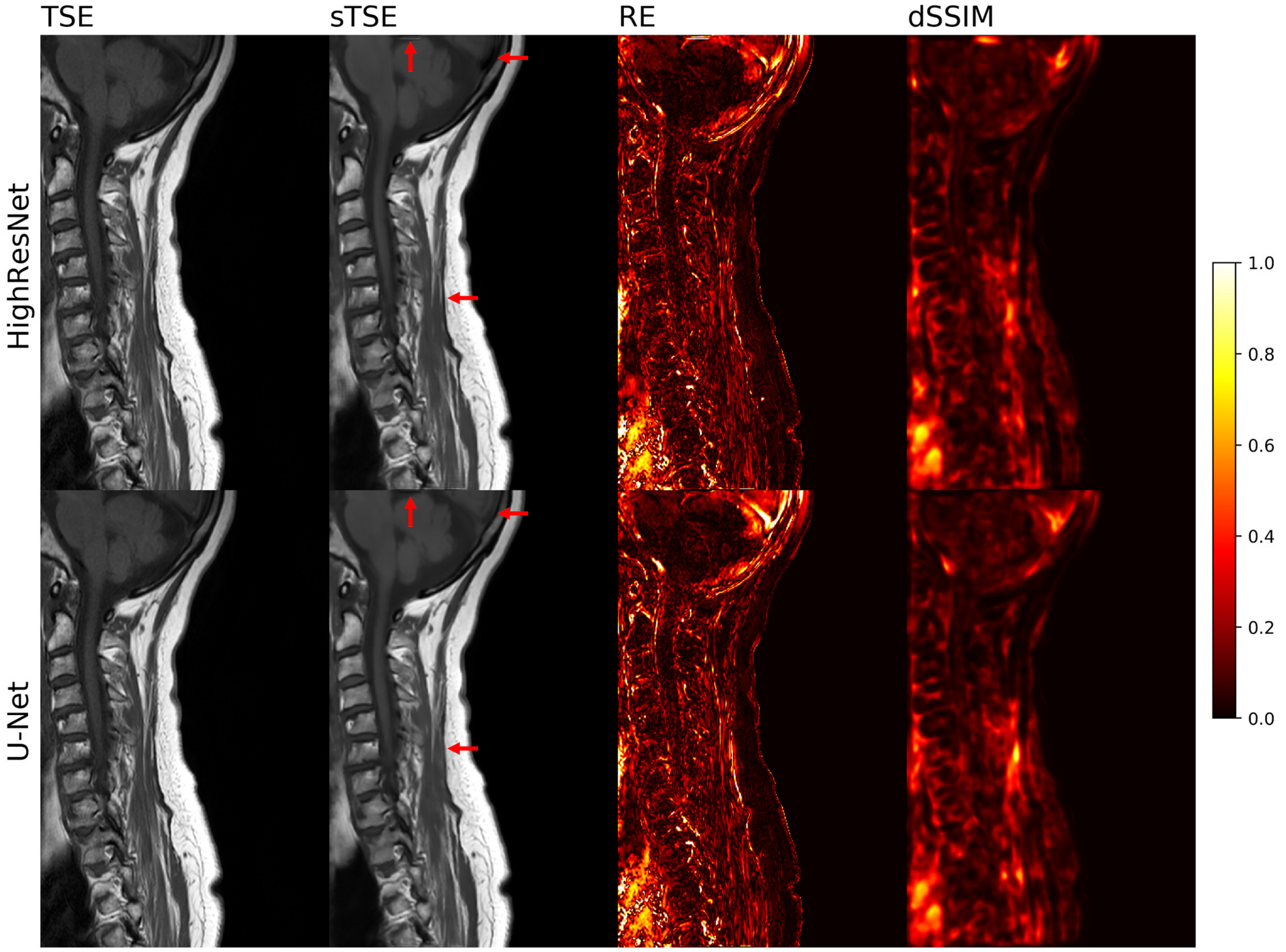


Fig. 5. Qualitative results from HighResNet and U-Net trained using the Low-to-low approach. The ground-truth TSE and predicted sTSE are shown along with the Relative Error (RE) and Structural Dissimilarity Index Measure (dSSIM), where the latter two are masked by the body mask. Each of the image slices (excluding the RE and dSSIM maps) is individually normalized by clipping the intensities to the 1st- and 99th-percentile and subsequently mapping to the interval 0-1.

III. B. High-to-low approach

Both HighResNet and U-Net were trained using the High-to-low approach, with U-Net+BN left out because of the unsignificant performance change compared to U-Net on five of the six evaluation metrics. The results of the $sTSE_{LR}$ to TSE comparison from five repetitions are summarized in Table IV. Downsampling was performed using a fixed kernel with weights [1/2, 1/2, 0, 0]. The table clearly shows that (1) the two networks perform similarly when using only the L1-loss, except for a significant difference in the MAE. (2) Using the MIND loss, U-Net actually performs significantly better in five out of the six metrics. (3) The MIND term degrades performance on the (low resolution) TSE prediction.

Although U-Net outperforms HighResNet on one out of the six metrics using L1-only and on five out of the six metrics using MIND, this performance comes at a cost to the $sTSE_{HR}$ images. The cost is demonstrated in Fig. 6, where slice-dependent behaviour is clearly present in both image pairs, although the MIND term of the loss regularizes the network and slightly prevents it.

Because of these artifacts, which make the $sTSE_{HR}$ hard to use for clinical interpretation, the rest of the experiments were only performed using HighResNet.

TABLE IV

EVALUATION METRICS FOR NETWORKS TRAINED USING THE HIGH-TO-LOW APPROACH WITH TWO DIFFERENT LOSS FUNCTIONS. SIGNIFICANTLY BETTER CONFIGURATIONS ARE INDICATED IN BOLD PER LOSS FUNCTION AND METRIC.

Loss	Network	MAE	MAE$_{bone}$	MI	MI$_{bone}$	SSIM	SSIM$_{bone}$
L1	HighResNet	0.037 ± 0.005	0.033 ± 0.002	1.473 ± 0.070	1.168 ± 0.081	0.888 ± 0.016	0.902 ± 0.017
	U-Net	**0.035 ± 0.004**	0.033 ± 0.003	1.482 ± 0.059	1.163 ± 0.078	0.892 ± 0.015	0.902 ± 0.018
L1 + 0.1×MIND	HighResNet	0.038 ± 0.005	0.035 ± 0.002	1.431 ± 0.066	1.131 ± 0.081	0.873 ± 0.017	0.888 ± 0.019
	U-Net	**0.035 ± 0.004**	**0.034 ± 0.002**	**1.461 ± 0.059**	1.147 ± 0.080	**0.882 ± 0.015**	**0.894 ± 0.019**

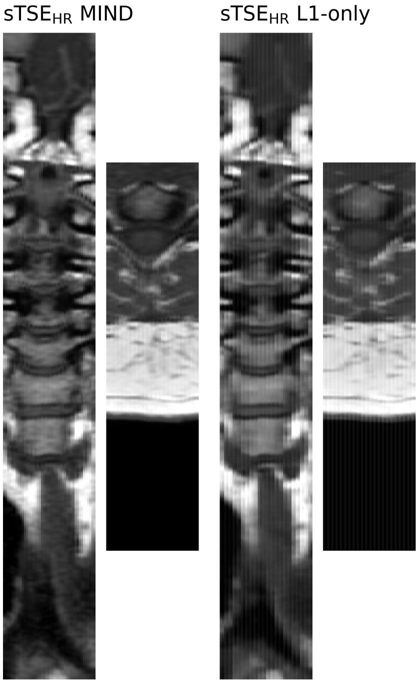


Fig. 6. Coronal and axial slices from training U-Net with MIND and with L1-only using the High-to-low approach. Both image pairs show strong slice-dependent behaviour, although it is suppressed by the MIND term.

1) Comparison to the Low-to-low approach

To see if the extra information from the higher resolution input in the High-to-low approach helps performance on the contrast transformation, HighResNet was retrained with all hyperparameters and configurations matched to the High-to-low approach. Fig. 7 shows the comparison between the two methods, in which the downsampling kernel in the High-to-low approach has been trained with the network weights frozen. The resultant kernel weights are shown in Table A1.

Fig. 7 demonstrates that (1) the High-to-low approach significantly outperforms the Low-to-low approach in four out of the six metrics when used with L1-only. (2) The addition of the MIND term significantly degrades the High-to-low performance to below that of the Low-to-low approach in terms of the SSIM in the entire body and bone.

Not only does the High-to-low approach increase performance (when using L1-only) on the TSE synthesis because of the better utilization of the available data, it also creates a synthetic TSE at a higher resolution than the ground-truth TSE. These are shown along with the aIP GE (which is used for the MIND loss computation and is part of the input images) and the ground-truth TSE in Fig. 8.

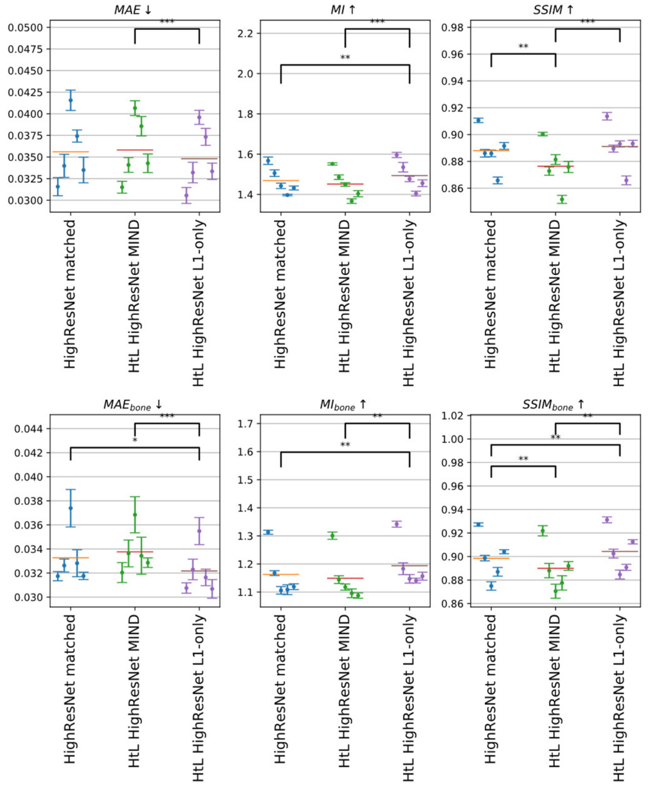


Fig. 7. Results of training HighResNet with the Low-to-low approach matched to the High-to-low approach parameters vs. HighResNet trained using the High-to-low (abbreviated as HtL) approach. Each datapoint represents a patient with the standard deviation over five repetitions and horizontal bars indicate the mean over all patients and repetitions.

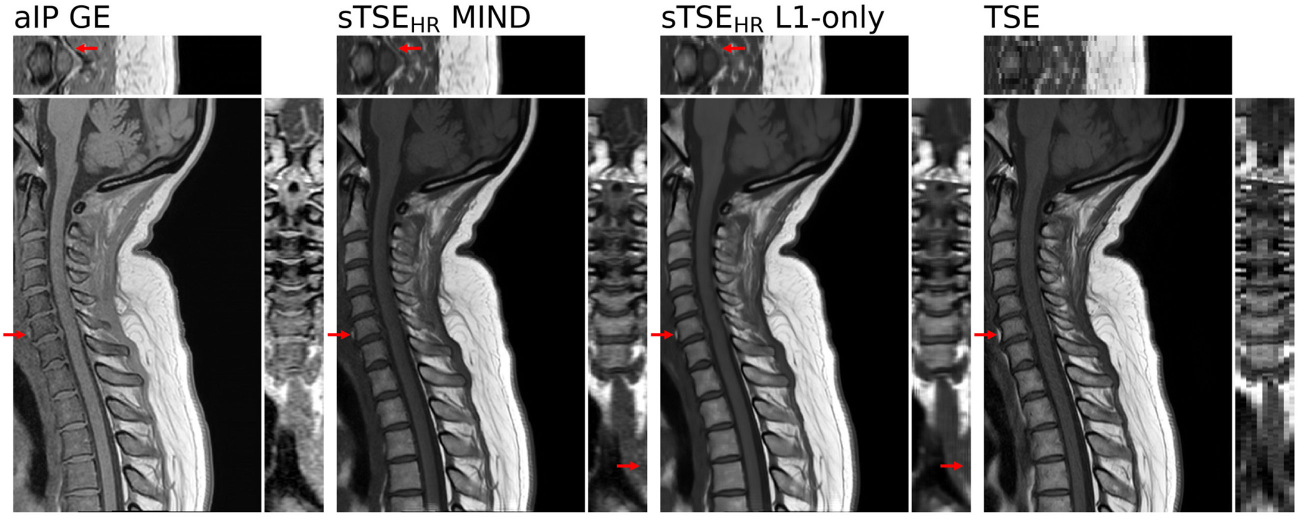


Fig. 8. Comparison between the two High-to-low approach variants using HighResNet with L1-only and with MIND. The input aIP GE is also shown along with the ground-truth TSE, where the latter is of a four times lower resolution in the left-right direction as evident in the coronal and axial slices. Each slice is individually normalized by clipping the intensities to the 1st- and 99th-percentile and subsequently mapping to the interval 0-1.

Starting with the axial slices, we clearly see the effect of the MIND term in helping with maintaining some of the structure present in the aIP GE, where the red arrow points at a V-shaped muscle structure bordering the spinal cord, which is much better represented in the MIND variant than in the L1-only variant. This same mechanism can also lead to negative effects, however, which is shown in the sagittal slices with the arrows. Here, the fat anterior to the vertebra is hard to distinguish in the aIP GE, which propagates into the sTSE$_{HR}$ when using the MIND term. Using L1-only, the contrast and shape correspond better to the actual TSE. Finally, the arrow on the coronal slice shows that even with the random offsets and flipping augmentation, the L1-only variant still exhibits some slice-dependent behaviour. Again, this behaviour is much less prominent/almost absent when using MIND.

The results of the other four folds are shown in the appendix in Fig. A1 for comparison, where it can be seen that the severeness of this slice-dependent behaviour varies across the runs. It also shows two images containing incorrectly predicted hyperintensities in the posterior subcutaneous fat for both the MIND and L1-only variants.

Going back to the slice-dependent behaviour, there is a simple solution to this problem: going from a stride of (18, 18, 4) to (18, 18, 1) in the sTSE$_{HR}$ reconstruction. The smaller stride in the left-right direction smears out the artifact, making it less visible. Examples of this technique are shown in Fig. A2 in the appendix for the same folds as in Fig. 8 for all test set patients, with the second row corresponding to the same patient as in Fig. 8. Note that although this stride increases the quality of the sTSE$_{HR}$ images, it masks an underlying issue in the reconstruction and does not allow for a fair comparison to the Low-to-low approach because of the smaller stride, which would correspond to an undefined stride of (18, 18, 1/4) for the sTSE$_{LR}$ reconstruction.

2) Validation of the MIND loss

To validate the use of the MIND loss, the TSE images were downsampled with a factor four in the superior-inferior direction to have a ground-truth TSE$_{HR}$ image for evaluation and a TSE$_{LR}$ image to train on. The results of the comparison between the TSE$_{HR}$ and sTSE$_{HR}$ (note that the former is simply the original TSE) are summarized in Table V.

Table V shows what is also reflected in the earlier results in regard to the MIND variant, which is that it underperforms on the TSE synthesis. Not only does this apply to the sTSE$_{LR}$, but also to the sTSE$_{HR}$ as shown here, which is to be expected.

TABLE V
COMPARISON BETWEEN THE TWO LOSS FUNCTIONS FOR HIGHRESNET TRAINED USING THE HIGH-TO-LOW APPROACH. SIGNIFICANTLY BETTER CONFIGURATIONS ARE INDICATED IN BOLD PER LOSS FUNCTION

	MAE	MI	SSIM
L1-only	**0.038 ± 0.004**	**1.419 ± 0.055**	**0.876 ± 0.015**
MIND	0.038 ± 0.004	1.397 ± 0.056	0.865 ± 0.015
	MAE$_{bone}$	**MI$_{bone}$**	**SSIM$_{bone}$**
L1-only	**0.037 ± 0.003**	**1.079 ± 0.074**	**0.879 ± 0.020**
MIND	0.037 ± 0.002	1.065 ± 0.078	0.873 ± 0.020

IV. DISCUSSION

IV. A. Experimental results

In this study, we investigated the possibility of translating from GE to TSE images. A dataset of cervical spine data was used to train a well-established network architecture, U-Net, and a network architecture designed specifically for the GE to TSE translation, HighResNet. Additionally, a new approach was implemented to keep the higher through-plane resolution offered by the GE images.

The results demonstrate that HighResNet outperforms U-Net and U-Net+BN when using the Low-to-low approach (Fig. 4). This confirms our hypothesis that the use of residual connections and/or the lack of downsampling of the feature maps (e.g. by max-pooling) boosts performance for the GE to TSE translation. Although U-Net is used in much of the image synthesis research due to its good performance and flexibility [9], [12], [18], [19], we have shown that a specifically tailored network architecture like HighResNet might be better in some cases.

As mentioned in section II. C. 1), U-Net is often used inside a GAN. The reason GANs were not used in this study is that although the adversarial loss helps with creating more visually appealing results, they can also hallucinate features that are not in the actual source data [25]. An example could be a GAN that has only been trained on image data from healthy patients, which subsequently removes a tumour from an image during inference, because the network has been trained to match the healthy training set distribution [26]. This type of behaviour is of course unwanted when dealing with medical images that are meant for clinical interpretation.

The hypothesis that HighResNet's architecture allows it to outperform U-Net does not hold true in the High-to-low approach, where U-Net and HighResNet perform similarly with an L1-only loss. Using the MIND term U-Net even outperforms HighResNet significantly on a majority of the similarity metrics (Table IV).

An explanation of this could be that it actually seems advantageous for the networks to create the slice-dependent behaviour to optimize for performance on the sTSE$_{LR}$ synthesis. This is why even with the random offsets and flipping of the training data to combat this behaviour, it is still present in the sTSE$_{HR}$ for both loss variants of U-Net and for the L1-only variant of HighResNet (Fig. 6 and Fig. 8). The behaviour is much better supressed for HighResNet on the other hand, which might actually hinder its performance on the sTSE$_{LR}$, but it creates qualitatively better sTSE$_{HR}$'s than U-Net.

For both networks the benefit of using the High-to-low approach are two-fold: (1) the performance on the TSE synthesis is improved when using the L1-only loss as compared to the Low-to-low approach (Fig. 7) and (2) it results in a fourfold increase in the amount of sagittal slices as compared to the ground-truth TSE (Fig. 8). The improvement in performance on the TSE synthesis shows indirectly that, although there is no ground-truth high resolution TSE to compare the sTSE$_{HR}$ against, it still contains reliable information. Additionally, the fact that two of the four GE slices are more similar to the corresponding TSE slice shows that the GE contains more information than is present in the TSE image. This same principle applies to the sTSE$_{HR}$, which is also reflected in the trained weights of the downsampling kernel, where these two slices are weighted ~5-6× as much as one of their neighbouring voxels (Table A1).

The performance boost of the High-to-low approach compared to the Low-to-low approach is negated when using the MIND term in the loss function and results in a significant decrease in performance compared to the Low-to-low approach in terms of the SSIM inside the body and the bone (Fig. 7). This performance decrease is related to the fact that the GE images are generally noisier, which is smoothed out when using only the L1 loss, but is preserved more when adding the MIND loss, because the MIND loss increases the similarity to the aIP GE. The decrease in performance in relation to the L1-only variant would most likely also be present in the sTSE$_{HR}$ if a ground-truth image were available (Table V).

Although the MIND loss decreases the correspondence between the sTSE and TSE, it does give more realistic looking images by keeping some of the noise and detail present in the GE.

IV. B. Limitations

The reason of the performance decrease from the MIND loss is tied to one of the main limitations of the GE to TSE translation, namely that some of the features present in the TSE are completely absent and therefore unlearnable in the GE. This is most clearly visible in Fig. 8, where the thin fatty structures anterior to the vertebra with the red arrow and anterior to the three vertebrae below it in the TSE are almost completely absent from both sTSE$_{HR}$'s. Note that although four of the sTSE$_{HR}$ slices correspond to the one TSE slice and only one of these is shown, this feature is absent among all four of them.

The missing fatty structures are a direct result of this feature being mostly absent in the aIP GE (and also in the aOP GE, not shown in Fig. 8). In some cases the L1-only network might be able to reconstruct features like these based on the aOP GE image as these sometimes do vaguely show features absent in the aIP GE image. The network with the MIND loss term is forced to keep its sTSE images close to the aIP GE image, however, which further exacerbates this mismatch in features.

Another example of this feature mismatch is the dark edge between the intervertebral disks and the vertebrae (discovertebral junction), which extends out further on the TSE than on the aIP GE by about one to two voxels. The MIND term strongly enforces the edges of the sTSE$_{HR}$ and aIP GE to be aligned, so this means that in the sTSE$_{HR}$ these dark edges are in the wrong location. The network trained with L1-only on the other hand is able to work around this by learning to recognize these edges and slightly increasing their size, which is reflected by the more defined and thicker dark edges in this variant (Fig. 8).

Although the L1-only network is better able to reconstruct some of the TSE features, the fact still remains that some features are simply absent from the GE images, making them impossible to reconstruct for all networks and for both the Low-to-low and High-to-low approach.

Finally, the reason the MIND loss was applied to the Sobel filtered image patches instead of to the patches directly is that the former does not allow for mismatches between the edges of the aIP GE and sTSE$_{HR}$ as much. This edge mismatch was an issue when applying the MIND loss directly to the image patches, where the L1 loss would force the edges between e.g. the intervertebral disk and vertebra to be in the same location as in the TSE, but the MIND loss would force it to be in a different location as in the GE, giving an artificial double edge. Applying the MIND loss to the Sobel features instead fixes this issue, but it simply masks the underlying problem of the GE and TSE feature mismatch mentioned before.

The SSIM and MI between the aIP GE and sTSE$_{HR}$ were also used as a structural consistency loss, but when applied directly to patches extracted from both of these images (so without the Sobel filters), both losses created artifacts. The MI tended to create images with piecewise-linear intensity profiles with very sharp edges between e.g. fat and muscle, which should not be present. We hypothesized that this behaviour originated from the finite number of histogram bins (32 or 64) used, which seemed to force the network to push each voxel intensity to the centre of one of these bins, creating a more sparse histogram and therefore less realistic images.

The SSIM presented some artifacts as well, because it is maximized when both patches are exactly the same [24]. In practice, this meant that the SSIM term pulled the synthetic TSE images more towards the original GE aIP contrast, thereby hindering the contrast transformation.

IV. C. Future research

The sTSE in its current state lacks clinically important features due to some of the limitations outlined in the previous section, so we do not recommend the sTSE($_{LR}$)'s be used in their current form clinically. An example of these are the small fatty rims along the vertebrae, which are mostly absent from the GE images and therefore also from the sTSE.

For the complete removal of the TSE from the protocol, the quality of the sTSEs needs to be improved, which would require changes to the BoneMRI sequence, because it simply does not represent some of the features needed for the TSE reconstruction. The most important change would be an increase in the contrast between muscle and fat. This would require changes to the repetition time and/or flip angle, but that could in turn influence the performance of the sCT synthesis and might force the sCT network to be retrained.

Therefore, if this research direction were explored further, the sTSEs resultant from the altered sequence would need to be thoroughly clinically validated along with the sCT's generated from this data using a retrained network, while taking total scan time into account.

The quality of the sTSEs can also be slightly improved by changing the pre-processing of the data. In both the Low-to-low- and High-to-low approaches the in-plane resolution of the GE was upsampled to 0.488 mm × 0.488 mm, which is an 1.5× increase in the number of voxels in both in-plane directions, created from interpolation. This creates voxels that do not carry any real information, so the network is not just tasked with contrast transformation, but also with learning to transform an image, where roughly 57% of the voxels come from interpolation (ignoring the through-plane/left-right resampling), to a non-interpolated image.

A better approach would therefore be not to resample the GE in the in-plane orientation, but to resample the TSE to the in-plane GE resolution of 0.744 mm × 0.744 mm or, better yet, to acquire the TSE with the same in-plane resolution to maximize the correspondence between the two and remove the influence of interpolation artifacts.

A TSE with the same in-plane resolution would also alleviate some of the feature mismatch discussed in the previous section. For example, the mismatch between the dark edges between the vertebrae and intervertebral disks is partly determined by partial volume effects in the GE, which are exacerbated by the resampling in the in-plane direction.

Finally, the quality of both the sTSE$_{LR}$ and sTSE$_{HR}$ from the High-to-low approach can be improved by simply reducing the stride in the left-right direction to 1 as discussed in the results and demonstrated for the sTSE$_{HR}$ in Fig. A2. Although this is a simple change for the sTSE$_{HR}$ generation, it requires a rework of the creation of the sTSE$_{LR}$, because the sTSE$_{HR}$ is being shifted by 1/4th of the sTSE$_{LR}$ voxel size with each step in the left-right direction. This means that for each of the four configurations k, k+1, k+2, k+3, as shown in Fig. 9, a different downsampling kernel has to be used.

To determine the kernel weights, four different downsampling kernels could be trained for each of the four different sets of patches with the network parameters frozen. Note, however, that the top-most slice out of the four (the left-most in the actual image) is used in the sTSE$_{LR}$ reconstruction four times, whilst the one below it is only used thrice, biasing the prediction towards the top-most of these two slices. It might therefore be better to leave out every fourth patch for the sTSE$_{LR}$ reconstruction, corresponding to Patch k+3.

A less computationally expensive method than training a kernel for each of the four (or three, ignoring the fourth configuration to remove the bias) configurations could be to train the network on patches in the Patch k+1 configuration, resulting in weights [a, b, c, d], which could then be transferred to Patch k as [b, c, d, 0] and to Patch k+2 as [0, a, b, c], where the last two would have to be normalized.

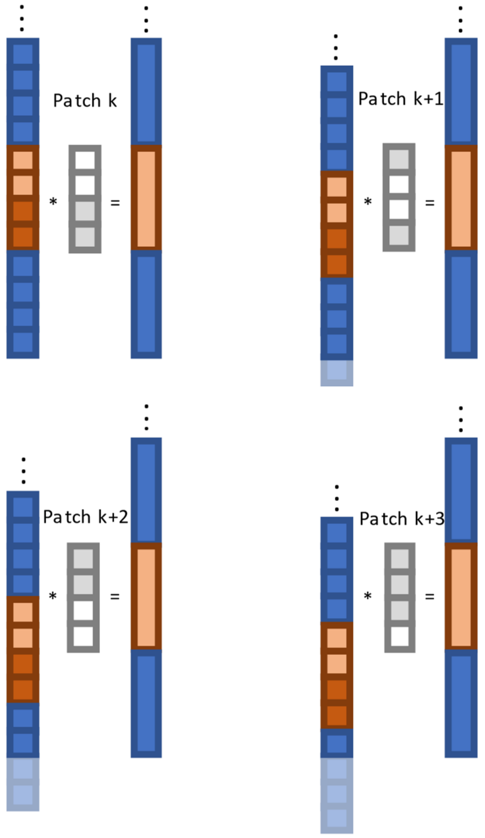


Fig. 9. Schematic overview of the convolutional downsampling from the sTSE$_{HR}$ to the sTSE$_{LR}$ patches with a stride of one. The orange sTSE$_{HR}$ slices that correspond most to the orange sTSE$_{LR}$ slice are highlighted along with the corresponding kernel locations.

V. Conclusion

We studied the synthesis of TSE images from GE images and proposed the High-to-low approach. The approach was found to be capable of creating significantly higher quality synthetic TSE images as compared to the Low-to-low approach, for which the input GE images were downsampled to the TSE resolution. Furthermore, the High-to-low approach was capable of generating synthetic TSE images at a higher resolution than the ground-truth data. Finally, we introduced a fully convolutional network architecture, HighResNet, which presented fewer artifacts in the high resolution synthetic TSE images than a conventional U-Net and outperformed the U-Net quantitatively in the Low-to-low approach.

I. Acknowledgment

I would like to thank my supervisors, dr. ir. Peter R. Seevinck and dr. Marijn van Stralen for all of their helpful feedback on my work and insights during meetings and for keeping the meetings fun.

I would also like to acknowledge dr. Mateusz C. Florkow for providing a lot of helpful code as a starting point for my project, teaching me the basics of helpful tools like MeVisLab and always being willing to help with small questions that would come up during the day.

References

[1] D. Ma *et al.*, "Magnetic resonance fingerprinting," *Nat. 2013 4957440*, vol. 495, no. 7440, pp. 187–192, Mar. 2013, doi: 10.1038/nature11971.

[2] O. van der Heide, A. Sbrizzi, P. R. Luijten, and C. A. T. van den Berg, "High-resolution in vivo MR-STAT using a matrix-free and parallelized reconstruction algorithm," *NMR Biomed.*, vol. 33, no. 4, p. e4251, Apr. 2020, doi: 10.1002/NBM.4251.

[3] J. B. M. Warntjes, O. Dahlqvist Leinhard, J. West, and P. Lundberg, "Rapid magnetic resonance quantification on the brain: Optimization for clinical usage," *Magn. Reson. Med.*, vol. 60, no. 2, pp. 320–329, Aug. 2008, doi: 10.1002/MRM.21635.

[4] A. Hagiwara *et al.*, "SyMRI of the Brain: Rapid Quantification of Relaxation Rates and Proton Density, With Synthetic MRI, Automatic Brain Segmentation, and Myelin Measurement," *Invest. Radiol.*, vol. 52, no. 10, p. 647, Oct. 2017, doi: 10.1097/RLI.0000000000000365.

[5] S. U. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image Synthesis in Multi-Contrast MRI with Conditional Generative Adversarial Networks," *IEEE Trans. Med. Imaging*, vol. 38, no. 10, pp. 2375–2388, Feb. 2018, Accessed: Aug. 30, 2021. [Online]. Available: https://arxiv.org/abs/1802.01221v1.

[6] P. Welander, S. Karlsson, and A. Eklund, "Generative Adversarial Networks for Image-to-Image Translation on Multi-Contrast MR Images - A Comparison of CycleGAN and UNIT," Jun. 2018, Accessed: Sep. 05, 2021. [Online]. Available: https://arxiv.org/abs/1806.07777v1.

[7] Q. Yang, N. Li, Z. Zhao, X. Fan, E. I.-C. Chang, and Y. Xu, "MRI Cross-Modality NeuroImage-to-NeuroImage Translation," Jan. 2018, Accessed: Sep. 05, 2021. [Online]. Available: https://arxiv.org/abs/1801.06940v2.

[8] H. Yang, J. Sun, L. Yang, and Z. Xu, "A Unified Hyper-GAN Model for Unpaired Multi-contrast MR Image Translation," pp. 127–137, Sep. 2021, doi: 10.1007/978-3-030-87199-4_12.

[9] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-GANs: Edge-Aware Generative Adversarial Networks for Cross-Modality MR Image Synthesis," *IEEE Trans. Med. Imaging*, vol. 38, no. 7, pp. 1750–1762, Jul. 2019, doi: 10.1109/TMI.2019.2895894.

[10] A. M. Dinkla *et al.*, "MR-Only Brain Radiation Therapy: Dosimetric Evaluation of Synthetic CTs Generated by a Dilated Convolutional Neural Network," *Int. J. Radiat. Oncol.*, vol. 102, no. 4, pp. 801–812, Nov. 2018, doi: 10.1016/J.IJROBP.2018.05.058.

[11] M. C. Florkow *et al.*, "Deep learning–based MR-to-CT synthesis: The influence of varying gradient echo–based MR images as input channels," *Magn. Reson. Med.*, vol. 83, no. 4, p. 1429, Apr. 2020, doi: 10.1002/MRM.28008.

[12] Y. Lei *et al.*, "MRI-Only Based Synthetic CT Generation Using Dense Cycle ConsistentGenerative Adversarial Networks," *Med. Phys.*, vol. 46, no. 8, p. 3565, Aug. 2019, doi: 10.1002/MP.13617.

[13] M. C. Florkow, K. Willemsen, V. V. Mascarenhas, E. H. G. Oei, M. van Stralen, and P. R. Seevinck, "Magnetic Resonance Imaging Versus Computed Tomography for Three-Dimensional Bone Imaging of Musculoskeletal Pathologies: A Review," *J. Magn. Reson. Imaging*, 2022, doi: 10.1002/JMRI.28067.

[14] B. E. Dewey *et al.*, "Deep Harmonization of Inconsistent MR Data for Consistent Volume Segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11037 LNCS, pp. 20–30, Sep. 2018, doi: 10.1007/978-3-030-00536-8_3.

[15] E. Lin and A. Alessio, "What are the basic concepts of temporal, contrast, and spatial resolution in cardiac CT?," *J. Cardiovasc. Comput. Tomogr.*, vol. 3, no. 6, p. 403, Nov. 2009, doi: 10.1016/J.JCCT.2009.07.003.

[16] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imaging*, vol. 29, no. 1, pp. 196–205, Jan. 2010, doi: 10.1109/TMI.2009.2035616.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, May 2015, vol. 9351, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[18] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, "Deep learning based synthetic-CT generation in radiotherapy and PET: A review," *Med. Phys.*, vol. 48, no. 11, pp. 6537–6566, Nov. 2021, doi: 10.1002/MP.15150.

[19] J. Chen, J. Wei, and R. Li, "TarGAN: Target-Aware Generative Adversarial Networks for Multi-modality Medical Image Translation," pp. 24–33, Sep. 2021, doi: 10.1007/978-3-030-87231-1_3.

[20] W. Li, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10265 LNCS, pp. 348–360, Jul. 2017, doi: 10.1007/978-3-319-59050-9_28.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.

[22] O. S. Kayhan and J. C. van Gemert, "On Translation Invariance in CNNs: Convolutional Layers can Exploit Absolute Spatial Location," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 14262–14273, Mar. 2020, doi: 10.1109/CVPR42600.2020.01428.

[23] M. P. Heinrich *et al.*, "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Med. Image Anal.*, vol. 16, no. 7, pp. 1423–1435, Oct. 2012, doi: 10.1016/J.MEDIA.2012.05.008.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[25] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, p. 101552, Dec. 2019, doi: 10.1016/J.MEDIA.2019.101552.

[26] J. P. Cohen, M. Luck, and S. Honari, "Distribution Matching Losses Can Hallucinate Features in Medical Image Translation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11070 LNCS, pp. 529–536, Sep. 2018, doi: 10.1007/978-3-030-00928-1_60.

Appendix

Table A1
Average (± standard deviation) weights of the downsampling kernels after training for all five repetitions of the High-to-low runs

Index→ **Loss ↓**	**0**	**1**	**2**	**3**
L1-only	0.4436 ± 0.0585	0.4570 ± 0.0318	0.0790 ± 0.0672	0.0189 ± 0.0390
MIND	0.4359 ± 0.0973	0.4648 ± 0.0365	0.0810 ± 0.0755	0.0052 ± 0.0283

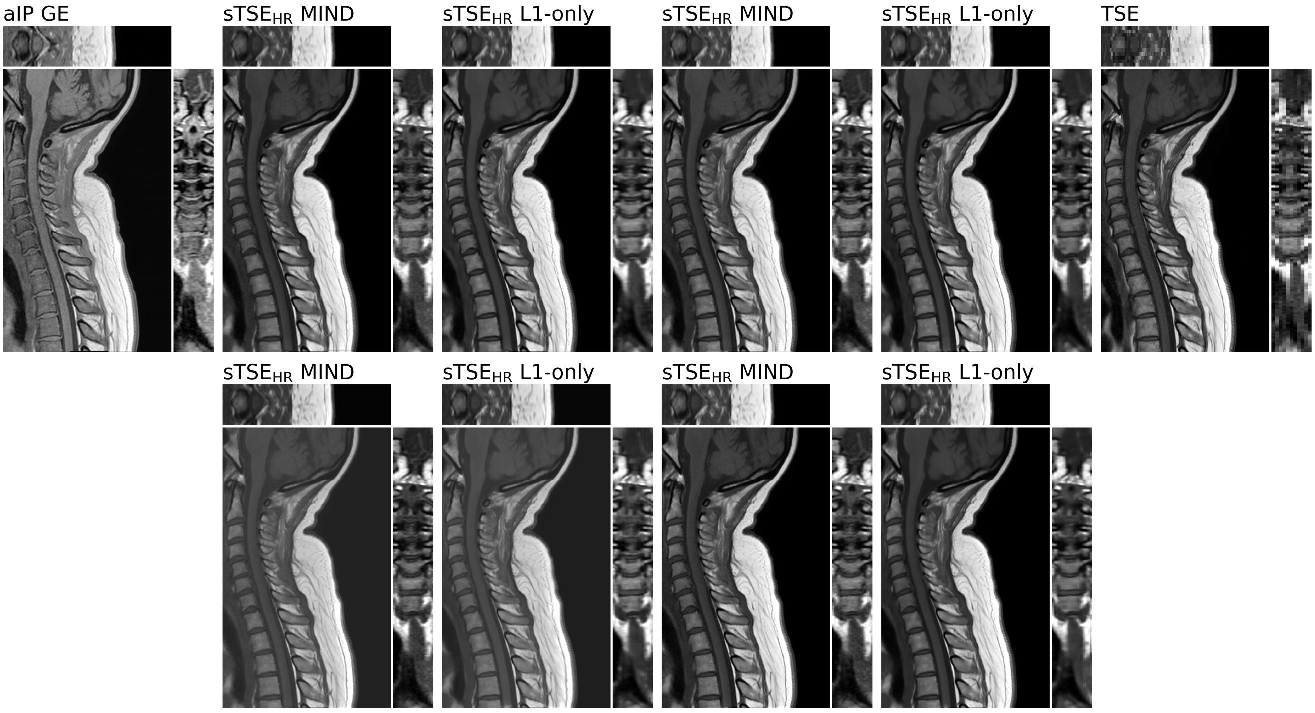
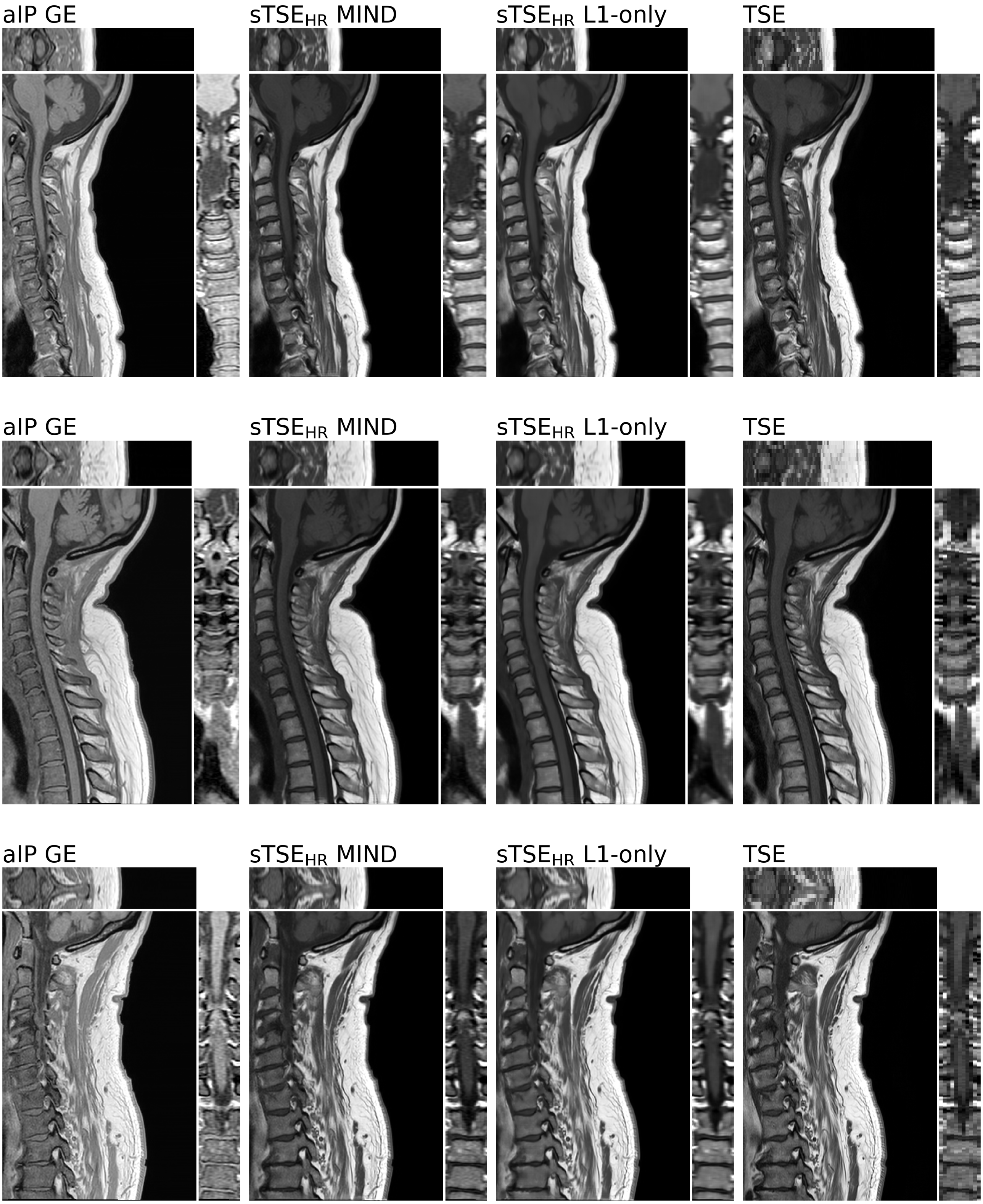


Fig. A1. Comparison between the two High-to-low approaches with L1-only and with MIND, showing the other four out of the five repetitions missing in Fig. 8.


aIP GE
sTSE$_{HR}$ MIND
sTSE$_{HR}$ L1-only
TSE
aIP GE
sTSE$_{HR}$ MIND
sTSE$_{HR}$ L1-only
TSE
aIP GE
sTSE$_{HR}$ MIND
sTSE$_{HR}$ L1-only
TSE

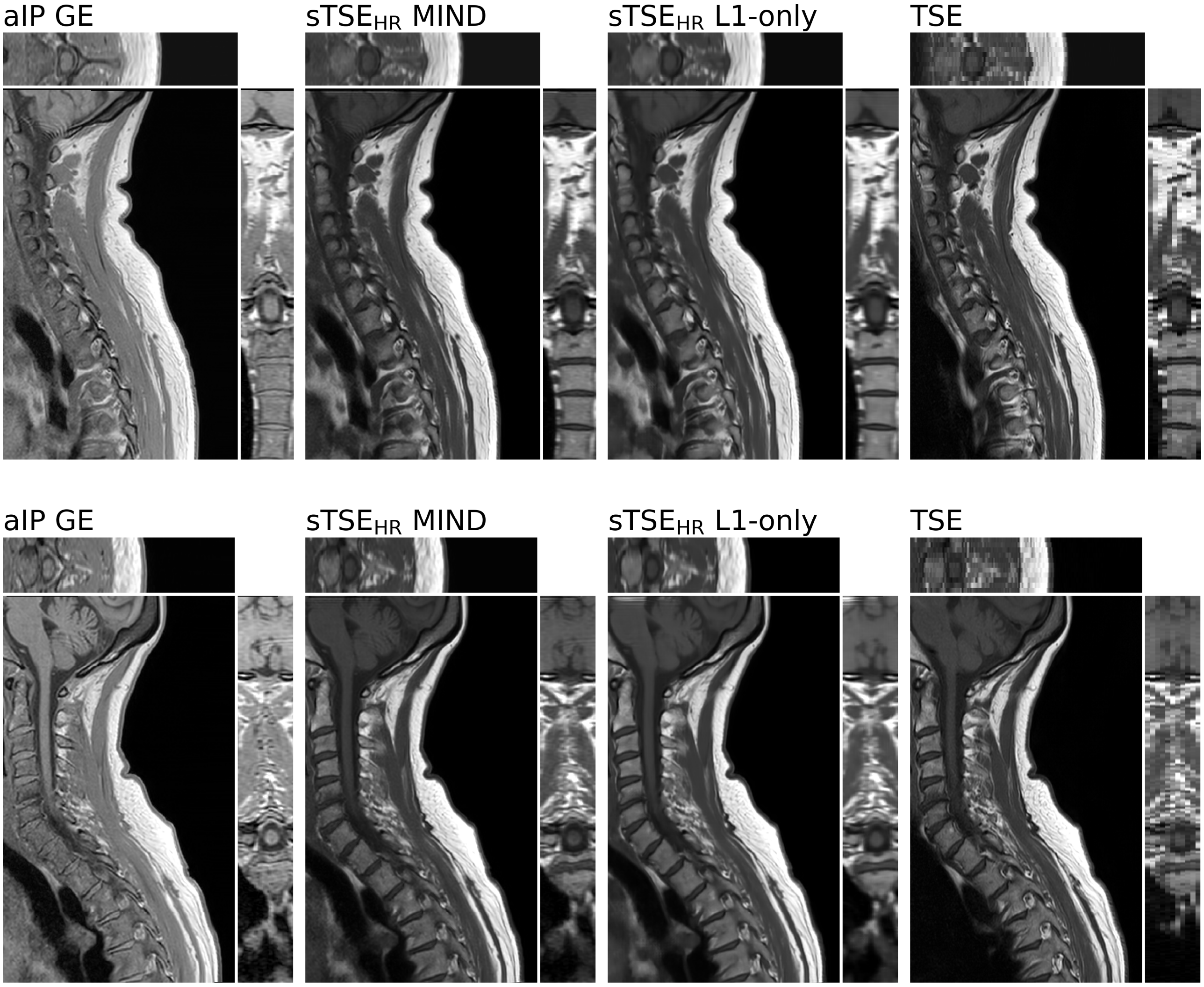


Fig. A2. Comparison between the two High-to-low approaches with L1-only and with MIND. All sTSE$_{HR}$ images are reconstructed with a stride of (18, 18, 1) instead of a stride of (18, 18, 4) as used in Fig. 8, which shows the same patient as in the second row of this figure. Note that the slice-dependent behaviour is much less pronounced than in Fig. 8, which is especially apparent at the bottom of the coronal sTSE$_{HR}$ L1-only slice.