Investigation of the influence of MR physics-based versus random intensity-based data synthesis methods for a generalizable spine segmentation network

LAYMAN'S SUMMARY

Paula Arias Martínez

Spine segmentation is the process of delineating the shapes of the vertebrae in medical images, creating a new image that indicates the exact location of the vertebrae. Particularly, Magnetic Resonance Imaging (MRI) is a highly suitable technique for the assessment of many spinal conditions (e.g., scoliosis, fractures or metastases) where spine segmentation is a valuable task, since it is a radiation-free imaging modality that allows for the acquisition of three-dimensional (3D) scans at a high resolution.

Currently, automatic segmentation approaches based on deep learning are gaining more and more popularity. These approaches employ neural networks, which are algorithms that mimic the functioning of the human brain by recognizing relationships and patterns in the data. To train a neural network for segmentation, it must be provided with MR images of the spine and their corresponding segmentation, which is usually named 'ground truth'. During several training loops, the network learns complex features in the data so that, when presented with new MR images, it should be able to segment the structure of interest.

Neural networks frequently face problems to generalize or, in other words, to perform well in new types of data. Particularly in this context, networks usually have problems to adapt to the vast variety of contrasts that MR images present and to the different patient populations. To attain generalization, deep learning models should be trained with abundant and diverse data, but these extensive datasets are often difficult to obtain in the clinical setting. One approach that can address this issue and facilitate generalization is data augmentation, a process that increases the dataset variation by creating new samples from the available ones. In this work, two data augmentation approaches were developed and compared after being applied in the training of a spine segmentation network.

The first one, named SynthMRI, consisted of the generation of synthetic scans with contrasts that simulate those encountered in real MR images by utilizing quantitative MRI (qMRI) maps. qMRI is a common approach to quantify physical properties of the tissues with MRI. In conventional MR images, these properties collectively influence the MR signal. This signal is characterized by a signal equation, which varies depending on the type of MR contrast used and contains parameters that reflect these physical properties along with time parameters employed in the image acquisition. By putting the values of the qMRI maps into this equation, while modifying the time parameters over a specific range, diverse synthetic images resembling real MR scans can be created, with varying tissue intensities and contrasts.

The second augmentation approach, named SynthSeg, consisted of the generation of synthetic images with random contrasts and intensities from anatomical label maps. These anatomical label maps have information from the real images and segmentation, but are used to create synthetic images unlike any real images. By using these label maps and varying the contrasts considerably beyond the real images, the idea is that the neural network learns to only look at the shapes inside the images and not the intensities

Results demonstrated that spine segmentation networks trained with both SynthMRI and SynthSeg exhibited good generalization capabilities when presented with data featuring new MR contrasts and spine conditions despite the limited initial dataset, comprised of images from only five healthy individuals. Statistical analysis revealed no significant differences in performance between the two augmentation strategies overall. However, splitting these results by the type of MR image showed that SynthSeg performed worse in a particular type of scan. This was due to methodological adjustments made to ensure a fair comparison between both methods, which prevented them from being fully optimized. Despite these issues, the primary conclusion of this study was the great potential of both SynthMRI and SynthSeg for achieving generalization in spine segmentation tasks.

Investigation of the influence of MR physics-based versus random intensity-based data synthesis methods for a generalizable spine segmentation network

Paula Arias Martínez
Master's Programme in Medical Imaging
Utrecht University
Utrecht, Netherlands
p.ariasmartinez@students.uu.nl

***Abstract*—Recent advances in deep learning have greatly improved the automation of segmentation tasks. However, challenges remain in achieving robust performance on new domains, evidencing the need for large and diverse training datasets. In this study, two approaches, SynthMRI and SynthSeg, were implemented to generate new images during training, using available magnetic resonance (MR) scans, on a lumbar spine segmentation network. The main objective was to evaluate and compare their ability to generalize to unseen data. SynthMRI followed a physics-based approach that employed a set of quantitative MR images and Turbo Spin Echo (TSE) scans to synthesize new MR-like images with varied weightings using signal equations. In contrast, SynthSeg followed a domain randomisation strategy, where new images with random contrasts and intensities were generated from a set of anatomical label maps derived from TSE scans and the vertebrae segmentation by clustering image intensities. The evaluation of the predictions generated by the segmentation network trained with each approach revealed the ability of both SynthMRI and SynthSeg to generalize to images with unseen contrasts and patient populations. Specifically, SynthMRI achieved a mean Dice Similarity Coefficient (DSC) of 0.843 and a mean 95th percentile Hausdorff distance (HD95) of 3.712 mm, while SynthSeg obtained a mean DSC of 0.810 and a mean HD95 of 5.008 mm. Overall, no significant differences in performance were observed between the two methods. However, splitting the results by modality revealed that SynthMRI exhibited better performance than SynthSeg in TSE images. In conclusion, the outcomes of this study showed the great potential of both data synthesis strategies for achieving generalization in segmentation tasks.**

Keywords—deep learning, segmentation, convolutional neural networks, Magnetic Resonance Imaging, image synthesis

I. INTRODUCTION

Spine segmentation, or the precise delineation of the spinal anatomy, is a valuable task in several medical contexts and, particularly, in the assessment and management of deformities such as adolescent idiopathic scoliosis (AIS), where a thorough follow-up over time is usually needed for planning of treatment approaches. Moreover, an accurate spine segmentation can also aid in the evaluation of other spinal conditions, such as fractures, metastases, or degenerative diseases, highlighting its significance within the scope of musculoskeletal health and orthopaedics [1][2].

For all these cases, magnetic resonance imaging (MRI) can play a significant role by providing high-resolution images that enable an accurate visualization of the three-dimensional (3D) nature of many spinal conditions. Although MRI provides a rich variety of contrasts that reflect the underlying properties of the tissues, its inherent diversity poses a challenge for automated segmentation strategies. Consequently, manual delineation by trained individuals remains the gold standard, despite being time-consuming and exhibiting significant inter-expert variability [3].

Numerous modern methods for automated segmentation focus on the use of convolutional neural networks (CNNs), which are capable of learning and extracting hierarchical features from the images. However, optimizing the performance of these deep learning models requires the utilization of extensive and diverse datasets during training, which are especially hard to obtain in the clinical setting [4]. Within the context of this study, the key is to train a model that is able to generalize to diverse MR images. Generalization in this setting refers to a network still performing well on new types of data, in this case, different MR contrasts or patient populations where it has not been trained on. One approach capable of improving generalization capabilities is data augmentation. Data augmentation increases the dataset diversity, so that training the network on more varied data facilitates a better generalization. One way of creating more variation in the data is by simulating different MR contrasts.

The inherent properties of the imaged tissues, such as the proton density (ρ) and magnetic relaxation times (T_1, T_2, T_2^*), contribute collectively to the formation of the MR signal, with the image intensities depending on these factors and being non-quantitative in most conventional MR images. However, it is possible to decompose the MR signal to obtain the contribution of these individual contrast factors and represent their spatial distribution in a quantitative way, as ρ, T_1, T_2 or T_2^* maps. This technique is known as quantitative MRI, or qMRI [5]. Accordingly, having a set of qMRI maps, along with a repetition time (TR) and echo time (TE) of choice, would enable the generation of a broad range of MR contrasts by means of the signal equations.

The gold standard MR sequences for obtaining T_1 and T_2 relaxation maps are Inversion Recovery Spin Echo [6] and Multi-echo Spin Echo [7], respectively, as well as Multi-echo Gradient Echo for T2* maps [8]. Another strategy was developed by In Den Kleef and Cuppen, in which a Multi-echo Spin Echo sequence, interleaved with Inversion Recovery pulses, is utilized to independently derive ρ, T1 and T2 relaxation values simultaneously [9]. This method has been incorporated into Philips scanners as a protocol named MIXED, which was used in this study to acquire qMRI maps to be used in MR signal simulation as one form of data augmentation.

Besides simulating diverse MR contrasts, a different way of creating more variation in the data is SynthSeg [10], which involves the generation of synthetic scans following a domain randomization strategy, meaning that all the parameters of the generative model, including orientation, contrast, resolution, etc., are fully randomized. As opposed to MR signal simulation, these images do not reflect realistic scan contrasts and instead cover a much wider variety of image contrasts beyond what would be seen in real MR images. As a result, the network is forced to learn high-level features such as object shapes, instead of focusing on aspects such as intensities or textures. It should also be noted that these training images are generated from anatomical label maps that are created from the ground truth delineations in an automatic way, ensuring a perfect correspondence between the synthetic scans and the target segmentations. In addition, the image generation process is carried out on the fly, yielding a different image at every training step. Ultimately, the authors demonstrated SynthSeg's high robustness and generalizability to wide variations in contrast and resolution, which was evident not only in MR scans from different modalities but also in computed tomography (CT) images [10].

The authors of SynthSeg proved the generalization capabilities of a network trained on synthetic contrasts varying beyond real contrasts. They state that having enough diversity in the training data by generating random contrasts at every training step is paramount for generalizability. This counterintuitive result, which showed that synthetic contrasts are better than training on real MR images, shows the importance of having lots of variation in the training data. The question remains: does SynthSeg work better than an approach that tries to match its variation by also creating synthetic images on the fly, that stay within the range of plausible MR contrasts?

Hence, the aim of this study is to investigate and compare two data augmentation approaches, which will be applied for the training of a spine segmentation network. The first method is based on the generation of spine images with multiple MR-like contrasts, while the second one employs SynthSeg's strategy to spine images, obtaining volumes with random intensities. The ultimate goal is to determine which technique yields the most generalizability on the model for different, unseen domains.

This work is structured in the following way: after the introduction, chapter II describes the data acquisition, the different image synthesis methods, the pre-processing techniques, the segmentation network employed, and the metrics used for evaluation. Chapter III presents the results of the trainings with both strategies. Chapter IV includes a discussion on these outcomes and, finally, chapter V provides the conclusions of this investigation.

II. Materials and Methods

A. Data

1) Dataset for image synthesis and training

The dataset used in this project, aimed at generating new MR-like and random-intensities images, consisted of a collection of lumbar spine MR images obtained from five different volunteers, 2 female and 3 male, aged between 23 and 25 years old and without any noticeable spinal pathology. The subjects were scanned using a clinical 1.5T MR scanner (Philips Healthcare, Best, Netherlands, software release 5.7) using the base head coil and the built-in posterior coil. For each volunteer, the imaging protocol included a set of quantitative MRI (qMRI) scans, comprising a T_1 map, a T_2 map and a proton density (ρ) map, obtained using the MIXED sequence [9]. Additionally, T_1 -weighted and T_2 -weighted Turbo Spin Echo (TSE) images were obtained, along with a BoneMRI scan, which consists of a 3D T_1 -weighted RF-spoiled Gradient Echo (GRE) sequence, with two echoes with echo times such that one is almost in-phase and the other one is almost out-of-phase, considering the water-fat interference. Parameter settings for each of these acquisitions are included in Table 1. Furthermore, vertebrae segmentations were created, outside of this project, from CT-like BoneMRI images derived from the GRE scans (BoneMRI V1.6 Research Version, MRIguidance B.V., Netherlands).

TABLE I. Acquisition parameters for the different images

Parameter	**Acquisition**			
	2D MIXED	***2D T_1-weighted TSE***	***2D T_2-weighted TSE***	***3D Spoiled GRE***
Repetition time (TR)	TR SE: 1200 ms TR IR: 2000ms	436 ms	2435 ms	7 ms
Echo time (TE)	[15, 38, 61, 84] ms	8 ms	100 ms	[2.1, 4.2] ms
Number of Signal Averages (NSA)	1	3	1	2
Field of View (FOV)	220 × 278 × 101 mm^3	220 × 278 × 101 mm^3	220 × 278 × 101 mm^3	220 × 278 × 100 mm^3
Acquisition voxel size	1 × 1 × 4 mm^3	1 × 1 × 4 mm^3	1 × 1 × 4 mm^3	1 × 1 × 2 mm^3
Reconstruction voxel size	0.7 × 0.7 × 4.4 mm^3	0.7 × 0.7 × 4.4 mm^3	0.7 × 0.7 × 4.4 mm^3	0.7 × 0.7 × 1 mm^3
Acquisition duration	24 min 38 sec	5 min 14 sec	2 min 7 sec	4 min 30 sec

2) Dataset for validation and inference

A different collection of images was employed as validation and test data for the spine segmentation network. This dataset comprised seventeen 3D spoiled GRE and 2D TSE (T_1-weighted and T_2-weighted) scans, belonging to seven different patients, with a mean age of 54 (ranging from 18 to 74) years old and presenting a variety of suspected spinal pathologies, including fractures, deformities, infections, metastases and degenerative conditions. In addition, this dataset included corresponding vertebrae segmentations, also obtained from CT-like BoneMRI images derived from the GRE scans. Note that no data augmentation was applied to the validation and test images to create synthetic images, as was done for the training images. Four scans were used for validation, and the remaining thirteen scans, for training.

B. Physics-based synthesis of MR images (SynthMRI)

This section provides a comprehensive description on augmenting the available data through the implementation of a physics-based strategy. In particular, the signal equations from the turbo spin echo (TSE) sequence were employed to achieve a diverse range of weightings and, therefore, a variety of MR-like contrasts. This creates more diverse data, hopefully yielding better generalizability. For conciseness, this strategy is referred to as "SynthMRI" throughout the rest of this report.

For this approach, the set of qMRI scans from the MIXED sequence was employed, as well as the TSE images. Several ways of synthesizing physics-based contrasts were initially evaluated, which are grouped in two main methods: Bloch simulations and steady state equations.

Bloch simulations describe the behaviour of net magnetization due to the application of a pulse sequence, through three independent dynamics: T_1-relaxation, T_2-relaxation, and precession. After conducting several trials using the functions provided in [11], Bloch simulations were excluded from this project due to the lengthy computation time (the synthesis of a single image slice took approximately 42 seconds for a T_1-weighted TSE image and 99 seconds for a T_2-weighted TSE image, on the other hand, approximately 1.2 seconds were spent on synthesizing the entire 3D image using steady state equations).

For the second method, the steady state signal equation of the spin echo sequence was employed:

$$S = \rho \cdot \left(1 - e^{-TR/T_1}\right) \cdot e^{-TE/T_2} \tag{1}$$

where S is the spin echo signal, ρ is the proton density, TR is the repetition time, T_1 is the longitudinal relaxation time, TE is the echo time and T_2 is the transverse relaxation time. After qualitative evaluation in several tissue of interest (fat, muscle, CSF, spinal cord, bone marrow), it was found that similar contrasts to real TSE images were achieved by this method.

The most straightforward way of synthesizing new MR images is to directly plug in the ρ, T_1 and T_2 values from the qMRI maps in the steady state equation. By doing so, a wide range of contrasts or weightings (T1-weighted, T2-weighted, PD-weighted, combined…) can be simulated based on specific combinations of TR and TE values.

However, a different use of this equation was ultimately reached, that enabled the integration of the acquired (or experimental) TSE images to enhance the synthesized images. This method consisted of defining two steady state signal equations, one corresponding to a real TSE image (S_{exp}), and another one corresponding to a new synthetic image with a simulated contrast (S_s):

$$S_{exp} = \rho \cdot \left(1 - e^{-TR_{exp}/T_1}\right) \cdot e^{-TE_{exp}/T_2} \tag{2}$$

$$S_s = \rho \cdot \left(1 - e^{-TR_s/T_1}\right) \cdot e^{-TE_s/T_2} \tag{3}$$

here, T_1 and T_2 of each tissue of interest can be assumed to be the same in both cases. Nevertheless, the signal of the experimental/real image may be influenced by different factors, such as the hardware of the MR scanner, electromagnetic imperfections (B_0- and B_1- field inhomogeneities, RF receive coil sensitivities or gradient distortions) or the physical properties of the patient (e.g. RF interferences), that are not reflected in the ρ map from the MIXED protocol. Such factors were then contained in the signal equation inside the term *K:*

$$S_{exp} = K \cdot \rho \cdot \left(1 - e^{-TR_{exp}/T_1}\right) \cdot e^{-TE_{exp}/T_2} \tag{4}$$

Thus, using this experimental TSE signal, a new proton density map was defined that encompassed these factors:

$$PD = K \cdot \rho = \frac{S_{exp}}{\left(1 - e^{-TR_{exp}/T_1}\right) \cdot e^{-TE_{exp}/T_2}} \tag{5}$$

Substituting ρ in equation (3) by the estimated PD in equation (5) yields

$$\begin{aligned} S_s &= PD \cdot \left(1 - e^{-TR_s/T_1}\right) \cdot e^{-TE_s/T_2} \\ &= S_{exp} \cdot \frac{\left(1 - e^{-TR_s/T_1}\right)}{\left(1 - e^{-TR_{exp}/T_1}\right)} \cdot e^{\frac{-(TE_s - TE_{exp})}{T_2}} \end{aligned} \tag{6}$$

Therefore, synthetic images were influenced by the signal and time parameters of a real TSE experiment, which affects their contrast and improves the level of detail, achieving realistic results. The noise in the qMRI maps, that would be excessive if the ρ map from MIXED was used, was also almost entirely mitigated, except for some speckles caused by outliers in the maps that, due to the divisions and exponentials in the equations, resulted in values that are outside of the dynamic range of the image.

Several intermediate and mixed weightings were simulated by choosing a random TE_s within the range of 5 to 100 ms, and a random TR_s within the range of 200 to 3000 ms. These ranges match those commonly used in conventional TSE acquisitions, encompassing various weightings. Since T_1-weighted images are acquired with short echo times, it was decided to couple the random selection of a TE_s below 50 ms with the utilization of a PD map estimated from the experimental T_1-weighted TSE signal. Conversely, since T2-weighted images require longer echo times, the choice of a TE_s greater than 50 ms would be linked to the use of a PD map defined from the experimental T_2-weighted TSE signal. This criterium is graphically depicted in Figure 1.

It should be noted that choosing TR_s and TE_s equal to the experimental ones results in a synthetic image also identical to the experimental one, as equation (6) becomes

$$S_s = S_{exp} \tag{7}$$

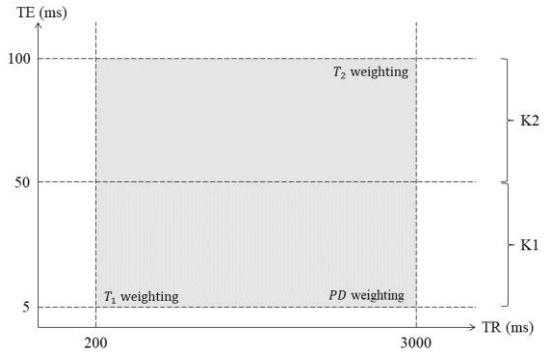


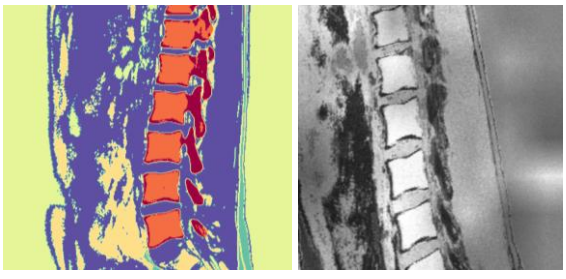
Fig. 1. Criterion for the automatic synthesis of new MR images. Synthetic images are automatically generated by randomly selecting a combination of TR and TE values contained inside the shaded area. In addition, depending on wether TE is lower or higher than 50 ms, we consider K_1 or K_2, which represent the scanner imperfections arising in a real T_1-weighted or T_2-weighted TSE image, respectively.

C. *Random intensity-based synthesis of images (SynthSeg)*

This section explores in detail the functioning of SynthSeg's [10] approach and its adaptation to this project, with the goal of using the available data to automatically generate new spine images with random intensities on each training step. In this way, the segmentation model is forced to focus only on object shapes, disregarding information related to intensity or contrast due to its high variation and randomness. For the sake of simplicity, the name "SynthSeg" is also utilized throughout this work.

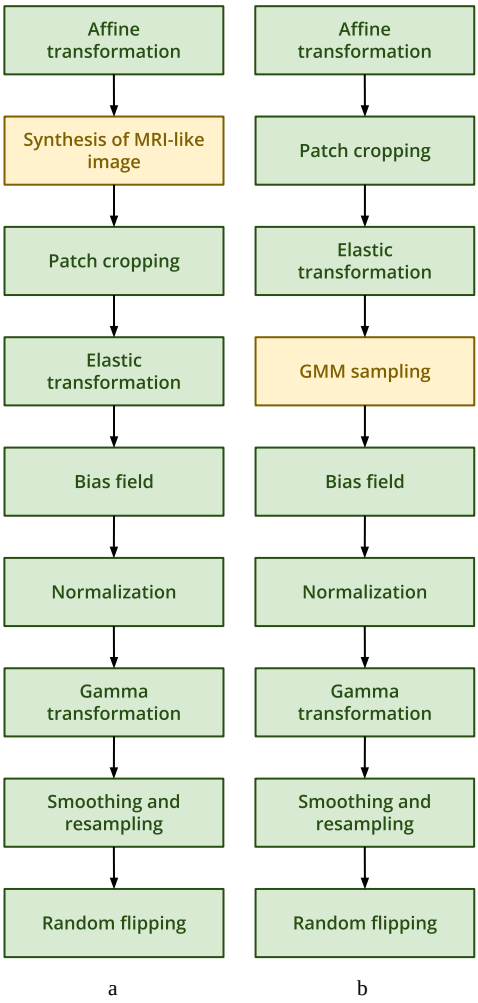
Synthesizing spine images with random contrasts required the utilization of anatomical label maps, derived from the ground truth vertebrae segmentations and an MRI image from that acquisition, such as a TSE. Section D in this chapter, titled "Data pre-processing", describes in more detail how these maps were generated. The label maps contain a variable number of clusters within and outside the foreground (i.e., the vertebrae), as depicted in Figure 2a, that represent different body structures and levels of granularity.

Label maps take their values from a set of K labels, $S_n(x, y, z) \in \{1, \dots, K\}$ [10], where K represents the number of clusters inside the volume. To synthesize the new images, a Gaussian Mixture Model (GMM) was used to generate random intensities for each of the K labels so that, every time, each label $k \in \{1, \dots, K\}$ was associated with a Gaussian distribution of intensities having mean μ_k, and standard deviation σ_k [12]. The function *RandomLabelsToImage*, from the Python library *TorchIO*, was employed for this task.



a b

Fig. 2. (a) Label map. (b) Synthetic image with random intensities derived from the label map.



a b

Fig. 3. (a) Order of augmentations for SynthMRI strategy. (b) Order of augmentations for SynthSeg strategy.

Moreover, spatial augmentations, random bias field artifacts, intensity augmentations and resampling were also applied to the new images being created on the fly to make training robust to such effects and to create even more variation. These augmentations were implemented in the same way in the images generated by the SynthMRI strategy, to make both approaches as comparable as possible. The order in which these transformations were executed for each of them is depicted in Figure 3 and has slight differences due to computational speed considerations. Figures A1 and A2 in the Appendix show examples of the final synthetic images generated by SynthMRI and SynthSeg after the application of these augmentations.

The spatial augmentations consisted of an affine and an elastic transformation. Between them, patches of 80×80×32 pixels were cropped from the images. To add random bias field artifacts, the approach used in [10] was followed, where a volume with size 4^3 was first sampled from a Gaussian distribution with zero-mean and standard deviation σ_B. This volume was then expanded to the full size of the image, and an exponential function was used on each voxel to ensure a smooth and non-negative field B. Subsequently, the multiplication of this field B by the original image resulted in the creation of a biased image. Next, the image was min-max normalized to have values between 0 and 1, and the intensity distribution of the synthetic scans was further augmented by

the application of a random Gamma transformation through voxel-wise exponentiation. Then, images were smoothed via Gaussian filtering and resampled, to simulate low-resolution scans. This resampling consisted of a random downsampling followed by upsampling, using trilinear interpolation, to return to the original patch dimension. Finally, image flips along the antero-posterior and left-right directions were randomly applied with a 50% probability. Figure 2b illustrates the image generated after all these steps, from its corresponding label map. For visualization purposes, a bigger patch of 300×300×32 pixels is depicted.

D. Data pre-processing

Pre-processing was required for the T_1 maps and the T_2 maps from the MIXED sequence, the T_1- weighted and T_2-weighted TSE scans and the vertebrae segmentations, as well as for all the images of the validation and inference dataset. This process will be described below.

Initially, all images were resized to a common resolution of 0.7×0.7×2.2 mm^3, using 3rd order B-spline interpolation. As noted previously, the ground truth segmentations were obtained from the CT-like BoneMRI images derived from the GRE scans, which had 1 mm slice thickness. While SynthSeg could theoretically be applied for training at this resolution, it would require upsampling the qMRI maps and TSE images for SynthMRI by a factor of 4.4 in the left-right direction. Therefore, to ensure a fair comparison between both approaches, a slice thickness of 2.2 mm, was chosen as an intermediate resolution between that of the 1 mm thick slices of the segmentation and the 4.4 mm thick slices of the TSEs.

1) SynthMRI

The resized images for the SynthMRI approach were pre-processed further. First, the T_1 and the T_2 maps from the MIXED sequence were filtered to remove the noise in a two-step process. The first one consisted of a thresholding operation applied to the abdominal region containing no vertebrae. Voxels exceeding a value of 625 ms, determined heuristically, were clipped to this number. The second step consisted of the application of a median filter of size 3×3 in the sagittal plane, as this provided the best noise reduction with minimal blurring.

After de-noising the maps, two proton density (PD) maps were estimated from the T_1 and the T_2 maps and the TSE signals, utilizing the experimental TR and TE, following equation (5). At training time, the T_1 and T_2 maps, and the PD maps from each volunteer were used to generate a synthetic scan with an MRI-like contrast using equation (6) and the criterion defined in Figure 1, which determined if the T_1 TSE-derived map or the T_2 TSE-derived map would be used.

2) SynthSeg

The pre-processing steps for the SynthSeg approach involved the creation of the label maps, which were obtained by clustering MRI intensities following the Expectation-Maximization (EM) algorithm, as mentioned in [10], where it was employed for the same purpose. In this project, the T_1-weighted TSE scans of each subject were initially separated into a vertebrae label map and a background (everything that is not vertebrae) label map, using the ground truth segmentation as a mask. The EM algorithm was applied to the vertebras using the *GaussianMixture* function from the Python library *scikit-learn*. In three of the volunteers, two clusters were determined for the vertebrae label map, corresponding to the vertebral cortex and bone marrow. However, two of the subjects also presented some hyperintensities inside the bone marrow, which were included as a third cluster. For the background, eight different label maps were generated, determining for each of them a different number of clusters, that ranged from 3 to 10. This helps to introduce more data variation during contrast synthesis. Finally, the vertebrae label maps were combined with the background maps to produce the complete label maps.

E. Segmentation network and training

A UNet architecture [13] for 3D image segmentation was employed, depicted in Figure 4. This architecture was adapted from the *BasicUNet* implementation from *MONAI*, a deep

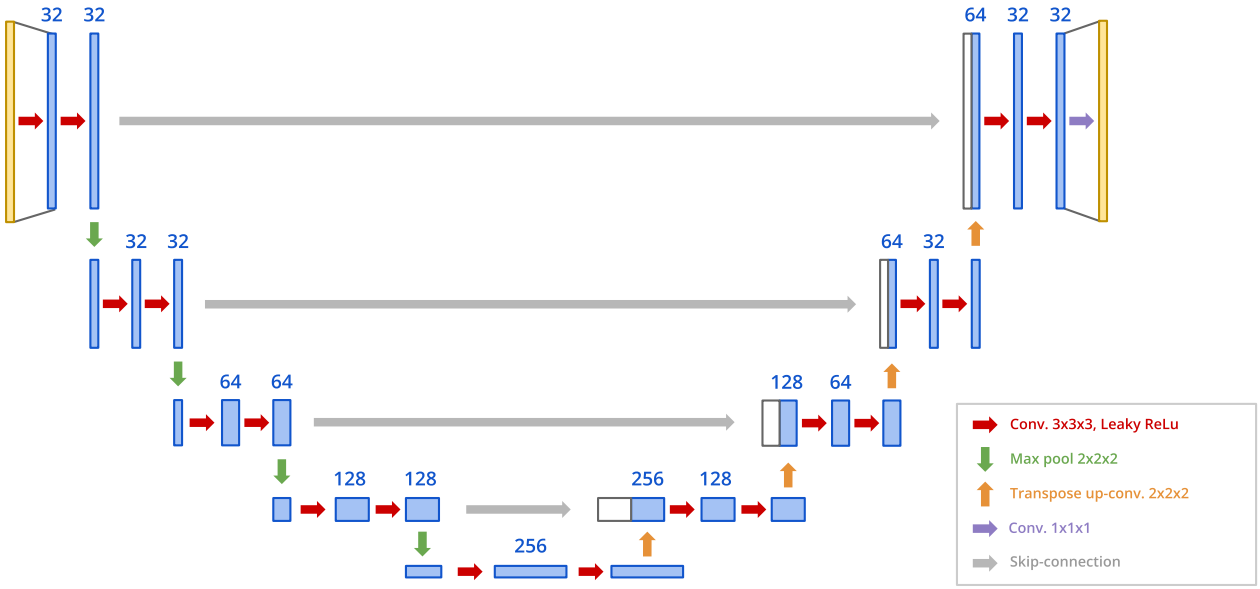


Fig. 4. Scheme of the 3D UNet architecture employed in this work, where the blue rectangles represent the feature maps with their corresponding number of features after every operation indicated by the arrows.

learning framework for medical imaging [14], and it consisted of five levels, along which the number of features was increased from 32 to 256. Max pooling operations were used for downsampling and transpose convolutions for upsampling, both using 2×2×2 kernels. Each level contained two convolutions with 3×3×3 kernels, followed by an instance normalization layer, a dropout layer with probability 0.1 and a Leaky ReLu activation function. In the last layer, a 1×1×1 convolution was employed to map the feature vectors to the number of classes which, in this case, is 2 (i.e., 0 for the background and 1 for the vertebrae). Same as in [10], the loss function employed for training was the soft Dice Loss.

The UNet was trained twice with both of the data synthesis strategies. For each one, all augmentations and image generation steps were executed on each learning step to generate new training samples. Both approaches utilized a learning rate of 0.0001 and a batch size of 40.

On each epoch, the validation images also underwent patch cropping, maintaining the same size as in training (80×80×32). During training, the model weights were saved periodically, every 560 iterations. The models attaining the three best validation metrics were also saved. The UNet trained with SynthMRI strategy was trained for 59,920 iterations, whereas the one trained with SynthSeg strategy was trained for 333,200 iterations.

F. Inference

To select a model for inference, the last three saved models and the three models achieving the best validation metrics were used to predict the validation images. Each model was evaluated against the ground truths using the Dice Similarity Coefficient. The model achieving the best scores was ultimately selected for the final evaluation on the test data. This procedure was applied to both the UNet models trained using SynthMRI and those trained using SynthSeg. For SynthMRI, the last saved model (which trained for 59,920 iterations) was selected, whereas for SynthSeg, the model with the second best validation loss (which trained for 329,168 iterations) was selected.

For inference, patches of size 80×80×32 were sampled from the test images with a stride of half the patch size. The overlapping predictions were fused with a Gaussian weighting to mitigate potential artifacts near the edges using *MONAI*'s *SlidingWindowInferer* implementation.

G. Evaluation metrics and statistical analysis

The performance of both methods was evaluated on the predicted segmentations against their respective ground truths by using the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff distance (HD95), given in millimetres (mm). The selection of HD95 over the conventional Hausdorff distance was justified by its reduced sensitivity to small outliers, which provides a more robust estimate of the maximum error or distance between the segmentations.

A posterior statistical analysis of the values derived from these metrics was conducted to investigate the presence of significant differences in performance between the SynthMRI and SynthSeg approaches. Because the test samples were not independent, the non-parametric Wilcoxon signed-ranks test was conducted for both the DSC and HD95, considering all test images but also per-modality, separately evaluating the performance of both methods on the GRE and TSE scans.

III. Results

The outcomes of the SynthMRI and SynthSeg strategies are presented hereafter in both quantitative and qualitative terms.

Boxplots in Figure 5 provide a visual comparison of SynthMRI and SynthSeg's performances, with the corresponding mean values shown in Table A1, in the Appendix. The DSC and HD95 values attained by both approaches are plotted for the GRE and TSE scans separately and taken together. Figure 5 shows that, despite both SynthMRI and SynthSeg achieving comparable metric values, the first one exhibits a greater consistency across all test images. This is particularly evident in the DSC values, which show less variation than those obtained by SynthSeg. However, there is no significant difference in overall performance between both strategies. When analysing the DSC coefficients per modality, no significant differences are observed between SynthMRI and SynthSeg in the GRE images, although SynthSeg attained slightly higher DSC values. Nevertheless, significant differences can be found in the TSE images, where SynthMRI outperforms SynthSeg. There is no significant difference between the HD95 of both strategies when evaluated on all image types. As with the DSC coefficients, a significant difference in HD95 values between

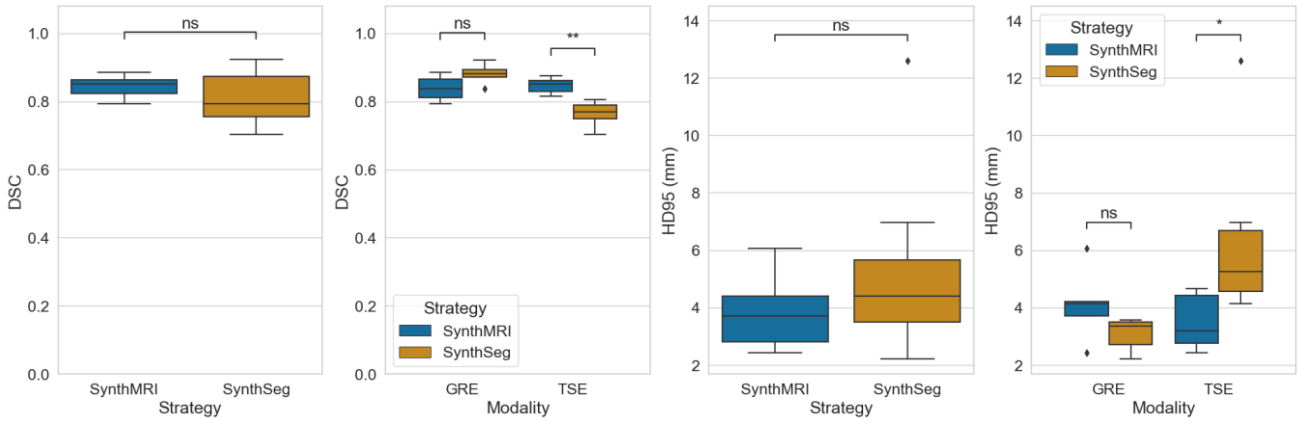


Fig. 5. Boxplots representing the DSC and HD95 metric scores attained by SynthMRI and SynthSeg on the test images. The first and third plots depict the DSC and HD95, respectively, over all predictions, while the second and fourth plots illustrate the DSC and HD95, respectively, grouping the predictions according to test image modality.

SynthMRI and SynthSeg is observed when considering only the predictions on the TSE images. Lastly, it is worth mentioning the disparity in performance exhibited by SynthSeg across the different modalities, where inference on GRE images achieves better metric scores than on TSE images.

Qualitative results of the two approaches are illustrated in Figure 6, which includes sagittal slices of the images with the highest and lowest performances of SynthMRI and SynthSeg, in terms of both the DSC and HD95. Both strategies achieve their best performance on the same test image, with a GRE contrast and no spinal pathologies. SynthMRI presents some inaccuracies in the vertebral bodies, where a few groups of voxels are classified as background. On the other hand, SynthSeg attains an almost complete match with the ground truth in the vertebral bodies but is less accurate in the spinous processes. SynthMRI performs worse in a GRE scan that reveals a pathology inside two of the vertebrae. Its prediction fails to classify the voxels inside and surrounding this pathology as foreground, or vertebra voxels. SynthSeg achieves its worst metric values in a TSE image where the vertebrae are deformed. SynthSeg fails considerably in the delineations of the spinous processes, completely missing some of them and merging others. In addition, it introduces some incorrect classifications of foreground voxels in background areas.

IV. Discussion

A. *Physics-based synthsesis of MR images (SynthMRI)*

The DSC and HD95 values achieved by the network trained with SynthMRI data clearly show that the variability introduced by this approach was high enough to attain good generalization capabilities. While a strong performance of this strategy on TSE scans could be expected, as it relied on the generation of images with a wide range of weightings for this contrast, it is also noteworthy its similar performance in GRE scans, which were not observed during training.

Probably the greatest limitation of this approach is the long scanning time of the MIXED sequence from which the qMRI maps are obtained, which made it necessary to acquire thick slices (4.4 mm), hindering the ability of the network to learn detailed features and possibly preventing SynthMRI from achieving better results.

B. *Random intensity-based synthesis of images (SynthSeg)*

Following the methodology in [10], SynthSeg was also able to achieve a good performance and generalizability in spine images. Even though it obtained favourable DSC and HD95 values for all test images, the differences between GRE and TSE scans are noteworthy. When examining qualitatively the predictions performed by SynthSeg on the TSE images, it is noticeable how this method fails specially in the segmentation of the spinous processes. These predictions are

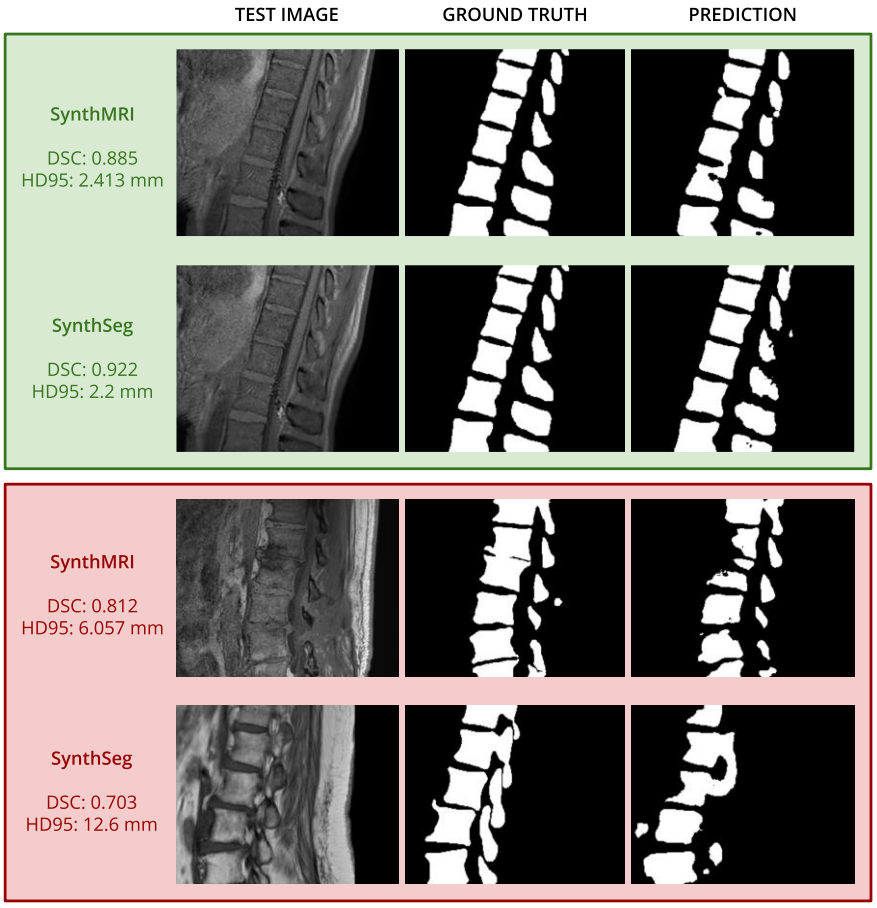


Fig. 6. Sagittal slices obtained from several predictions by the models trained using SynthMRI and SynthSeg. The green box contains the predictions with best metric scores for each method. The red box contains the predictions with worst metric scores for each method.

depicted in Figures A3 and A4 in the Appendix, along with their respective TSE scans and ground truths.

It can be observed that the edges between the vertebrae and the surrounding tissues are quite discernible in SynthSeg's synthetic images, shown in Figure A2, since the intensities inside and outside the vertebrae are always drawn from different Gaussian distributions. In contrast, the boundaries between these structures in the TSEs are more diffuse, especially in the transverse and spinous processes. This is likely due to partial volume effects, that average together the signals of bone and surrounding tissue due to the spinous processes having a smaller width than that of the slice (i.e., less than 4.4 mm), as well as the signals from the bone marrow and the thin layer of cortical bone, eventually showing almost no dark edge around the vertebrae. Moreover, given the already similar intensities of these tissues in the TSEs, the presence of partial volume effects further complicates their discrimination. Hence, this could offer a plausible explanation for SynthSeg's weaker performance in the TSE images and, particularly, in the spinous processes. On the other hand, the GREs originally presented a thinner slice thickness (i.e., 1 mm), which results in the cortical bone signal being less affected by partial volume effects. In this way, the edges between bone and surrounding tissues are clearer, facilitating the detection of these shapes and making these images more similar to the synthetic training data from SynthSeg.

The resampling of all images to the common voxel size of $0.7\times0.7\times2.2$ mm^3 could have also affected the creation of the label maps used for this strategy, where the ground truth segmentations were used as a mask on the TSE images to cluster the intensities inside and outside the vertebrae. Since these ground truths were computed from the CT-like BoneMRI images derived from the GRE scans, it is likely that a worse alignment between image and target segmentation occurred in the TSE scans. Therefore, parts of these vertebrae could have been clustered as background intensities and vice versa. Because SynthSeg focuses on learning the object shapes and other domain-independent features, not having accurate label maps could have prevented this method from performing better, overall.

C. Comparison between SynthMRI and SynthSeg

As stated in the Results chapter, no significant differences were found between SynthMRI and SynthSeg except when analysing solely the predictions on the TSE scans. Here, SynthMRI demonstrated superior results, mainly due to SynthSeg's worse performance on this type of data in contrast to the GRE scans. SynthMRI and SynthSeg's best prediction was made on the same patient, which did not present any noticeable spinal pathology and therefore showed vertebrae similar to those encountered during training.

A reason why SynthMRI performed better than SynthSeg on the TSE images could be due to the fact that this first approach learns from the intensities and contrasts of the images, and not just from the object shapes, making it more robust to the partial volume effects. Figures A3 and A4 show how SynthMRI was better able to segment the spinous processes than SynthSeg.

In addition, the resampling process in all images (training, validation and test) not only affected the creation of the label maps for SynthSeg, but it could also have impacted the learning process and the evaluation of the predictions for the two approaches. Both SynthMRI and SynthSeg did not train in their original (and probably most optimal) resolution and, as mentioned in the previous section, the alignment between the TSE scans and the ground truths possibly lacked accuracy. Nevertheless, this step was necessary to ensure a fair comparison between them.

Both data synthesis methods present distinct advantages and disadvantages, and the choice of using one or the other may depend on different factors. If generalization only needs to be attained on a specific type of MR sequence, and segmentations can be done at the original acquisition resolution, then SynthMRI could be more suitable, since employing information about the intensities and contrasts of these images would be useful in this case. However, qMRI maps and one or several MR images from this type of sequence, along with their corresponding segmentation, will need to be acquired specifically for this purpose. If this is not feasible, then SynthSeg would be a better approach, since it can be fully applied to existing MR data (as long as it has its respective ground truth), which can be more convenient in practice. In addition, because the qMRI maps for SynthMRI always present thick slices, SynthSeg can be a better fit for segmentations at high resolutions, as well as for generalization to very diverse MR sequences or, possibly, to other imaging techniques such as CT.

D. Comparison to other spine segmentation methods

A small and homogeneous training dataset typically presents a challenge for achieving generalization in segmentation networks. However, SynthMRI and SynthSeg were able to generate enough variation from a limited set of MR images and train models that generalized successfully to unseen scan types, populations and pathologies. This is further evidenced when comparing the results obtained by both approaches with those achieved by other recently developed spine segmentation methods. Latest works have attained mean DSCs of 0.828 (averaged over the vertebrae and intervertebral discs) [15], 0.93 [16], 0.735 [17] or 0.903 [18]. Note that [15] and [16] conduct per-vertebra segmentation, assigning different labels to each vertebra. SynthMRI was able to outperform [15] and [17], with a mean DSC of 0.843, whereas SynthSeg only surpassed [17], with a mean DSC of 0.81.

It is noteworthy that the datasets employed for training in these studies comprised hundreds of labelled spine images (except for [17], which employed 6 spine images together with 189 images from other bones), in contrast to the five images SynthMRI used for training in this work. In addition, these studies tested their models on similar data as the one used for training, in terms of image modalities and patient populations. On the other hand, SynthMRI and SynthSeg were trained with scans belonging to healthy volunteers with a mean age of 24 years old and were tested on images from patients with severe spinal pathologies and a mean age of 54 years old, still achieving good results. Additionally, even though they did not see any GRE scan during training, they were also able to provide accurate segmentations on these scans on the testing phase.

E. Future research

Several refinements and corrections can be applied to these methods in future research scenarios to address the different limitations.

One direction could involve expanding the variability of contrast types incorporated into the SynthMRI strategy beyond the current approach of just synthesizing different

weightings of a TSE sequence. In other words, this could entail introducing more signal equations. This would provide even more variation to the synthetic data, potentially leading to an improved generalization ability of the network. Note that the introduction of new signal equations would also require the acquisition of their corresponding real scans to estimate their corresponding *PD* maps.

Further enhancing both the quantity and variability of training data would also be favourable. This could entail including more images not only from healthy individuals but also from patients with diverse spinal conditions. These cases, which cannot be accurately reflected by deformations or other augmentations applied to the images, could aid in improving the robustness and generalizability of the network. Unfortunately, this task may be particularly challenging for the SynthMRI strategy since obtaining new qMRI scans with their respective TSE scans would pose significant logistical and time constraints.

Additionally, it would be advantageous to conduct the inference phase across more diverse imaging domains, introducing both healthy and pathological test images, as well as a wider range of contrasts and image domains. This could be especially insightful for SynthSeg, as it would allow to observe whether the difference in performance between scan domains also extends to other contrasts.

Furthermore, improving the accuracy of the ground truth images, by specially obtaining specific ones for the TSE images instead of adapting them from the segmentations derived from the GRE scans would be highly beneficial. Addressing this limitation could contribute to better learning by the network and a more reliable evaluation of the predictions.

Lastly, optimizing image resolution is vital for ensuring the efficacy of the segmentation models. While maintaining a fixed resolution for a fair comparison of both data synthesis approaches was important for this work, future investigations should adjust the resolution to suit the strengths of each strategy. While the resolution of the qMRI and TSE scans employed for the SynthMRI approach could be maintained, SynthSeg's resolution should keep the smallest possible voxel size in all three dimensions to enhance the learning of the shape features.

V. Conclusion

This study compared the performance of two approaches aimed at synthesizing data on the fly from existing MRI scans to train a generalizable spine segmentation network. The outcomes of this research revealed that SynthMRI, a physics-based method generating images within the range of weightings of a TSE sequence, demonstrated comparable generalizability to unseen domains when evaluated against SynthSeg, a random intensity-based data synthesis method. Overall, no significant differences in performance were observed between both approaches, except when splitting the results by modality, where SynthMRI outperformed SynthSeg in TSE scans. Nevertheless, the two methods effectively augmented the quantity and diversity of a limited training dataset, achieving a high generalization capability and demonstrating the promising potential of SynthMRI and SynthSeg in advancing spine segmentation techniques.

References

[1] G. Hille, S. Saalfeld, S. Serowy, and K. Tönnies, 'Vertebral body segmentation in wide range clinical routine spine MRI data', *Computer Methods and Programs in Biomedicine*, vol. 155, pp. 93–99, Mar. 2018, doi: 10.1016/j.cmpb.2017.12.013.

[2] M. Vania, D. Mureja, and D. Lee, 'Automatic Spine Segmentation using Convolutional Neural Network via Redundant Generation of Class Labels for 3D Spine Modeling'. arXiv, Nov. 29, 2017. Accessed: Feb. 09, 2024. [Online]. Available: http://arxiv.org/abs/1712.01640

[3] S. K. Warfield, K. H. Zou, and W. M. Wells, 'Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation', IEEE Trans. Med. Imaging, vol. 23, no. 7, pp. 903–921, Jul. 2004, doi: 10.1109/TMI.2004.828354.

[4] L. El Jiani, S. El Filali, and E. H. Benlahmer, 'Overcome medical image data scarcity by data augmentation techniques: A review', in *2022* International Conference on Microelectronics (ICM), Casablanca, Morocco: IEEE, Dec. 2022, pp. 21–24. doi: 10.1109/ICM56065.2022.10005544.

[5] H. Margaret Cheng, N. Stikov, N. R. Ghugre, and G. A. Wright, 'Practical medical applications of quantitative MR relaxometry', *Magnetic Resonance Imaging*, vol. 36, no. 4, pp. 805–824, Oct. 2012, doi: 10.1002/jmri.23718.

[6] N. Stikov, M. Boudreau, I. R. Levesque, C. L. Tardif, J. K. Barral, and G. B. Pike, 'On the accuracy of T 1 mapping: Searching for common ground: Accuracy of T 1 Mapping', Magn. Reson. Med., vol. 73, no. 2, pp. 514–522, Feb. 2015, doi: 10.1002/mrm.25135.

[7] N. Ben-Eliezer, D. K. Sodickson, and K. T. Block, 'Rapid and accurate T 2 mapping from multi-spin-echo data using Bloch-simulation-based reconstruction: Mapping Using Bloch-Simulation-Based Reconstruction', Magn. Reson. Med., vol. 73, no. 2, pp. 809–817, Feb. 2015, doi: 10.1002/mrm.25156.

[8] K. Egger *et al.*, 'T2* Relaxometry in Patients with Parkinson's Disease: Use of an Automated Atlas-based Approach', *Clin Neuroradiol*, vol. 28, no. 1, pp. 63–67, Mar. 2018, doi: 10.1007/s00062-016-0523-2.

[9] J. J. E. In Den Kleef and J. J. M. Cuppen, "RLSQ: T1, T2, and ρ calculations, combining ratios and least squares," Magn. Reson. Med., vol. 5, no. 6, pp. 513–524, Dec. 1987, doi: 10.1002/mrm.1910050602.

[10] B. Billot *et al.*, 'SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining', *Medical Image Analysis*, vol. 86, p. 102789, May 2023, doi: 10.1016/j.media.2023.102789.

[11] B. Hargreaves, "Bloch Equation Simulation." http://www-mrsrl.stanford.edu/~brian/bloch/

[12] B. Billot, D. Greve, K. Van Leemput, B. Fischl, J. E. Iglesias, and A. V. Dalca, 'A Learning Strategy for Contrast-agnostic MRI Segmentation', 2020, doi: 10.48550/ARXIV.2003.01995.

[13] O. Ronneberger, P. Fischer, and T. Brox, 'U-Net: Convolutional Networks for Biomedical Image Segmentation'. arXiv, May 18, 2015. Accessed: Feb. 09, 2024. [Online]. Available: http://arxiv.org/abs/1505.04597

[14] M. J. Cardoso *et al.*, 'MONAI: An open-source framework for deep learning in healthcare'. arXiv, Nov. 04, 2022. Accessed: Feb. 09, 2024. [Online]. Available: http://arxiv.org/abs/2211.02701

[15] Y. Deng *et al.*, 'An effective U-Net and BiSeNet complementary network for spine segmentation', *Biomedical Signal Processing and Control*, vol. 89, p. 105682, Mar. 2024, doi: 10.1016/j.bspc.2023.105682.

[16] J. W. van der Graaf *et al.*, 'Lumbar spine segmentation in MR images: a dataset and a public benchmark'. arXiv, Jun. 22, 2023. Accessed: Feb. 08, 2024. [Online]. Available: http://arxiv.org/abs/2306.12217

[17] H. Gu *et al.*, 'SegmentAnyBone: A Universal Model that Segments Any Bone at Any Location on MRI'. arXiv, Jan. 23, 2024. Accessed: Feb. 08, 2024. [Online]. Available: http://arxiv.org/abs/2401.12974

[18] L. Li *et al.*, 'ICUnet++: an Inception-CBAM network based on Unet++ for MR spine image segmentation', *Int. J. Mach. Learn. & Cyber.*, vol. 14, no. 10, pp. 3671–3683, Oct. 2023, doi: 10.1007/s13042-023-01857-y.

APPENDIX

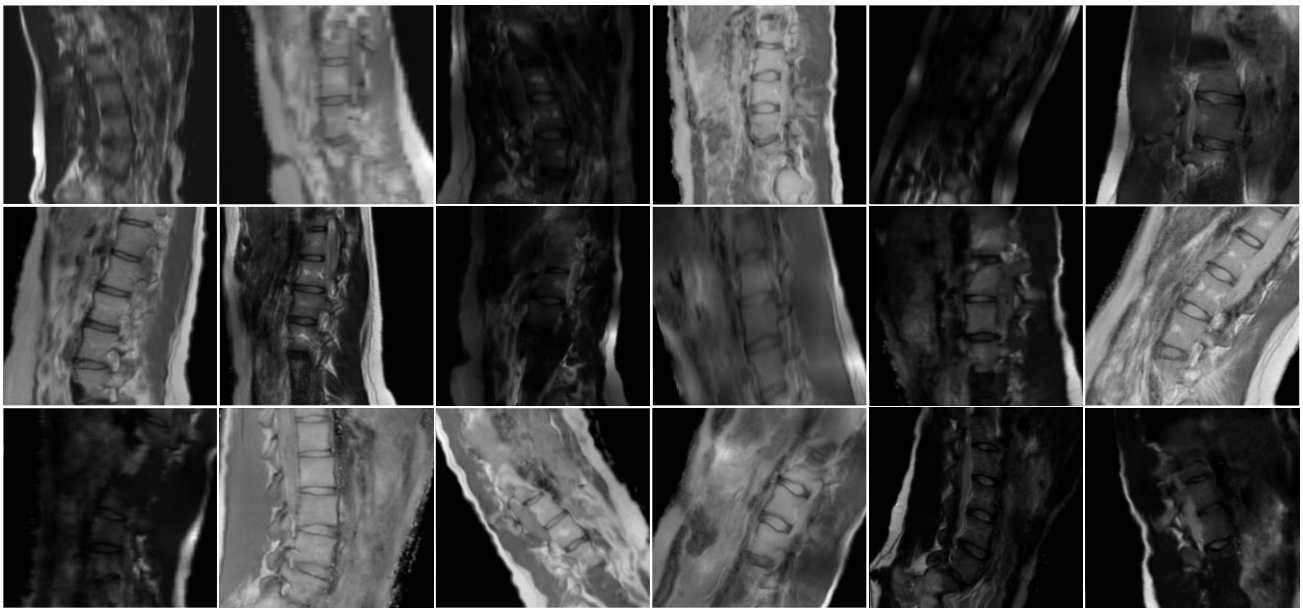


Fig. A1. Representative examples from the physics-based data synthesis approach (SynthMRI). Synthetic data presents variation not only inside the plausible range of TSE weightings, but also in terms of resolution, shape or bias field effects.

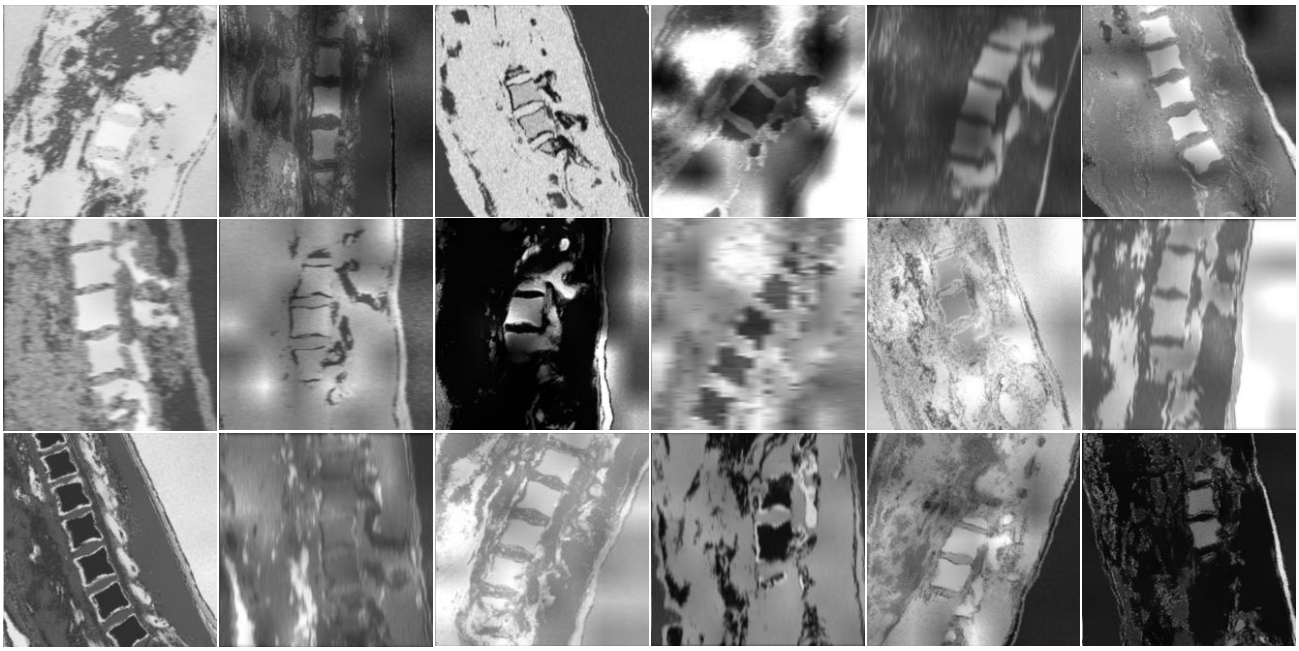


Fig. A2. Representative examples from the random intensity-based data synthesis approach (SynthSeg). Synthetic data presents variation not only in terms of intensities and contrasts, but also in terms of resolution, shape or bias field effects.

TABLE A1. MEAN VALUES FOR EVALUATION METRICS

		SynthMRI	*SynthSeg*
DSC	All images	0.843	0.810
	GRE	0.838	0.882
	TSE	0.847	0.765
HD95 (mm)	All images	3.712	5.008
	GRE	4.100	3.062
	TSE	3.470	6.224

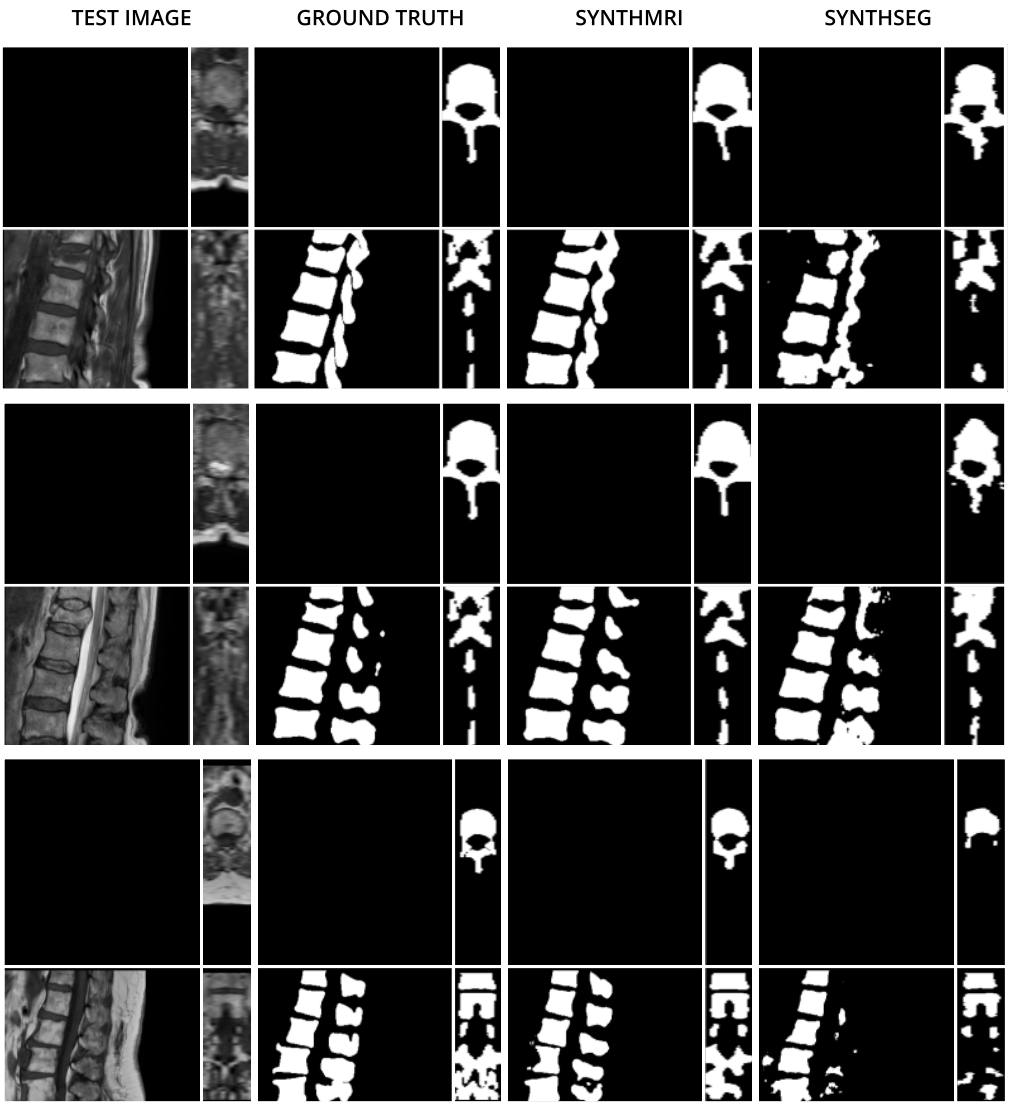


Fig. A3. Predictions made by SynthMRI and SynthSeg on the TSE images, part 1.

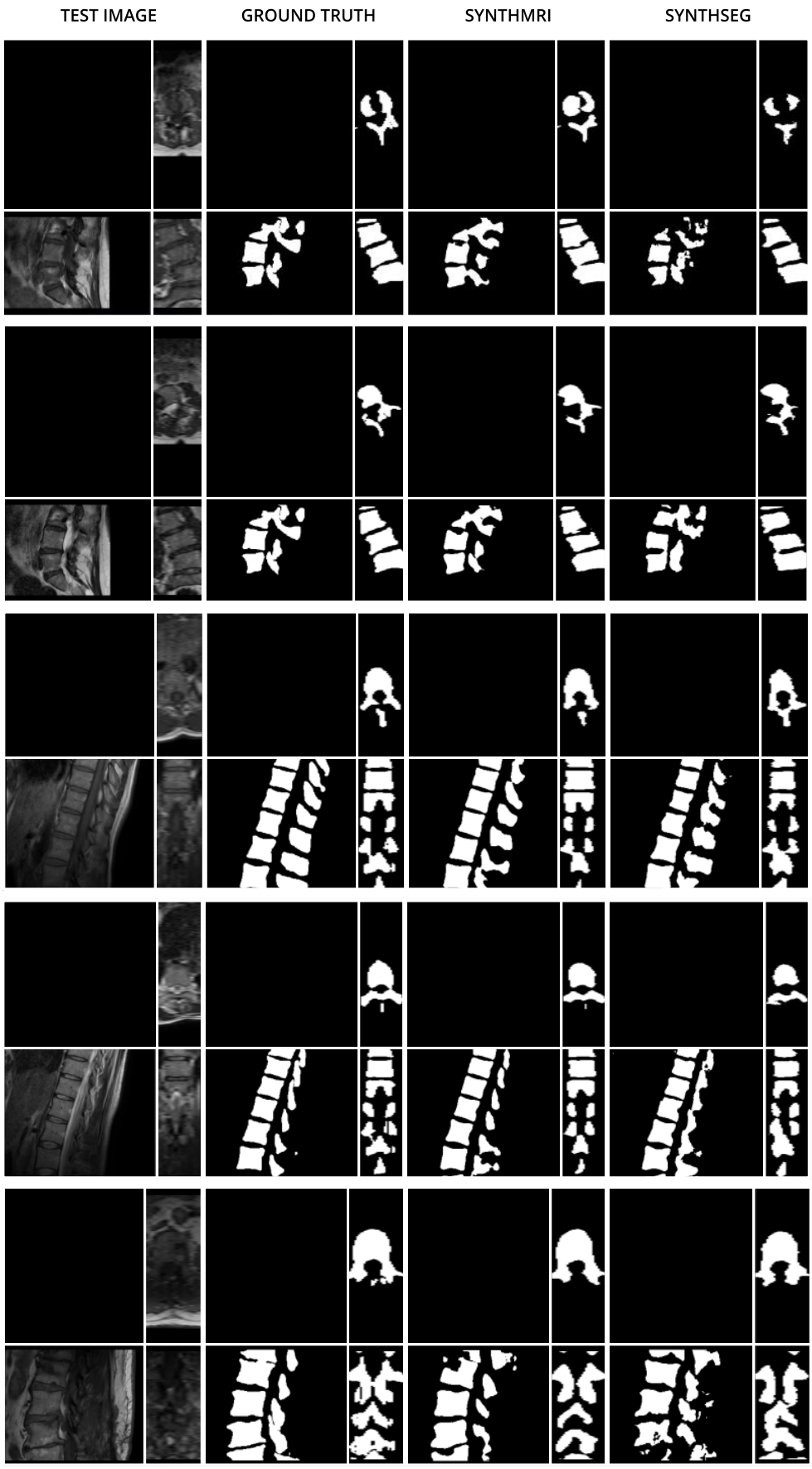


Fig. A4. Predictions made by SynthMRI and SynthSeg on the TSE images, part 2.