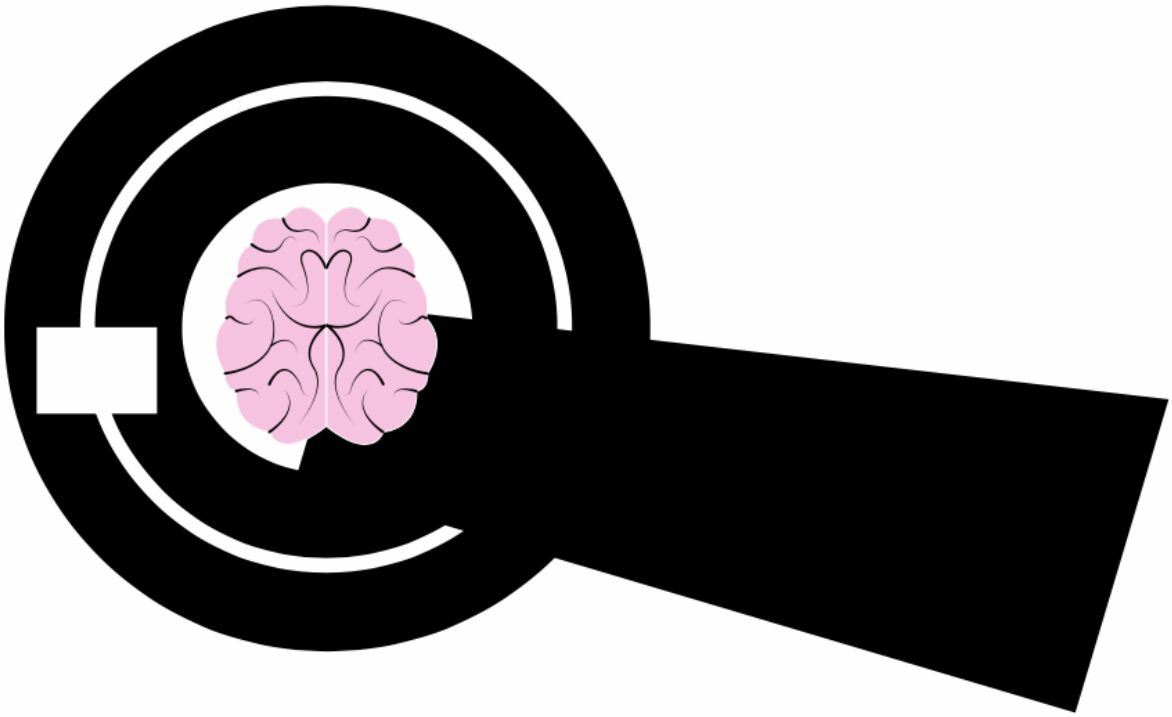
QUANTITATIVE MR INTER-SCANNER HARMONIZATION USING IMAGE STYLE TRANSFER



Patricia Enríquez Calzada

TU Delft

Erasmus MC
Universitair Medisch Centrum Rotterdam

Quantitative MR inter-scanner harmonization using image style transfer

By

Patricia Enríquez Calzada

in partial fulfilment of the requirements for the degree of

Master of Science
in Biomedical Engineering

at the Delft University of Technology,
to be defended publicly on Wednesday January 27, 2021 at 13:30 PM.

Thesis committee:	Dr. F.M. Vos	TU Delft
	Dr. J.A Hernández-Tamames	Erasmus MC
	Dr. J.C van Gemert	TU Delft
	Dr. M de Bruijne	Erasmus MC

TU Delft

Preface

This report is the result of my master thesis project and concludes my Master of Science in Biomedical Engineering at the Delft University of Technology in Delft, The Netherlands. This work was developed within the Radiology Department and the BIGR at the Erasmus Medical Center in Rotterdam, together with the Imaging Physics group department of TU Delft.

I would like to thank everyone who has been there for me these past twelve months. First, I would like to express my deepest gratitude to my supervisors. To Juan Antonio Hernández Tamames, who gave me the chance to work on this project and always trusted my capabilities. To Frans Vos, who always provided valuable feedback and suggestions in our progress meetings. I learned so much from you these months.

I would like to thank Marleen de Bruijne and Sebastian Van der Voort for generously taking their time to help me with their Deep Learning knowledge when I was stuck. Also, to Laura Núñez and everybody at the BIGR for making me feel so welcomed at the hospital and helping me get started with the project.

Thank you to my friends, especially those with whom I have shared my last years in Delft. I want to thank the best roommates in the world: Andrei, Farkas, Joel, Matthieu and Sara. I can't believe I tricked you into buying me a boat. You made Delft as much of a home to me as Madrid. To Henar, I could write a separate thesis about how important your friendship has become to me this past year. My happy place is going for a walk with you to "our" favorite street. To Júlia, I can always count on you for both the stupidest and the deepest stuff. My first single is as yours as it is mine. To Jacob for being my sticker muse, I am sorry annoying you is my hobby. To Priyanka who I admire more every time we catch up and she tells me another achievement. To Edamamens for never asking me to pay rent nor to lower my terrible singing voice. To Pingonas for always making me laugh and feel supported. I want to thank Bioguays, for still feeling so close even through the distance, I can't wait for our reunion trip. To Carlos and Iñigo, you are home to me, thanks for always listening. To Pariseo. To Juan. To Johan.

I lastly want to thank my family. Mom, dad, Silvia and Tomasa. Thank you for always making me feel safe and loved (even when you do it in special ways, like eating my hair).

To all of you, I sincerely thank you.

Patricia Enríquez Calzada
Delft, January 2021

Contents

Abstract

Quantitative MR obtains images containing meaningful physical and chemical characteristics of the tissues, allowing for comparison between patients. The procurement of this type of MRI through specific correction acquisition sequences tries to eliminate the maximum possible standardization problems that the technique inherently possesses. However, this does not provide perfect results. Deep learning could be used as a tool to finalize this harmonization after acquisition is finished. In this project, a CycleGAN is used to transfer the style of one specific MRI scanner into the images of another specific scanner.

The aim is to achieve a better harmonization that eases posterior image analysis and to possibly solve other issues like hardware obsoleteness. Inspired by the literature, which has never applied image style transfer to this type of images, different methodologies are tested. Some have been applied to other MRI modalities, like an extra similarity measure in the loss function. One novel implementation is tested. It consists on an extra discriminator that tries to reinforce the classification of the original and fake/generated images of one scanner as one class, as opposed to the class formed by the original and fake images of the other scanner.

Validation is based on visual inspection; histogram comparison; SSIM, NRMSE and correlation measures and CNN classification of the generated images (in a network trained to distinguish the origin of the scanners).

Experiments show the inconclusiveness of the possibility to apply the general CycleGAN loss function to a set of images with such visual similarity. A further study on the specific details that a discriminator uses in order to classify images as coming from a given scanner could help design a specific loss which's optimization generates the desired results

1 Introduction

Many efforts have been aimed towards solving the standardization problems that conventional MRI inherently presents. Quantitative MRI facilitates measuring physical and chemical properties of the imaged tissues which is lacking in common qualitative MRI. However, there are still standardization problems caused by hardware and processing algorithms that affect the images differently between scanners. Artificial intelligence could be cheap and efficient instrument to fill in these gaps.
This project aims at using deep learning to create a network such that, after scanning one patient in a scanner, images can be passed through that network and they will look (and contain the characteristics) as if they had been taken in another scanner. This targets to be a practical solution to the mentioned harmonization issues that are usually faced in day to day medical research.
This chapter includes a brief introduction to MRI and the basics of deep learning necessary to understand the following chapters. More detail can be found in the full literature study in Annex A.

1.1. Magnetic Resonance Imaging (MRI)

Nuclear magnetic resonance (NMR) is a frequently used medical imaging technique that allows for high contrast between different soft tissues. A crucial advantage over x-ray imaging is that it exploits the magnetic properties of the nuclei of atoms by applying a strong magnetic field instead of using ionizing radiation (Brunner & Ernst, 1979)(Kransdorf & Murphey, 2000).
To acquire an image, clinical MRI focuses on hydrogen atoms, which constitute 63% of the body content. The spins of these particles can be conceived as tiny magnets that tend to align with a scanner's main magnetic field direction. As such, a net magnetic moment results from large population of such spins in rest situation. This net magnetization is disturbed through the application of radiofrequence magnetic fields (Joseph P. Hornak, 1996). As a consequence, a signal is generated while the spins recover their resting position after the disruption. The two most common types of images we can find in clinical applications are called T1-weighted and T2-weighted images respectively. T1 weighted images capture the longitudinal nuclear spin magnetization (M_Z) at a particular timepoint during recovery, while T2 weighted images represent decay of transverse magnetization ($M_{x,y}$) (Andrew J. Taylor et al., 2016)(Chavhan et al., 2009). Figure 1 helps understand the two modalities.

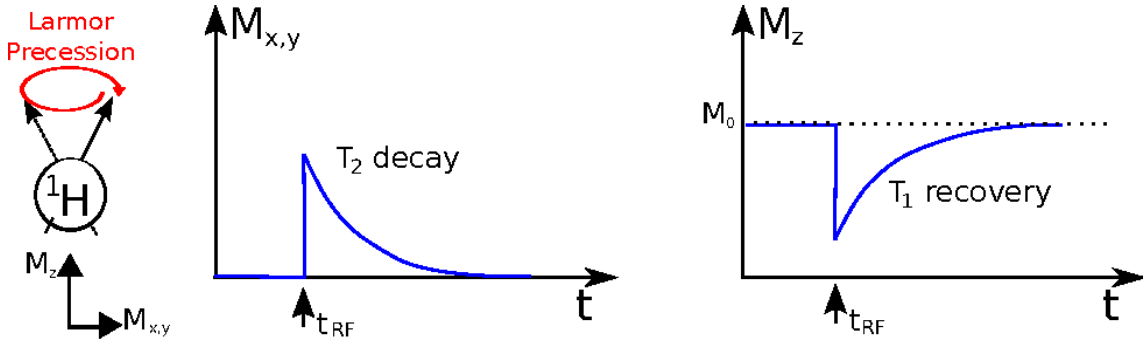


Figure 1. From left to right. A) Sketch of the precession of a hydrogen atom. B,C) Plots of the transverse magnetization (Mx,y),and the axial magnetization (Mz) after an excitation RF pulse is applied. Figure from (Murphy, 2011)

The signal of weighted MRI acquisition is affected by factors that are intrinsic to the tissue but also factors that are experiment-dependent, i.e. spin-spin interactions and sequence parameters (Pierpaoli, 2010). The versatility of MRI derives from its ability to highlight different tissue properties. An important issue, however, is that the signal intensity depends on physical effects that are hard to predict and control such as the strength or homogeneity of the static magnetic fields, tissue susceptibility, etc (Fullerton, 1987) (Jackson et al., 1997). Consequently, the conventional weighted maps provide signal values with no direct physical or chemical meaning and that cannot be compared across tissues or patients (Bergeest & Jäger, 2008)(Pierpaoli, 2010). Alternatively, Quantitative MRI is the procurement of maps of certain physical or chemical characteristics that are more easily reproducible and can be compared between subjects and experiments (Pierpaoli, 2010).

Particularly, Synthetic MR (*SyntheticMR*, n.d.) has commercialized software that is already in clinical use, which can create quantitative T1 and T2 maps. It does so by fitting a model to the relaxation of several, quickly acquired, weighted images. This allows for system imperfections such as magnetic field inhomogeneities to be taken into account (Warntjes et al., 2008). However, several publications have demonstrated that there is still a percentage of variability (Hagiwara et al., 2017) (Deoni et al., 2008) (Bauer et al., 2010) (Hagiwara et al., 2019) in the maps obtained with different scanners.

Artificial intelligence-based image analysis techniques are known to be sensitive to the typical variabilities present in MRI. A study by (AlBadawy et al., 2018) concludes that training models with data from different institutions than the data used to develop the algorithms produces a dramatic deterioration in model performance. Their main hypothesis is that the differences in scanners and their parameters are the main reason behind this. Eliminating these variances could enhance numerous post-acquisition procedures. The methodology to achieve this is generally referred to as scanner harmonization.

A study of the literature is initially done to find the optimal deep learning way of tackling the harmonization of parametric maps of MR images coming from different scanners. This will be done by implementing style transfer from one scanner into the other. To understand style transfer, a basic introduction of machine learning is done next.

1.2. Machine Learning

The big challenge of Artificial Intelligence consists on being able to solve problems that are intuitive for humans but hard to describe formally (Goodfellow et al., 2016). In particular, machine learning is the field of research that studies algorithms to detect a patterns by generalizing from given example data (Domingos, 2012). Within this field, Deep Learning is the specific use of neural networks to solve such issues. These are structures that do not need a manual extraction of the features; instead, they receive raw data and learn relevant features automatically. The learning process is achieved by building complex constructs by a combination of simpler ones (LeCun et al., 2015).

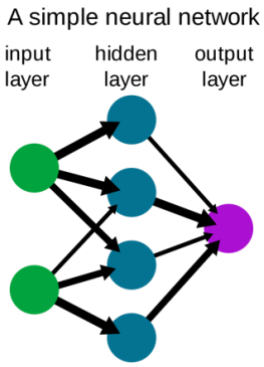


Figure 2. A simple neural network

Neural networks are a succession of layers composed by nodes as depicted in Figure 2. Information flows through the nodes that contain parameters defining a linear function f(x)=ax+b in which a and b are the parameters to adjust during training and x is the input information to the nodes. A further explanation on neural networks and, in specific Convolutional Neural Networks (CNNs), the most common architecture to deal with images, can be found in Annex A.

Generative Adversarial Networks (GANs). GANs are based on the simultaneous training of two neural network models. There is a Generator, G, which tries to approach the data distribution of the training set from a random noise input, and a Discriminator, D, that generates a value expressing the probability that two inputs are drawn from the same distribution (e.g. images from the same scanner) (Goodfellow et al., n.d.). Both networks are trained simultaneously, in competition to perform better than the other and using each other's information. Convergence is achieved when the discriminator cannot tell the difference between real and generated samples. To put a practical example, if a GAN is trained with images of faces, it will learn the characteristics of a face and generate a randomized combination of those features to build a fake face that the discriminator believes is a real one.

Based on this concept, several algorithms were proposed such as Pix2Pix (Isola et al., 2018), Cycle-GAN(Zhu et al., 2020), StarGAN (Chen et al., 2018) or MedGAN (Armanious et al., 2020), Disco-GANs (T. Kim et al., 2017), Fila-sGAN (Zhao et al., 2017), amongst others. These networks have been studied for many purposes in medical imaging like artifact elimination (Wang et al., 2018)(Liao et al., 2018), segmentation (Dong et al.,

2018) or, the most relevant application for this project, image translation between scanners (Ben-Cohen et al., 2018)(Yang et al., 2018)(Dar et al., 2018). Below, a brief summary is given for CycleGAN, designed for unpaired image style transfer and the chosen model for the task of this project.

CycleGAN. A CycleGAN network is able to consistently map back and forth the distributions of two sets of unpaired images. That they are unpaired means that there isn't a need for pixel to pixel correspondence between the pairs of images used for training. The main objective of the networks is to optimize Equation 1:

$$G^*, F^* = \arg min_{G,F}\, max_{D_X D_Y} \mathcal{L}(G, F, D_X, D_Y) \qquad \text{Equation 1}$$

The loss function included in Equation 1 is detailed in Equation 2, containing two aspects: loss functions corresponding to GANs that generate data in either direction (i.e. from X to Y and vice versa), Equation *3*, and the cycle consistency loss, Equation 4.

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F) \qquad \text{Equation 2}$$

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[1 - \log (D_Y(G(x))] \qquad \text{Equation 3}$$

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\| F(G(x)) - x \|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\| G(F(y)) - y \|_1] \qquad \text{Equation 4}$$

A set of two discriminators (D_X, D_Y) and two generators (G, mapping Y from X; F, mapping X from Y) are used to map from one distribution to the other and vice versa. Here, X and Y are the two distributions. What Equation *3* describes in the first part is maximization of the probability of D_Y to identify correctly if a sample is drawn from the Y distribution. At the same time, the second term tries to maximize the probability that the same discriminator is tricked into classifying G(x) (a fake, generated Y) as Y. Therefore, the Generator, G, is trained to imitate Y better and better when it receives an input from X. This is only half of the optimization though, as can be seen in Equation 2, since this is repeated in either mapping: one time in each direction (So D_X should be able to identify distribution X and F(y) should try to imitate x as good as possible). Finally, the cycle consistency, described in Equation 4, is used to ensure that forward and backward mapping are consistent.

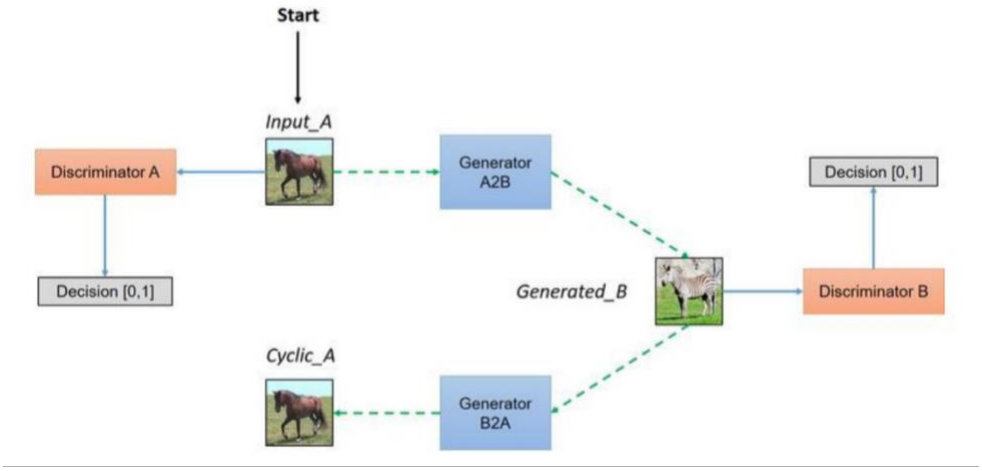


Figure 3.One-way CycleGAN schematic. Picture taken from (Zhu et al., 2020)

This is done by doing a pixel to pixel comparison of a sample of Y and a fake y generated from mapping to distribution X and then back to Y through F and G respectively (and the other way for x). This encourages a correspondence between the input and the output and limits the possible mappings. All terms are added up in a weighted sum as can be seen in Equation 2, where λ can define the importance of the cycle consistency loss as desired.

Finding the equilibrium for such complex optimization requires a large amount of data. A sketch of half of the training process of a CycleGAN is shown Figure 3. The same scheme is trained with an input “B” on the same Discriminators and Generators but switched.
Sometimes, an extra loss can be added to the CycleGAN to ensure that the output does not turn out too different from the input, for example, in terms of dynamic range. This is called the Identity loss and it is described in Equation 5. It consists on the minimization of the pixel to pixel comparison of a sample from X and its mapping to distribution Y and vice versa. It can be added as another term to Equation 5 with its own weight.

$$\mathcal{L}_{identity}(G,F) = \mathbb{E}_{y\sim p_{data}(y)}[\| \mathrm{G(y)} - \mathrm{y} \|_1] + \mathbb{E}_{x\sim p_{data}(x)}[\| \mathrm{F(x)} - \mathrm{x} \|_1] \qquad \text{Equation 5}$$

1.3. State of the Art

Image style transfer has been widely used in medical imaging. However, there is not an extensive literature in relation to MRI and, to our knowledge, it is nonexistent for the specific case of Quantitative MRI. In (Modanwal et al., 2020), the authors targeted harmonizing breast tissue with Dynamic contrast-enhanced (DCE) MRI images from a Siemens scanner to a GE Healthcare scanner. To do so, they used a Cycle GAN in which the discriminator is a PatchGAN (type of discriminator that makes classification decision based on patches instead of the whole image) with a smaller field of view than the usual (70x70 pixels) to help maintain anatomy. To check the performance of the model, the authors did a quantitative assessment of the mean intensity values distribution on dense tissues before and after harmonization. Given that the scanners produce images that can be differentiated by eye (see Figure 4)., this validation is considered sufficient to prove a good harmonization.
In (Dewey et al., 2019), the authors aimed at harmonizing images of different MRI modalities between protocols and scanners. Prior to using any deep learning tools, supervised pre-processing algorithms were applied to homogenize and align the images. Subsequently, for the harmonization purpose a U-net is used, which is a standardized CNN originally designed for image segmentation (Ronneberger et al., 2015). After harmonization, a second step is required for noise reduction. For validation the Mean Structure Similarity Index (SSIM) and Mean Absolute Error (MAE) criteria were used

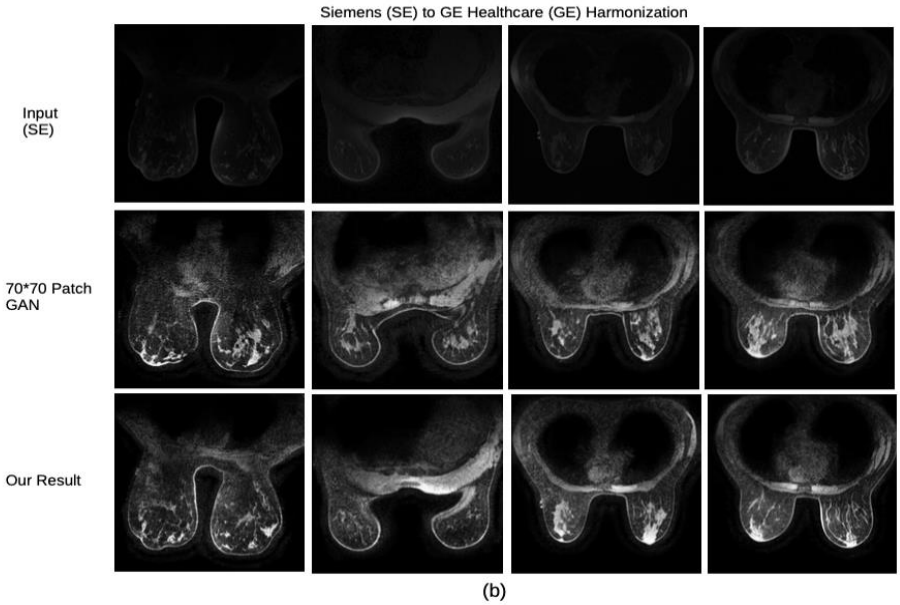


Figure 4.Proposed Image Harmonization results for Siemens to GE Healthcare in (Modanwal et al., 2020), Top row shows input image from Siemens scanner, Middle row shows results with the 70x70 field of view discriminator (standard) PatchGAN while the last row shows

Another publication (Xiang et al., 2018) attempted to harmonize data from a 3T and a 7T scanner. To do so, a CycleGAN with a modified generator for faster performance was used. A structural dissimilarity loss was added to the loss function that was optimized during training of the network. This was aimed towards

preserving anatomical details. Validation was performed based on three indicators: peak signal to noise ratio (PSNR), SSIM and Normalized Mean Square Error (NMSE). The authors concluded that satisfactory harmonization is achieved through the analysis of these statistical measures.

1.4. Objectives

The ultimate goal of the project is to optimize the application of a CycleGAN for the harmonization of quantitative MRI images. These quantitative maps can be used to subsequently derive any desired weighted image desired. This means that by harmonizing the original quantitative map, the derived weighted images would already be harmonized, avoiding the possible need to train networks for each specific weighted image harmonization between scanners. As far as we know, no attempt to apply this method to further harmonize corrected quantitative images has been tried in the literature. Therefore, the specific structure and training have to be designed from scratch as all the examples in the literature show the need for different methodological approaches depending on the image characteristics.

This will require to study the characteristics of the quantitative and corrected images that are currently being acquired in hospitals by the scanners in the market, to do experiments with different methodology to find the optimal CycleGAN for quantitative image harmonization and then to study the possibility to use this harmonization in order to overcome hardware unavailability (broken, busy, different magnetization scanners) and other common situations that could benefit from said process in hospitals.

2 Methods

A series of experiments was carried out to test the performance of a harmonizing CycleGAN implementation into the available images. Programming was done in Python (code can be accessed in: https://github.com/patham4/masterthesis) and experiments were run on the GPU cluster of the Biomedical Imaging Group (BIGR) group at the Erasmus Medical Center in Rotterdam.

2.1 The Data

The dataset was formed by the brain scan of 12 healthy subjects on two 1.5T GE Healthcare machines of different type: SIGNA and Optima MR450w. Many different acquisition protocols were run for every subject in each scanner. For the simplification, only the T1 quantitative maps were used in the experiments. Observe that any T1-weighted image can be synthesized later from this mapping. Therefore, harmonization of the quantitative maps would allow to generate already harmonized synthetic images.
For each of the subjects there are 27 axial brain slices. They are so-called 2D scans, having a thickness of 4 mm and a resolution of 256 by 256 pixels. 10 random subjects were used for training (270 images per scanner) of the CycleGAN while 2 were kept for testing (54 images per scanner). As slices are thick, no perfect correspondence exists between the same slice for the same patient in the two scanners, this lead to the need to apply unpaired image translation.

2.2 Pre-processing

A rigid registration was the only processing performed for some of the experiments in order not to modify the natural characteristics of the images. Registration was done to align the pairs of images that belong to the same patient slice in each scanner. The objective of this transformation was to help the network focus on details other than any possible anatomical variations when learning to transform from one scanner to the other. For this, the mutual information registration metric was used. Only shifts in lateral and vertical direction were applied (rotation is dismissed to avoid possible misalignments due to the symmetry of the brain and the selected registration measure) to modify both images equally and avoid bias, both were transformed to meet in the middle point. Interpolation was performed using cubic B-splines.
Images were normalized to the [0,1] range by dividing by the maximum value and eliminating every value below 0. Any value below 0 was always encountered in the background and therefore doesn't generate any information losses.

2.3 Image variability assessment

A necessary initial step was to do an assessment of the variability that the literature claims there is between quantitative images. For that, several techniques were applied:

- Visual inspection
- Histogram comparison
- A comparison based on the structural similarity index (SSIM) and Normalized Root Mean Square Error (NRMSE) between each pair of images
- A classifier that recognizes images coming from a particular scanner. Therefore, a very simple CNN classifier was built with 5 layers of 16, 32, 128, 256 and 512 filters with 3x3 kernels and in between pooling layers of 2x2. The training included data augmentation by horizontal flip, 10 epochs, batch size of 20 and a train-test split of 80-20% on the whole dataset of patients (the 12 of them, so 516 training images and 130 testing). The experiments were done both on the full image and on cropped versions since the background should not have any meaningful information and could be a confounding factor.
  Attention maps were derived using Grad-CAM on the last layer of the classifier to see what features are important to distinguish between the scanners (Grad-CAM is a technique by (Selvaraju et al., 2020). Grad-CAM computes the gradient of the loss function for a class (in our case classes are either scanner A, or 0, or scanner B, or 1). Going backwards from that output into the network, the neuron importance weights can be calculated, obtaining a map of feature dominance for such an input class).

2.4 CycleGAN experiments

An initial, base basic model was built from the base of a CycleGAN (Brownlee, 2019b) and Pix2Pix (Brownlee, 2019a) implementations. This base model included a 70x70 Patch GAN (a type of discriminator for GANs that takes a classification decision on patches of the image of size 70x70 pixels) as discriminator and a U-Net generator. This generator is chosen for its simplicity and supported by (Fetty et al., 2020) in a study that claims that there are no significant differences in results of CycleGAN for MRI applications when using U-Net, ResNets (the original CycleGAN generator) or Dense-Nets.

Network architecture details:

-PatchGAN: This discriminator is designed such that it does a prediction on many patches of a given size to make the discrimination decision. This is supposed to allow more accurate predictions. It consists on a CNN of 5 layers of 64, 128, 256, 512 and 512 filters with 4x4 kernels and 2x2 strides. Batch normalization and leaky ReLu activations appear in each of these layers. Then, a flattening layer with a sigmoid activation. Patches are 70x70 because of the size of the input, 256x256 pixel images, and the combination of the kernel size and the applied stride.

-U-Net: This widely used and well-known generator network contains a downsampling and then upsampling path as can be seen in Figure 5.

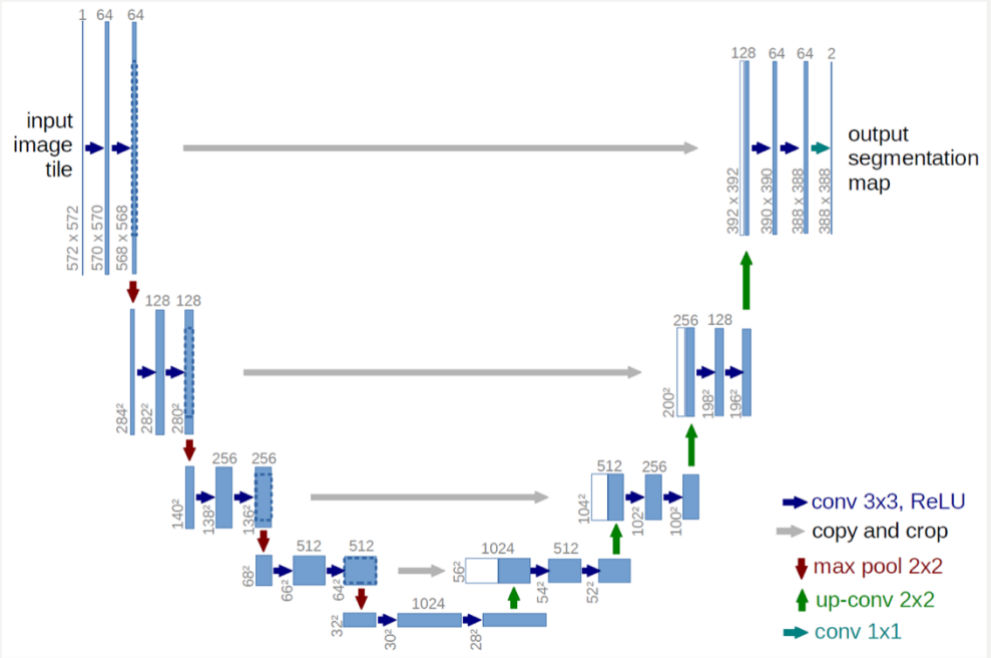


Figure 5. U-Net architecture (Ronneberger et al., 2015)

Many tests were carried out to optimize results. A summary of the main modifications done to the basic CycleGAN architecture are collated in Table 1. More experiments were performed (one sided label smoothing, different generator or discriminator, etc.) but the main findings derived from the different loss definitions. For every new experiment, only one variable was changed to make sure there are no confounding factors as to what is producing the changes in results.

Table 1. Main CycleGAN experiments performed. Highlighted in blue are the main changes from each previous experiment can be found.

	Batch n	Epochs	Discriminator	Generator	Loss function	Optimization details	Data augment	Objective
Original CycleGAN	1	200	70x70 Patch GAN	Res-Net Instance normalization	1x Discriminator loss 10x Cyclic loss (5x Identity loss)	Adam, lr=0.0002		

Experiment 1	20	50	70x70 Patch GAN	U-net	1x Discriminator loss 10x Cyclic loss 5x Identity loss	Adam, lr=0.0002	No (Training with unpaired batches)	Reference CycleGAN as described in the literature
Experiment 2	20	50	70x70 Patch GAN	U-net	1x Discriminator loss 10x Cyclic loss 1x Identity loss	Adam, lr=0.0002	No (Training with unpaired batches)	As images are very close already, reduce the condition to keep output close to input
Experiment 3	1	20	70x70 Patch GAN	U-net	1x Discriminator loss 10x Cyclic loss 1x Extra discriminator loss	Adam, lr=0.0002	Yes (Training with unpaired batches)	Favor the validation objective of making A-fakeA and B-fakeB the two classes distinguishable by a classifier
Experiment 4	5	20	70x70 Patch GAN	U-net	1x Discriminator loss 10x Cyclic loss 1x SSIM loss	Adam, lr=0.0002	Yes (Training with unpaired batches)	Test out the performance of this loss which also favors a possible validation
Experiment 5	5	20	70x70 Patch GAN	U-net	1x Discriminator loss 10x Cyclic loss 5x SSIM loss	Adam, lr=0.0002	Yes. Registration (Training with batches of corresponding slices of patients)	Make it easier for the network to focus on changes that do not have to do with anatomy or registration
Experiment 6	Experiment 5 but trained to harmonize unregistered T1 quantitative maps and some T2 maps from one of the scanners. This experiment is aimed at checking whether the CycleGAN performs properly in a situation in which the images to harmonize look very different.							

During training in each experiment, the followed criteria to save specific epochs was applied:

-Every 5,10 or 20 steps

-Whenever the loss of the discriminator on fake images was more than 0.3

-Whenever the loss of the generator is lower than 1 on both models (B to A and A to B)

Given the amount of data and the number of experiments to analyse, one single optimal training epoch was selected for the validation of each experiment. This optimal epoch was selected among the ones saved in the second half of training. Between images with good visual results, the epoch selected was a random one that fulfilled the following requirements:

-Generator loss was balanced for both models (AtoB and BtoA) and lower than 1.5

-Discriminator loss percentage was good (0.01-0.15) for real images and, if possible, higher than 0.1 for fake image classification.

In (Xiang et al., 2018), an extra loss is added in order to facilitate detail attention during mapping based on structural similarity (SSIM). This loss, defined in Equation 6, tries to minimize the dissimilarity between a distribution and the forward and backward mapping towards that same distribution (so between X and F(G(X))

and between Y and G(F(Y))). Inspired by this encouragement during training of a characteristic that will later be used for validation, a new method was proposed to favor validation based on passing the generated images through a classifier.

$$\mathcal{L}_{DSSIM}(G,F) = \mathbb{E}_{x\sim p_{data}(x)}\left[\frac{1-SSIM(x,F(G(x)))}{2}\right] + \mathbb{E}_{y\sim p_{data}(y)}\left[\frac{1-SSIM(y,F(G(y)))}{2}\right] \quad \text{Equation 6}$$

For this, an extra discriminator was added to the training loss in the same way the cyclic and the identity loss are added, using a weighting factor (see Equation 2). This extra discriminator reinforces the condition that x and artificially generated x images are considered the same class (as opposed to another class formed by y and artificially generated y images) (see Equation 7).

$$\mathcal{L}_D = \mathbb{E}_{x,y}\left[\log\left(D(x,y)\right)\right] + \mathbb{E}_{x,y}\left[\log 1 - \left(D(F(y),G(x))\right)\right] \quad \text{Equation 7}$$

3 Results

In this chapter, the most relevant results obtained during the initial variability assessment and the experiments described in the chapter before are presented.

3.1 Initial variability assessment

Observe that all images are T1 quantitative maps and corrected for inhomogeneities. A random sample can be seen in Figure 6.

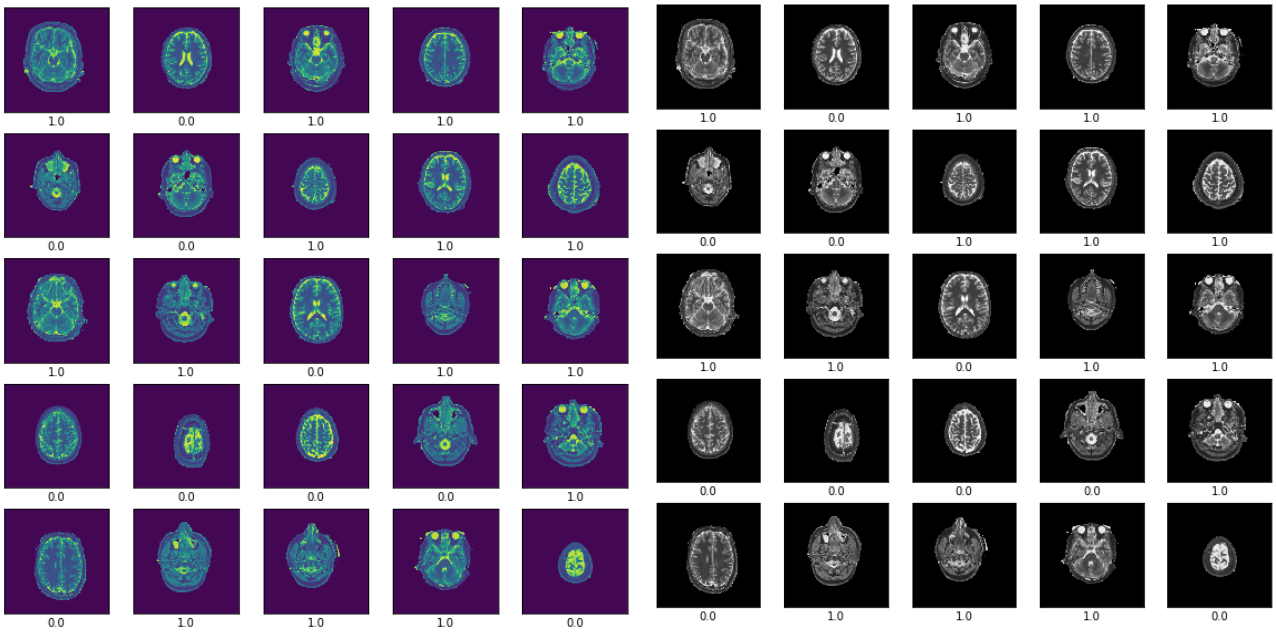


Figure 6. A random selection of images from the training set coming both from one or the other scanner (labels 0 (scanner A) and 1 (scanner B)). Both sets, left and right, are the same but represented differently. Original images are in grayscale. However, color allows spotting differences more easily, therefore it will be used in result representation.

A comparison on the average histograms for both scanners before and after registration can be seen in Figure 7. Here, the original range of values (from 0 to 2000) can be observed; later on, the analysis is done after a normalization to [0,1] as it is a more appropriate way to input the information into neural networks. Strong similarity between the two sets can be appreciated.

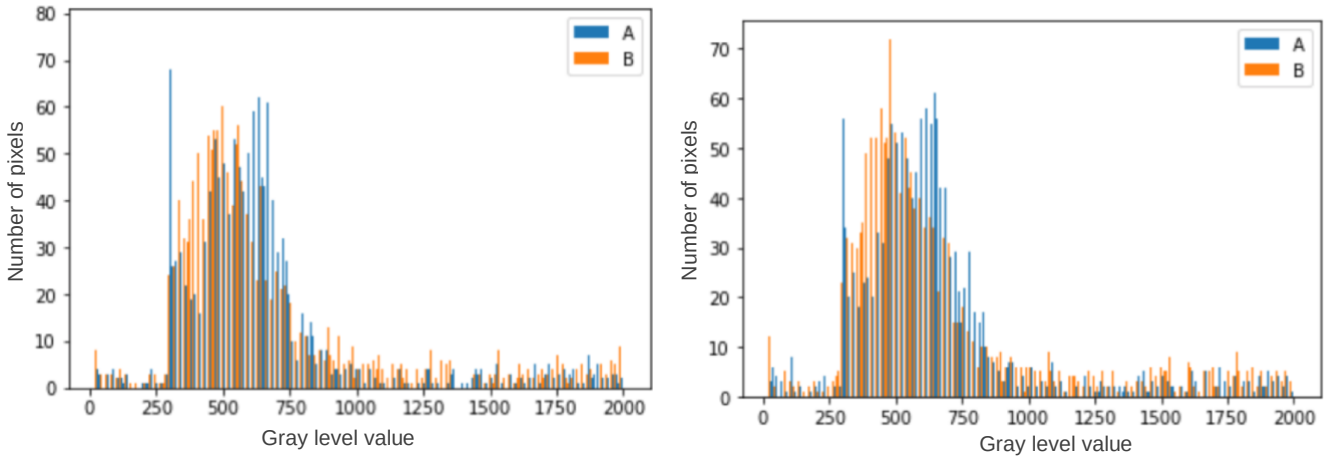


Figure 7. Comparison of average histograms before normalization (500 bins) of all images coming from scanner A (or 0) and scanner B (or 1): before registration (left) and after registration (right)

In Table 2 the average of the SSIM and NRMSE for every pair of images coming from both scanners is given (i.e. the average of the differences as computed by these two measures between the same slice in the same patient scanned in each of the scanners).

Table 2. Average SSIM and NRMSE between the pairs of images coming from each scanner

	Before registration	After registration
$SSIM_{ave}$	0.80	0.81
$NRMSE_{ave}$	0.56	0.58

Lastly, the images are passed through the classifier designed to distinguish the scanner of origin of an input image. Table 3 shows the performance of this model on a random sample of 130 test images 5 times. This allows to check for performance consistency in the discrimination. It can be observed that accuracy keeps consistency before and after cropping to the area of interest. Also, it is observed to be higher before than after registration. Figure 8 shows some test classification examples on the last run.

Table 3.Test accuracy of classifier trained to distinguish between scanners before and after registration

	Full image		Cropped image	
	Test accuracy (Before registration)	Test accuracy (After registration)	Test accuracy (Before registration)	Test accuracy (After registration)
Run 1	0.90	0.72	0.93	0.81
Run 2	0.88	0.76	0.88	0.82
Run 3	0.86	0.48	0.85	0.70
Run 4	0.90	0.84	0.93	0.77
Run 5	0.90	0.76	0.88	0.66
Average	0.89	0.71	0.89	0.75

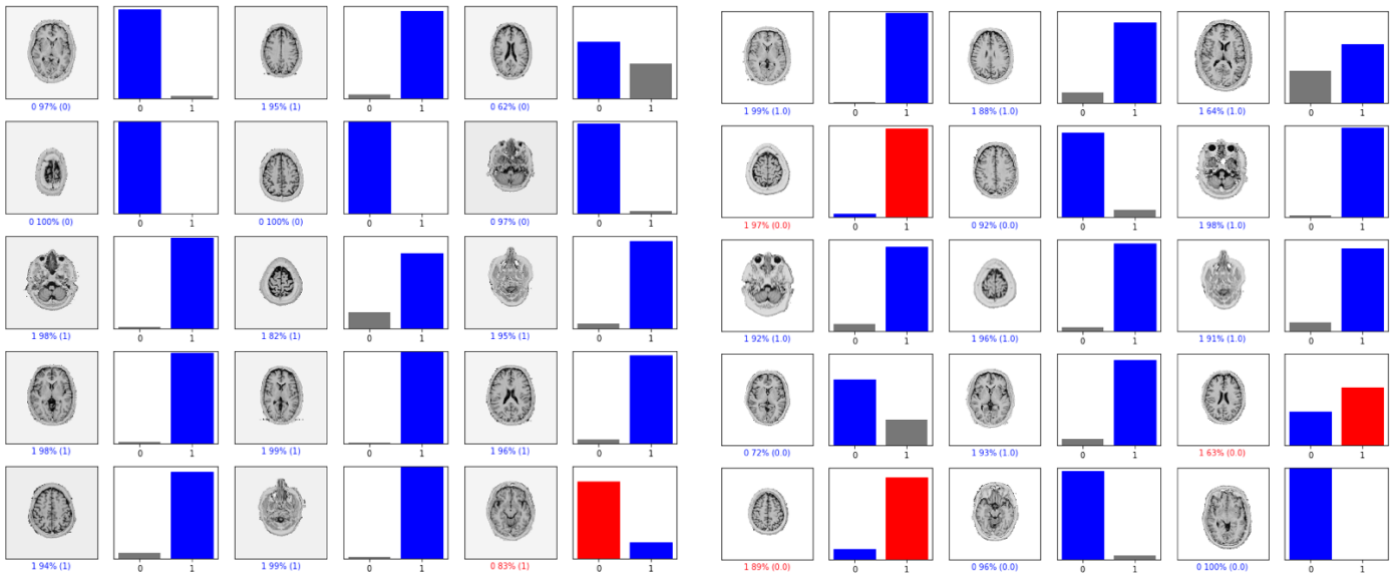


Figure 8.A random sample illustrating the test accuracy on some uncropped images in the classifier trained with the whole dataset identifying the of each image. Before registration is shown on the left and after registration on the right. The bars show the percentage with which the classifier thinks the image comes from scanner 0 (or A) or 1 (or B). In blue the correct classifications while red show incorrect ones.

Some examples of what the attention maps point to as decisive areas in classification of uncropped unregistered images in the last run (out of the 5 performed) can be seen in Figure 9. In Figure 10 can be seen the results for uncropped registered images. Figure 11 and Figure 12 show results for cropped images.

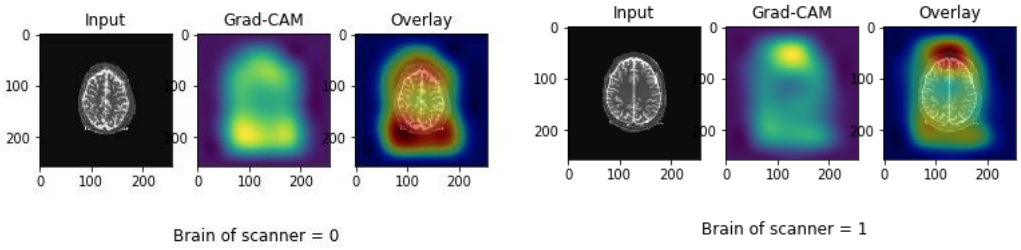


Figure 9. Random examples of attention maps for uncropped unregistered images. Image from scanner A or 0 to the left, image from scanner B or 1 to the right

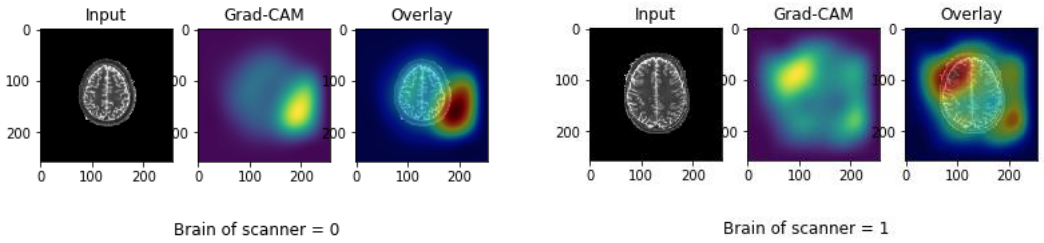


Figure 10.Random examples of attention maps for uncropped registered images. Image from scanner A or 0 to the left, image from scanner B or 1 to the right

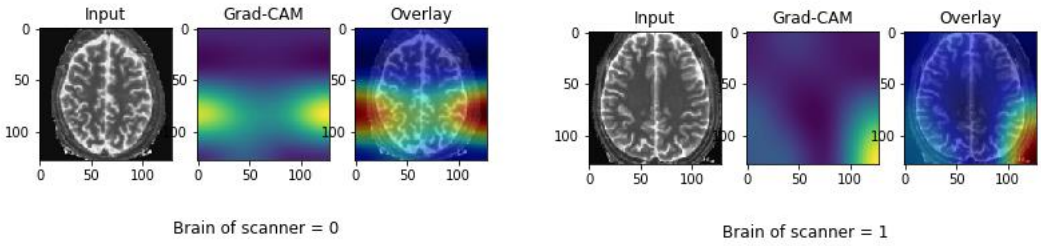


Figure 11. Random examples of attention maps for cropped unregistered images. Image from scanner A or 0 to the left, image from scanner B or 1 to the right

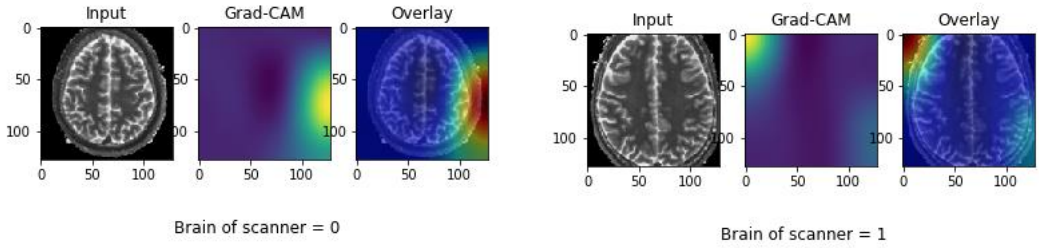


Figure 12. Random examples of attention maps for cropped registered images. Image from scanner A or 0 to the left, image from scanner B or 1 to the right

3.2 CycleGAN experiments

The following information is available for each experiment:

- Training loss numbers
- Training visual results
- Test visual results
- Test histograms and mutual information
- SSIM and MSNE values for test set
- Results of classification on test set generated images

All numerical results are presented first in tables. Then, figures will be presented separately for each experiment.

In Table 4 shows the training loss numbers for the epochs chosen to be tested. It can be checked that generator losses are low and that discriminators perform well in all cases.

Table 4.Numerical results of loss function during training on epoch chosen for testing

	epoch	D_A (A,fakeA)	D_B (B,fakeB)	D (A,fakeA,B,fakeB)	G (AtoB,BtoA)	D (AtoB,BtoA)	SSIM (AtoB)	SSIM (BtoA)	cycleLoss (AtoB)	cycleLoss (BtoA)
Exp1	720	0.14, 0.13	0.10, 0.15	-	0.69, 0.70	-	-	-	-	-
Exp2	720	0.02, 0.11	0.07, 0.08	-	0.62, 0.78	-	-	-	-	-
Exp3	3230	0.16, 0.17	0.03, 0.02	0.10, 0.15, 0.18, 0.23	1.48, 1.40	-	-	-	-	-
Exp4	1290	0.02, 0.04	0.07, 0.12	-	0.75, 1.19	0.31, 0.69	0.06, 0.05	0.06, 0.07	0.02, 0.02	0.02, 0.02
Exp5	1474	0.16, 0.16	0.07, 0.14	-	0.71, 0.78	0.17, 0.21	0.03, 0.03	0.04, 0.03	0.01, 0.01	0.01, 0.01
Exp6	1370	0.18, 0.13	0.04, 0.02	-	0.96, 1.32	0.28, 0.50	0.05, 0.04	0.07, 0.04	0.02, 0.01	0.01, 0.01

where:

- D_X(x,y): percentage loss obtained from inputting a batch of x and a batch of y in a classifier with the objective of distinguishing between real and fake of type X. A lower value indicates that the classifier is good at detecting the origin of the image, whichever it is. For convergence of the model, the loss of discriminating of fake images should get high.
- D (A,fakeA,B,fakeB): percentage loss of discriminating a batch of A, fakeA, B, fakeB in a discriminator made to distinguish A and fakeA as a class and B and fakeB as another class. Lower percentage means a better performance of the discriminator.
- G: loss function of the generator for A to B or B to A respectively. This includes the summation, according to the weights specified in the experiment description, of the following terms:
    - D: percentage loss of comparing the result classification of fakeB or fakeA respectively to the correct labels through mean squared error. Lower value indicates better performance
    - SSIM: DSSIM (structural dissimilarity) between the generated outputs and the original images. The lower the dissimilarity, the more similar the compared images are.
    - cycleLoss: Mean absolute error between the generated fake outputs and the original images. The lower the error, the more similar the compared images are.

Table 5 shows the correlation between all the images on the test set and the generated ones. It can be seen that the higher cross correlation can always be found between A-fake B and B-fake A.

Table 5.Cross correlation between sets of test images and generated images

	A vs B	A vs fakeA	A vs fakeB	B vs fakeA	Bvs fakeB	fakeA vs fakeB
Exp1	0.90	0.90	0.98	0.98	0.90	0.89
Exp2	0.90	0.88	0.96	0.94	0.90	0.86
Exp3	0.90	0.91	0.99	0.99	0.90	0.90
Exp4	0.90	0.89	0.96	0.96	0.88	0.87
Exp5	0.91	0.90	0.97	0.98	0.90	0.84
Exp6	0.65	0.77	0.78	0.75	0.54	0.60

Similarly, Table 6 shows the results for the average SSIM and NRMSE for all the experiments comparing images from the test set and the generated ones. For experiments 1 to 5 the higher SSIM and the lower

NRMSE can be found for the pairs A-fake B and B-fake A, while, for experiment 6, pairs A-fake A and B-fake B, hold those characteristics.

Table 6. Average SSIM and NRMSE for all possible combinations of images. A and B correspond to the test set images from the two scanners. FakeA and fakeB are the generated images from whose images.

	SSIM						NRMSE					
	A vs B	A vs fakeA	A vs fakeB	B vs fakeA	B vs fakeB	fakeA vs fakeB	A vs B	A vs fakeA	A vs fakeB	B vs fakeA	B vs fakeB	fakeA vs fakeB
Exp1	0.87	0.86	0.95	0.96	0.85	0.85	0.40	0.39	0.20	0.21	0.41	0.48
Exp2	0.87	0.83	0.92	0.90	0.84	0.83	0.340	0.44	0.24	0.30	0.40	0.49
Exp3	0.87	0.87	0.96	0.97	0.85	0.85	0.40	0.38	0.19	0.18	0.42	0.48
Exp4	0.87	0.84	0.93	0.92	0.84	0.82	0.40	0.42	0.24	0.26	0.42	0.46
Exp5	0.87	0.85	0.94	0.95	0.86	0.83	0.37	0.40	0.28	0.20	0.39	0.43
Exp6	0.74	0.77	0.75	0.77	0.91	0.75	0.87	0.58	0.86	3.69	0.88	0.87

Table 7 shows the classification accuracy of the generated images when passed through the discriminator for distinguishing between the two scanners. As in Table 3, classifiers showed high training accuracy. Tests with generated/fake images show a complete lack or very low accuracy in all cases.

Table 7. Test accuracy of the fake B generated images from each experiment passed through the classifier trained to distinguish the original scanner from which an image comes

	Test accuracy	Training of classifier accuracy	Epochs of training
Exp1	0.0	0.88	12
Exp2	0.0	0.88	12
Exp3	0.0	0.88	12
Exp4	0.06	0.88	12
Exp5	0.30	0.82	12
Exp6	0.0	1.0	1

Experiment 1

For the chosen epoch, representative results after training can be visually inspected in Figure 13 and Figure 14.

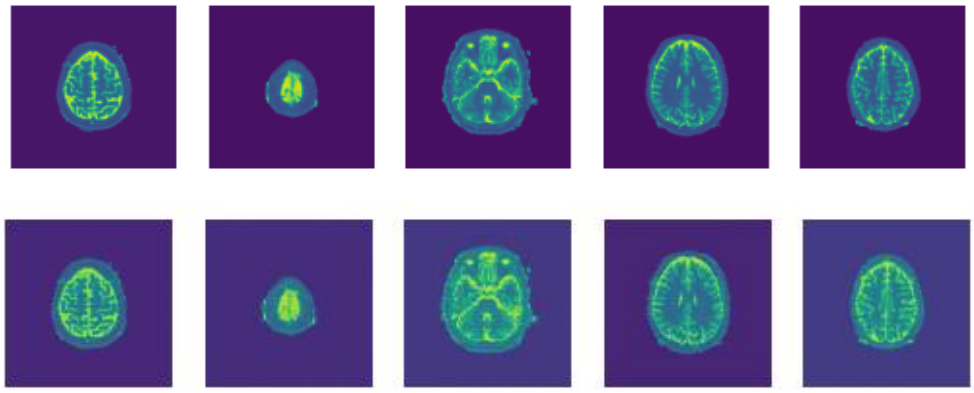


Figure 13. A to B training set results for Experiment 1. Top row is a random selection of real A images, bottom row is fake B generated from the images in the top row

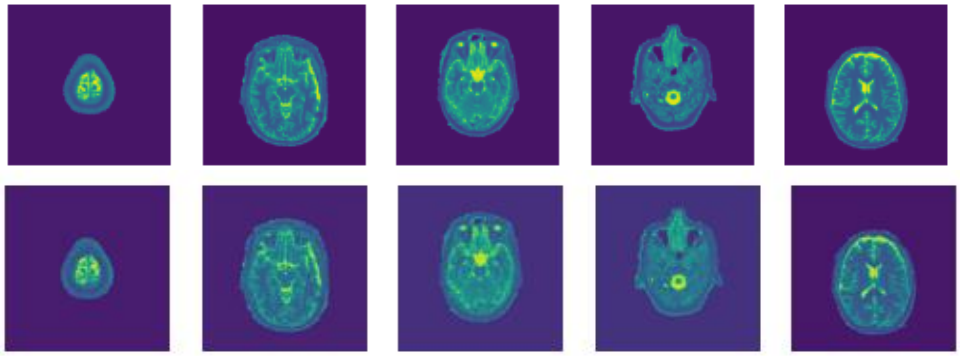


Figure 14. B to A training set results for Experiment 1. Top row is a random selection of real B images, bottom row is fake A generated from the images in the top row

After training, test set images are passed through the network to see its performance on unseen data. In Figure 15 a random choice of fake B images is shown in the top row; the corresponding fake A images can be seen in the lower row.

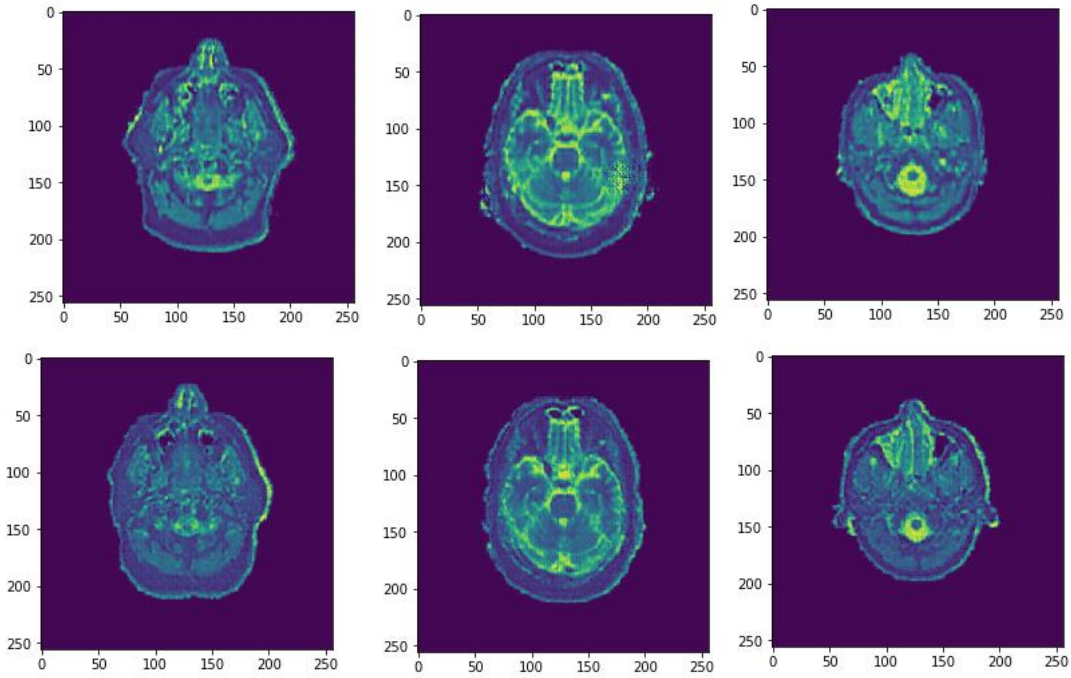


Figure 15. Fake B (top row) and corresponding fake A (bottom row) generated from test set

Lastly, an analysis of the histograms can be seen in Figure 16.

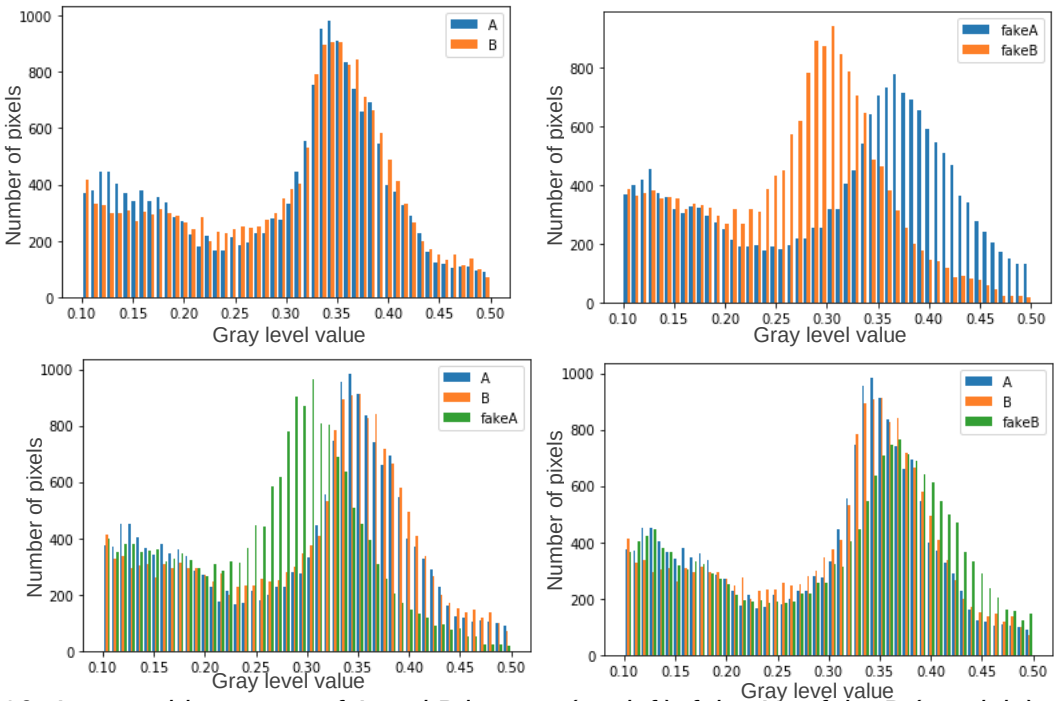


Figure 16. Average histograms of A and B images (top left), fake A vs fake B (top right), and for further classification of A, B and fake A (bottom left) and a comparative of A, B and fake B (bottom right)

Experiment 2

Representative results from the training set can be visually appraised in Figure 17 and Figure 18.

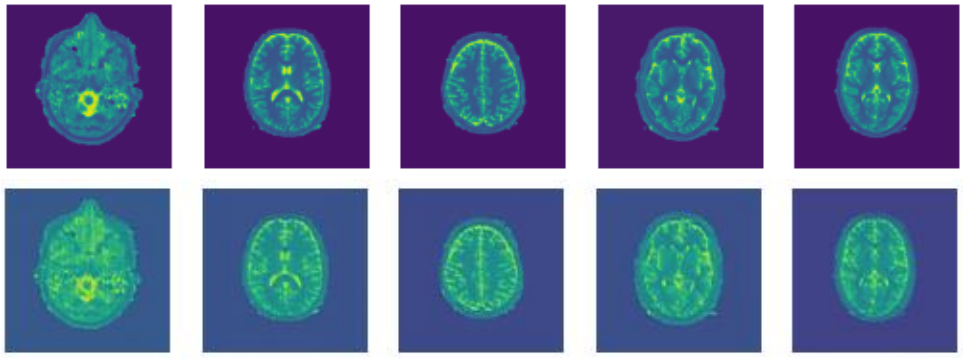


Figure 17. A to B training set results for Experiment 1. Top row is a random selection of real A images, bottom row is fake B generated from the images in the top row

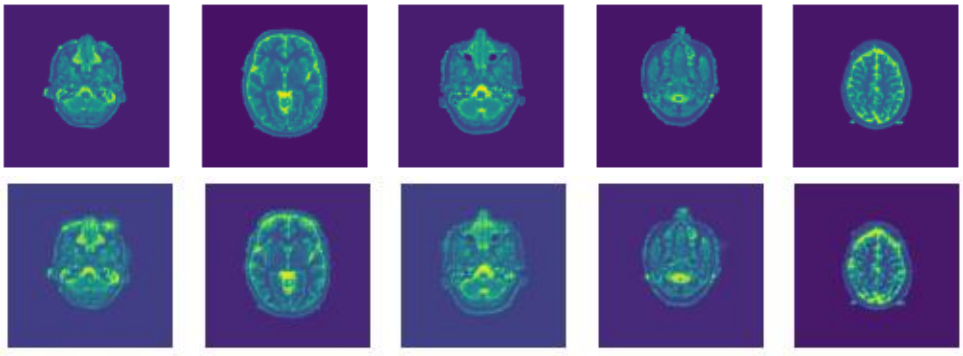


Figure 18. B to A training set results for Experiment 1. Top row is a random selection of real B images, bottom row is fake A generated from the images in the top row

In Figure 19 a random choice of fake B images is shown in the top row; the corresponding fake A images can be seen in the lower row.

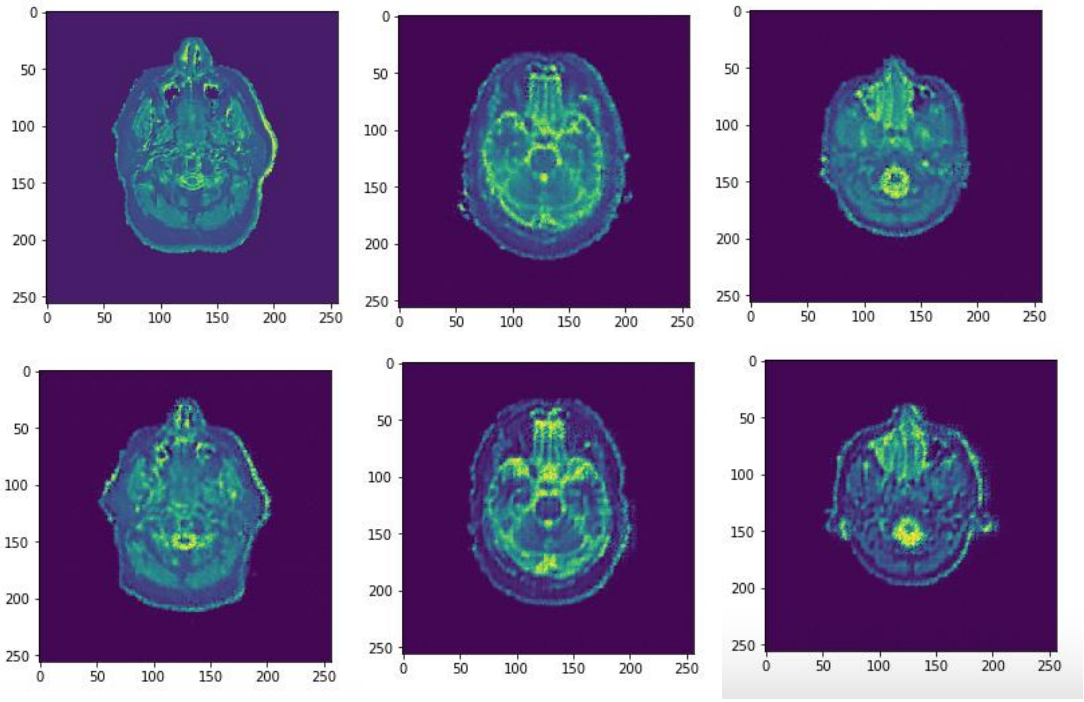


Figure 19. Fake B (top row) and corresponding fake A (bottom row) generated from test set

Lastly, an analysis of the histograms can be seen in Figure 20.

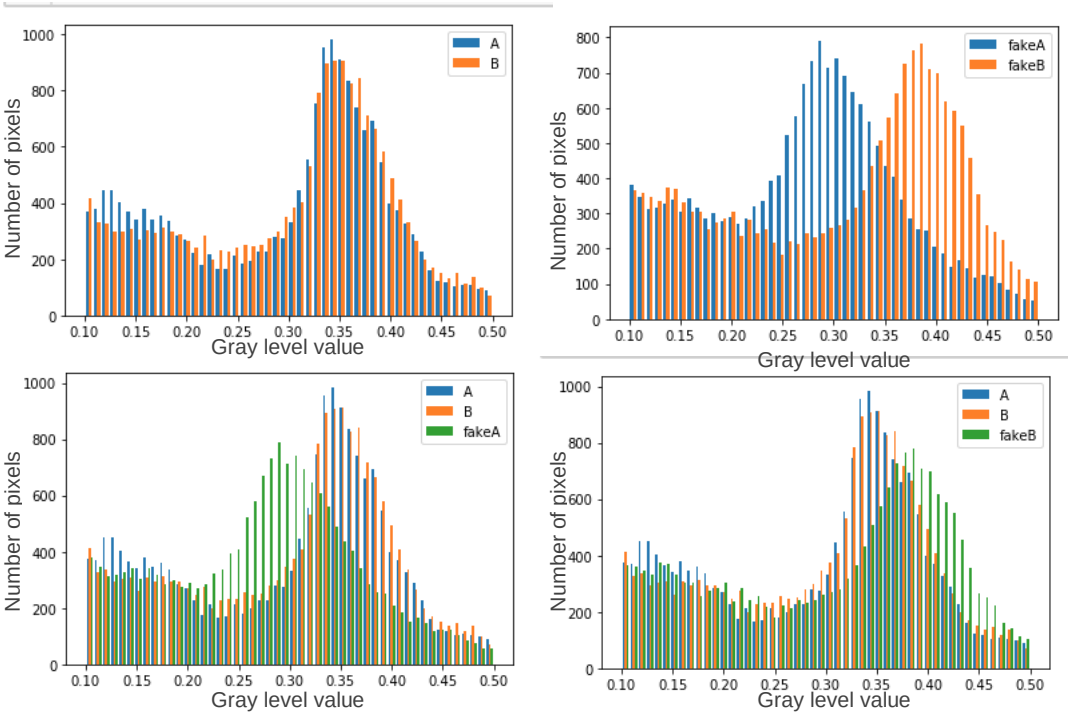


Figure 20. Average histograms of A and B images (top left), fake A vs fake B (top right), and for further classification of A, B and fake A (bottom left) and a comparative of A, B and fake B (bottom right)

Experiment 3

Results of training set can be visualized in Figure 21 and Figure 22.

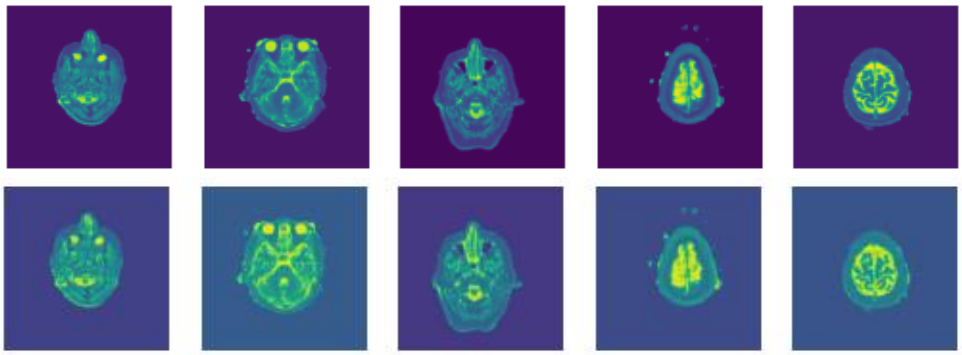


Figure 21. A to B training set results for Experiment 1. Top row is a random selection of real A images, bottom row is fake B generated from the images in the top row

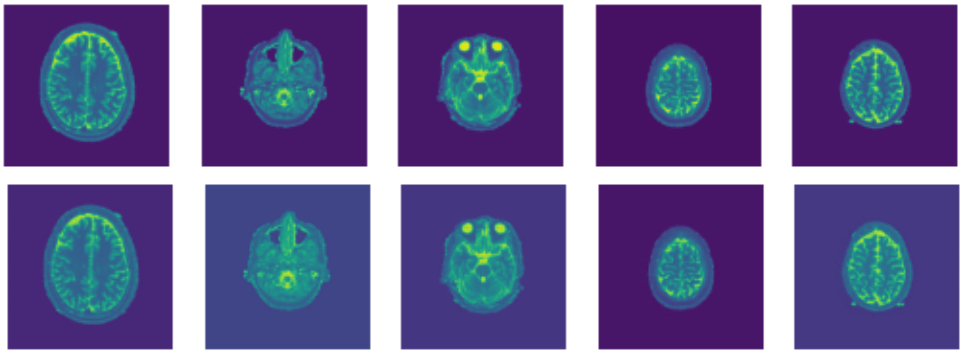


Figure 22. B to A training set results for Experiment 1. Top row is a random selection of real B images, bottom row is fake A generated from the images in the top row

In Figure 23 a random choice of fake B images is shown in the top row; the corresponding fake A images can be seen in the lower row.

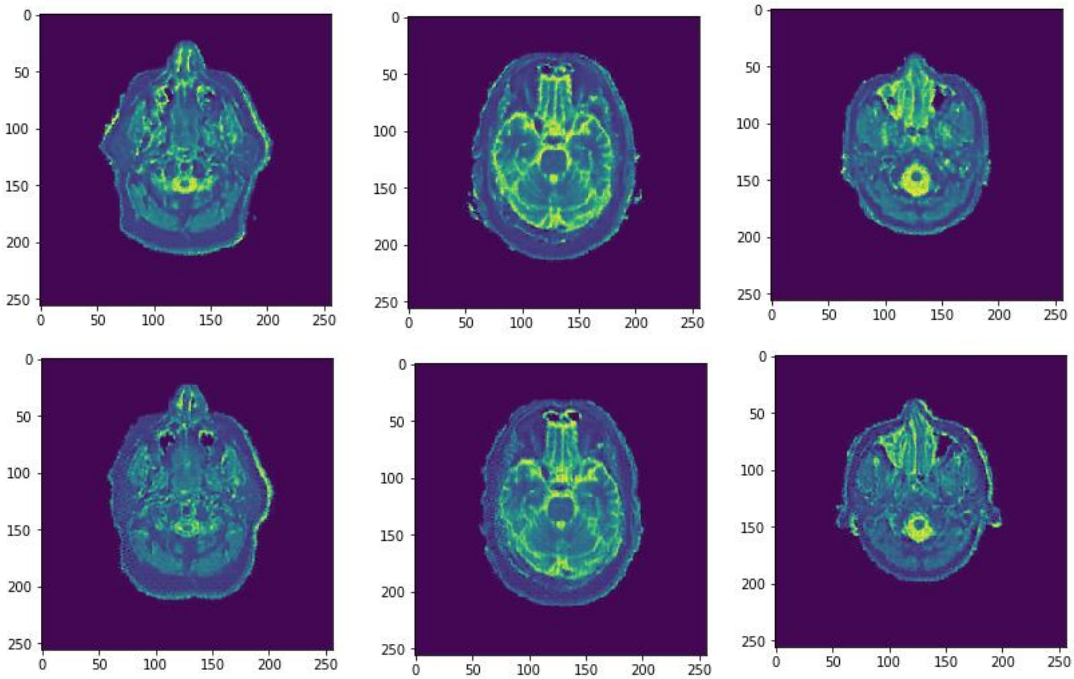


Figure 23. Fake B (top row) and corresponding fake A (bottom row) generated from test set

Lastly, an analysis of the histograms can be seen in Figure 24.

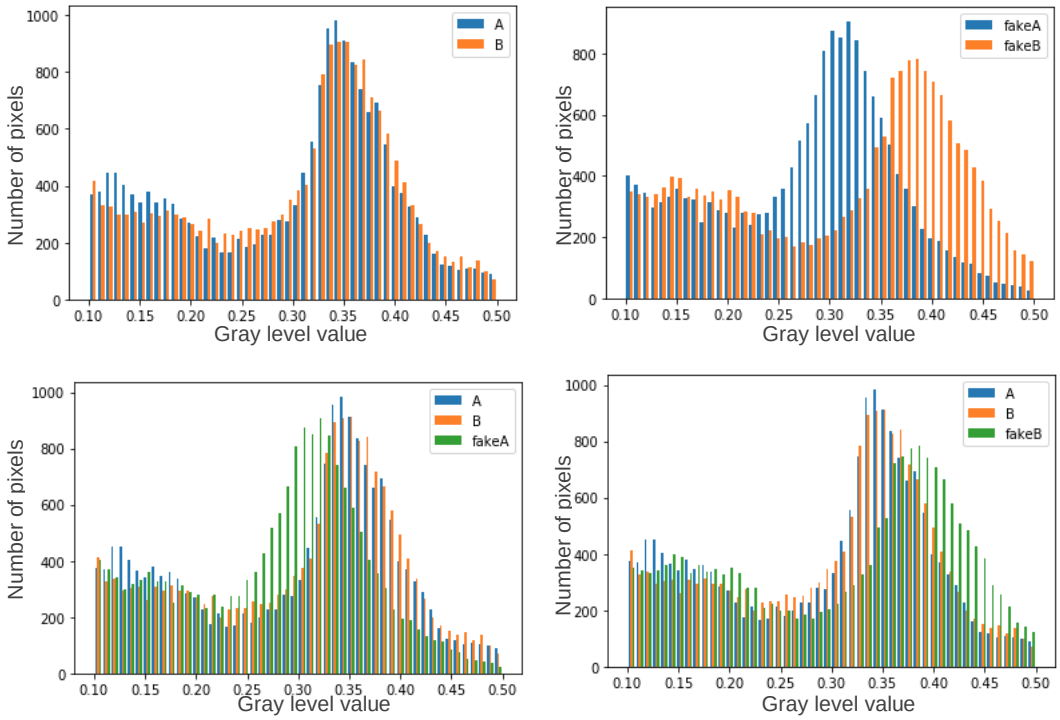


Figure 24. Average histograms of A and B images (top left), fake A vs fake B (top right), and for further classification of A, B and fake A (bottom left) and a comparative of A, B and fake B (bottom right)

Experiment 4

Representative results from the training set can be visually appraised in Figure 25 and Figure 26.

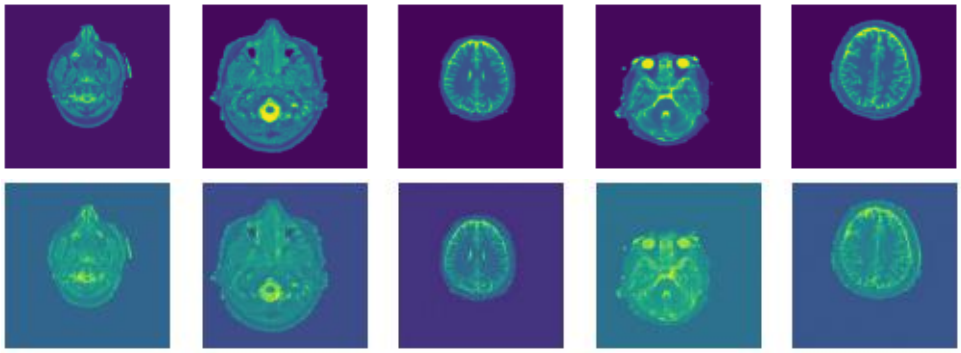


Figure 25. A to B training set results for Experiment 1. Top row is a random selection of real A images, bottom row is fake B generated from the images in the top row

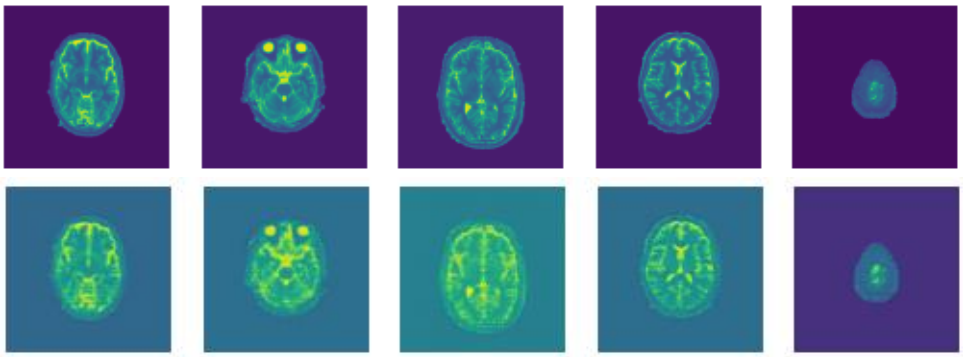


Figure 26. B to A training set results for Experiment 1. Top row is a random selection of real B images, bottom row is fake A generated from the images in the top row

In Figure 27 a random choice of fake B images is shown in the top row; the corresponding fake A images can be seen in the lower row.

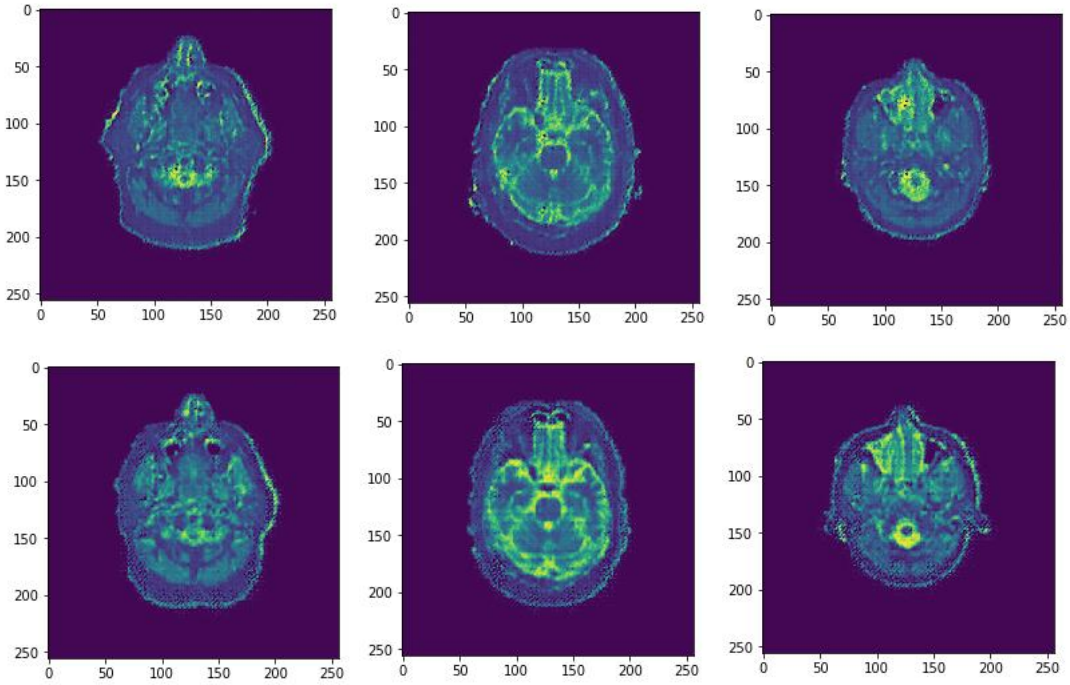


Figure 27. Fake B (top row) and corresponding fake A (bottom row) generated from test set

Lastly, an analysis of the histograms can be seen in Figure 28.

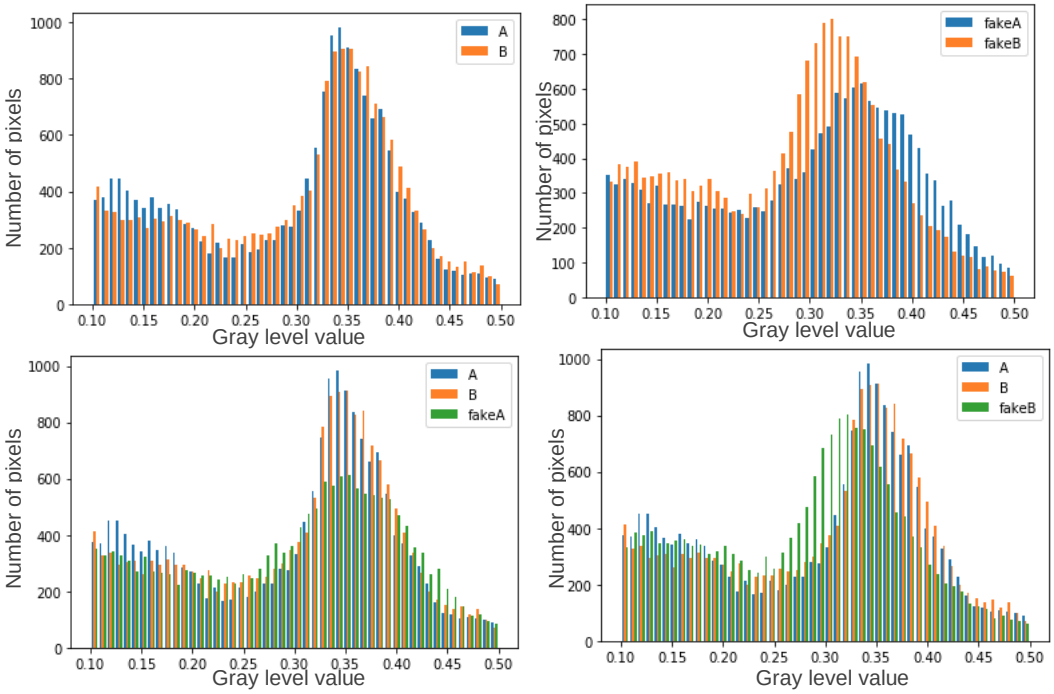


Figure 28. Average histograms of A and B images (top left), fake A vs fake B (top right), and for further classification of A, B and fake A (bottom left) and a comparative of A, B and fake B (bottom right)

Experiment 5

Representative results from the training set can be visualized in Figure 29 and Figure 30.

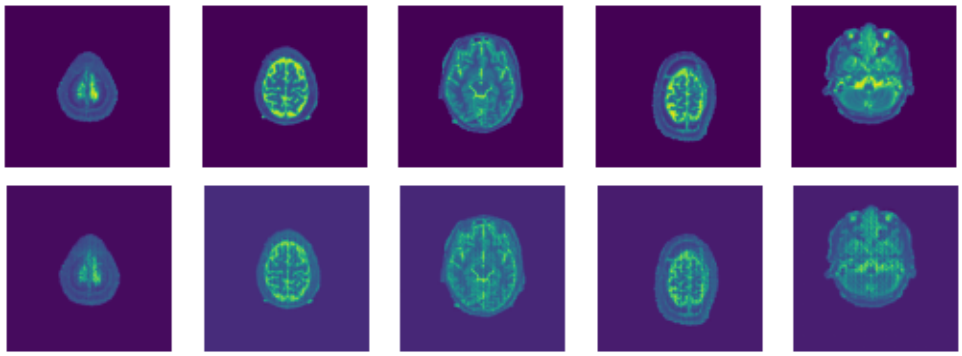


Figure 29. A to B training set results for Experiment 1. Top row is a random selection of real A images, bottom row is fake B generated from the images in the top row

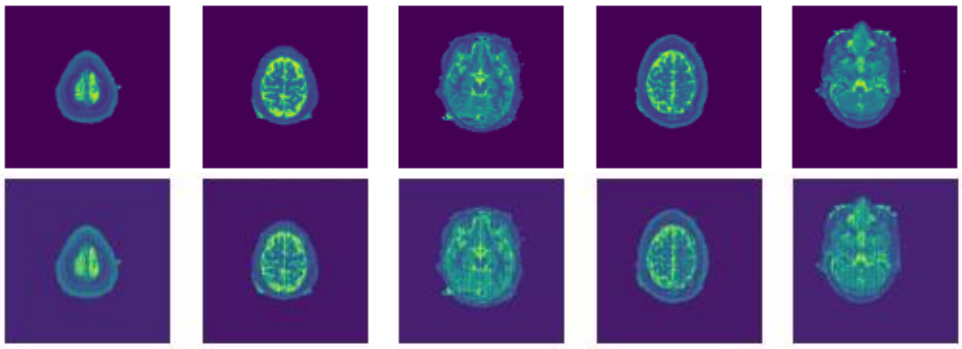


Figure 30. B to A training set results for Experiment 1. Top row is a random selection of real B images, bottom row is fake A generated from the images in the top row

In Figure 31 a random choice of fake B images is shown in the top row; the corresponding fake A images can be seen in the lower row.

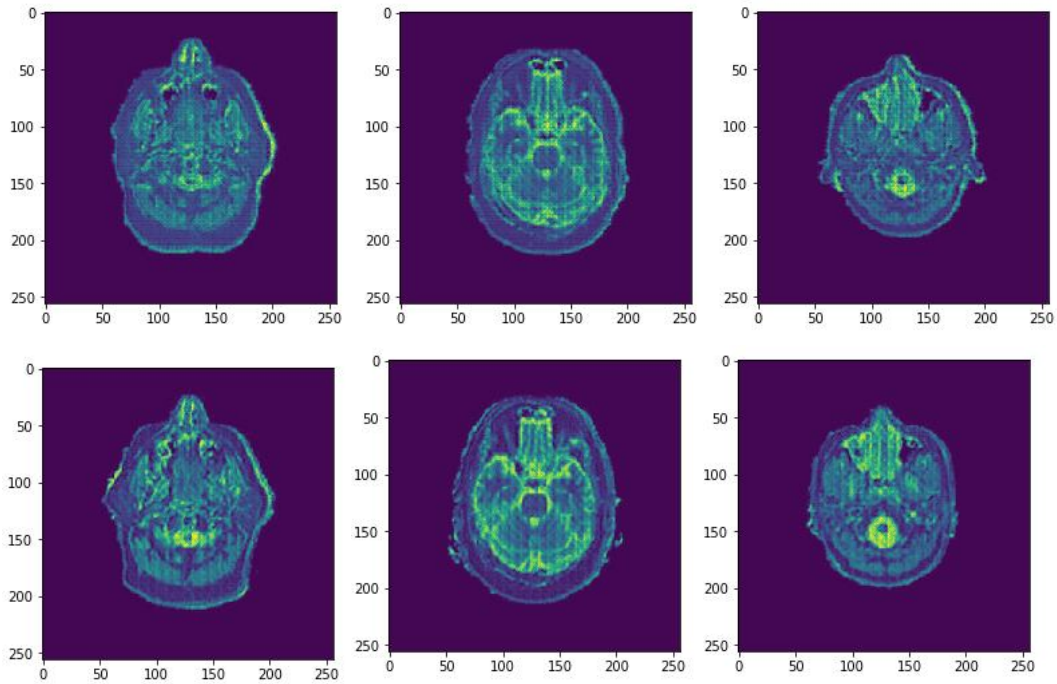


Figure 31. Top row shows some random examples of fake B images and bottom row shows pairing fake A images.

Lastly, an analysis of the histograms can be seen in Figure 32.

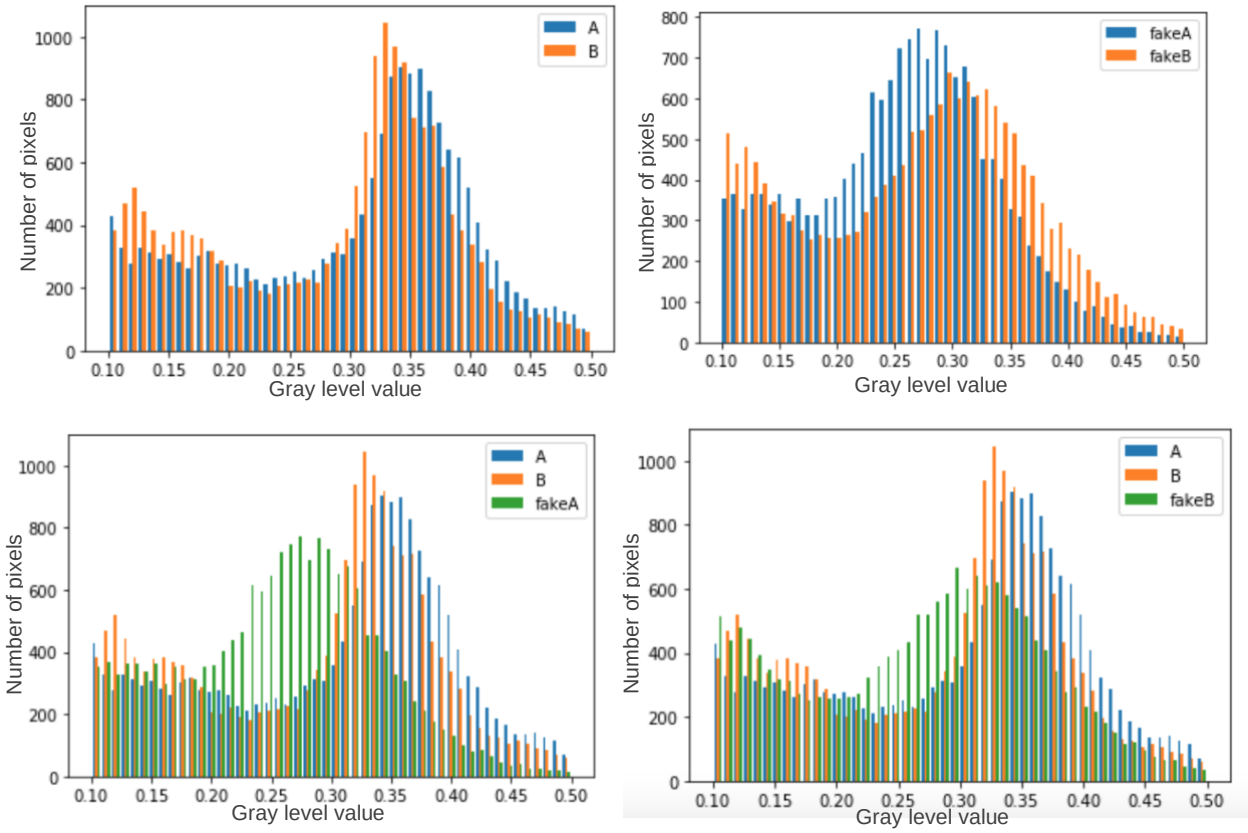


Figure 32. Average histograms of A and B images (top left), fake A vs fake B (top right), and for further classification of A, B and fake A (bottom left) and a comparative of A, B and fake B (bottom right)

Experiment 6

Representative results from the training set can be seen in Figure 33 and Figure 34.

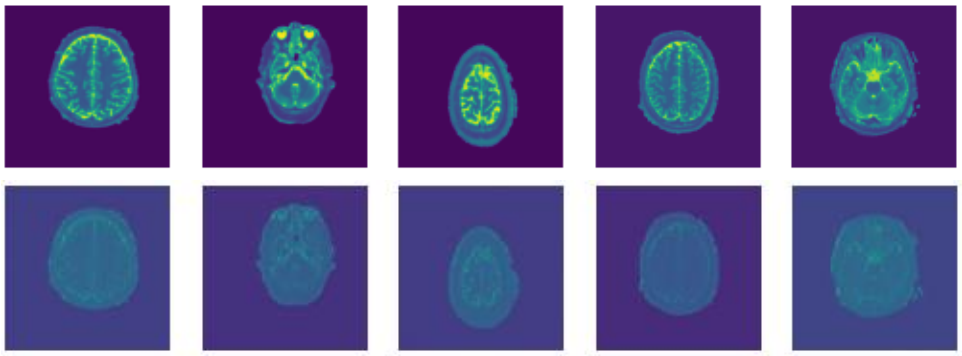


Figure 33. A to B training set results for Experiment 1. Top row is a random selection of real A images, bottom row is fake B generated from the images in the top row

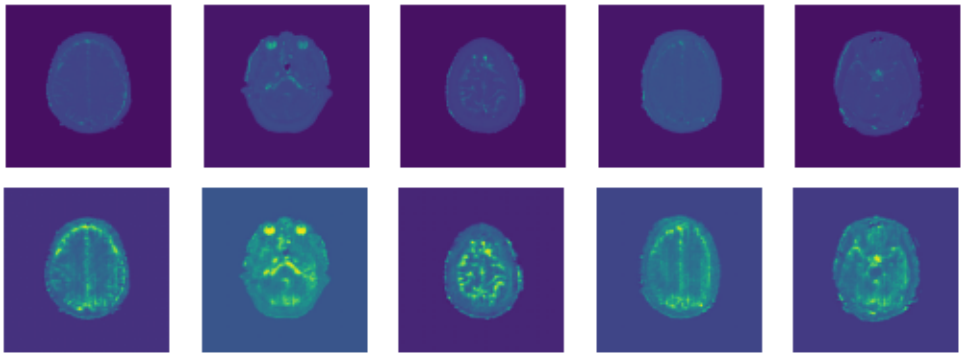


Figure 34. B to A training set results for Experiment 1. Top row is a random selection of real B images, bottom row is fake A generated from the images in the top row

In Figure 35 a random choice of fake B images is shown in the top row; the corresponding fake A images can be seen in the lower row.

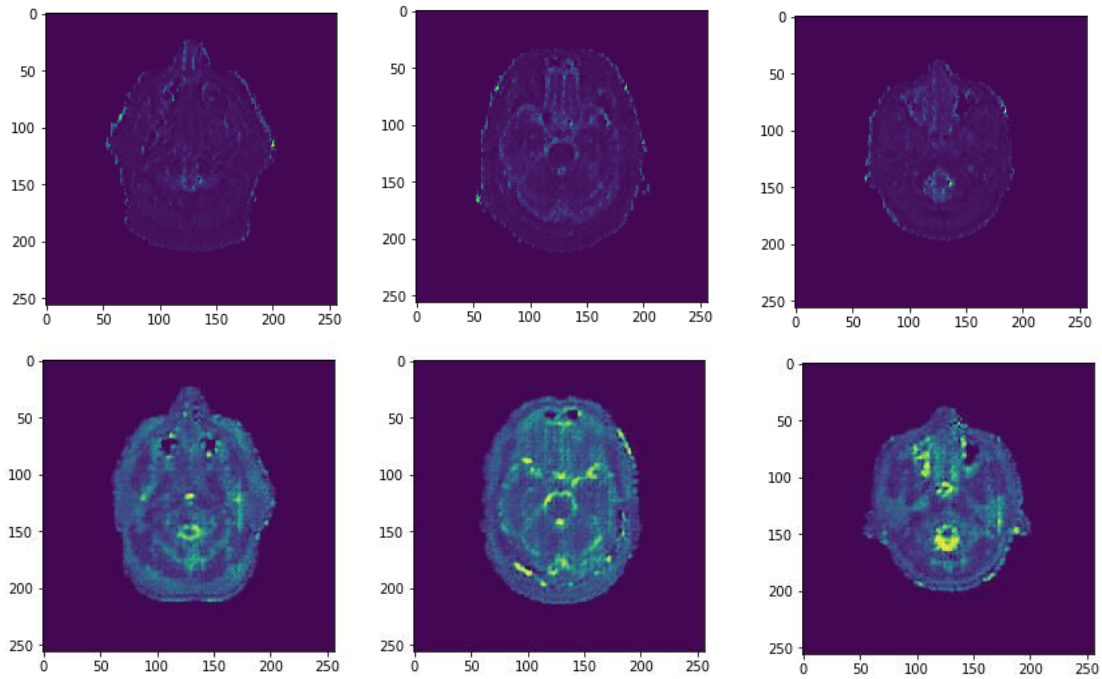


Figure 35. Fake B (top row) and corresponding fake A (bottom row) generated from test set

Finally, an analysis of the histograms can be seen in Figure 36.

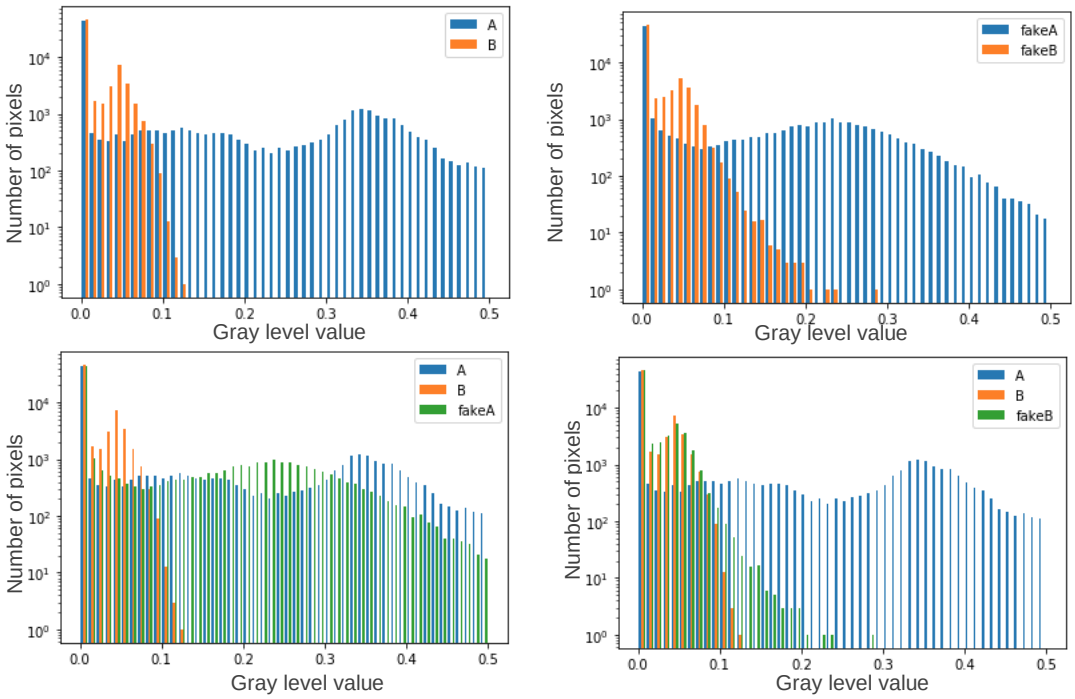


Figure 36. Average histograms of A and B images (top left), fake A vs fake B (top right), and for further classification of A, B and fake A (bottom left) and a comparative of A, B and fake B (bottom right). Logarithmic scale is applied for better distinction of the differences.

4 Conclusions

In this chapter, the main conclusions from the results of the project are explained. Some future recommendations are also presented for any possible further research into the topic.

4.1. Initial variability assessment

As can be observed in Figure 6, to distinguish the origin of the images is impossible through a visual inspection since they are all quantitative T1 maps. Therefore, unlike all the papers in the literature review, this is not considered an appropriate method of validation.
Statistical analysis is the other widely used technique to validate results. However, histograms shown in Figure 7 prove that differences in gray values are also minimal. The general distribution seems to be more or less aligned and differences don't go above around 50 pixels for a specific gray value. Given that the images are 256 by 256 pixels (so 65536 pixels in total), statistical analysis needs to be done carefully. Variations in means, standard deviations and any other measures might not show big enough variations to establish robust conclusions. After registration, changes on both sets are minimal. One might hypothesize that registration does not induce an important modification to the images that can affect the goal of the project.
To support these graphic assessments with numerical values, in Table 2, the SSIM and RMSE are shown. It would be reasonable to think that after registration, images are more similar and, therefore, SSIM is higher and RMSE is lower. Actually, both values are slightly higher after registration. Since both rises in numbers are very small it can be predicted that registration minimally changes images and should not have a huge effect on the process. It is still trusted that those minimal variations that registration has provided, allow the network to focus on any possible changes between scanners different from anatomical disparities.
Very little statistical variability is detected between the scanners with the two first assessments. This matches with the studies that mentioned a small percentage of variation. Simultaneously, this makes it difficult to assess the harmonization performance the same way it has been done typically in the literature. The last assessment, designed as an alternative based on a CNN classifier, seems to be the one finding it easier to make a clear distinction between the scanners.
With a very simple CNN and a small number of epochs and samples, very good discrimination performances are found (see Table 3 and Figure 8). To eliminate confounding factors, the background is set to 0 and images are cropped to the central region, where the brain is located. This still allows for a good classification performance. It can be seen that, after registration, this classification performance is a bit lower. This could be due to the smoothing of the image through the splines interpolation necessary to transform the images.
As a way to shed some light on what the discriminator triggers in the images of each scanner, some attention maps were drawn (see Figure 9-12). They seem to focus on random areas with a preference for edges. Some common MRI effects might make edges slightly different due to the inhomogeneities in the main static magnetic fields.
According to this initial assessment, the task of harmonization seems hard to achieve or, at least, to validate properly. Nevertheless, if the discriminator CNN is able to find differences so easily, this could be an indicator that the CycleGAN should be able to further harmonize the images.

4.2. CycleGAN experiments

The experiments were sorted in a chronological manner. Therefore, any changes performed were aimed at solving a prior difficulty found or aimed towards proving a hypothesis that a previous experiment raised.

Experiment 1

Initially, a normal CycleGAN as described in (Zhu et al., 2020) was tested. Loss numbers don't show a convergence since the discriminators are still too good at distinguishing the real from the fake images. No other epoch had much better results on that aspect.
Visual results for both training and testing images seem good, but the histograms do not reflect this. A shift is very noticeable in the case of fake A. This is a tendency that can be observed in many of the subsequent experiments. Notice that through epochs with similar characteristics, histograms show similar oscillating shifts. It is likely that the loss function allows for several gray value configurations (that is why the identity loss was initially created, see Equation 5) and a better match of histograms does not really mean a more optimized function.

In Table 5 and Table 6 it can be seen that statistical analysis does not reflect the proper transformation of the images. What could be expected is that the correlation and the SSIM is higher for A-fake A and B-fake B than for any other combination while RMSE is the lowest for these combinations. There could be a confounding factor in the differences in anatomy, though. Particularly, the highest correlation and SSIM, and the lowest RMSE, always corresponds to A-fake B and B-fake A. As can be seen in Table 7, when input into the classifier, the fake B generated images are always classified as A.

Experiment 2

This experiment lowers the weight of the Identity loss during training, which, in turn, lowers the contribution to maintaining the output close to the input. By lowering it, there is an expectation of minimizing the intuitively expected effect of Experiment 1 that the network doesn't transform the inputs enough.
However, no significant difference in any validation case was noticed. Training numbers are similar, images look good, histograms look a bit shifted, statistics show little change and CNN classification is always wrong.

Experiment 3

This experiment is designed to guide the training in the way of favouring the CNN classifier validation. By adding a discriminator that should believe A and fake A are one class and B and fake B are another, the final objective of classifier should be further included into its optimization.
The initial batch size of 20 tested in some experiments is changed to 1 as for low learning rates it is recommended to use lower batch sizes (Kandel & Castelli, 2020). Sharper images seem to result from this change. However, training appeared a bit more unstable which will lead to the use an intermediate value of 5 in the next experiments (see Annex B). Eventually, it turned out that total loss values went up and down during training without a clear convergence.

Experiment 4

As the literature shows (Xiang et al., 2018) (Dewey et al., 2019) the addition of a SSIM can help guide the CycleGAN into focusing on smaller parts of the image to do the transformation. The SSIM loss is initially added because it has been observed that CycleGANs can change anatomy during transformation (Cohen et al., 2018), which is a highly undesired effect. In our case, it could also contribute to noticing smaller details that the normal CycleGAN is not able to distinguish with such similar images as two quantitative T1 maps. However, results are again the same as in all previous experiments. A small modification was found in the posterior discrimination validation, as the accuracy is not exactly 0. This leads us to think that the SSIM can be key to the transformation.

Experiment 5

By giving more weight to the SSIM loss, the aim was to enhance the more promising results of Experiment 4 and the validation through this statistical measure.
To maximize the focus, training was now performed with matching images from the two distributions, meaning the same slice from the same patient are input at the same time. Together with registration, this aim is to focus the attention of the network on mapping scanner characteristics and not possible positional changes.
Results are, in general, still not good in statistical terms. There should be a special mention to the classification accuracy rising to quite a high number. This is curious given that images are visually less good looking visually. Nevertheless, this might be easily explained. The increase of SSIM, which helps transform in regions, generates those square-like patterns (Figure 30) that could be confusing the classifier.
After the failure of all the previous experiments, a conclusion is reached that no harmonization is possible on the given dataset with the loss functions are typically designed typically for a CycleGAN. Experiment 6 is designed to support this.

Experiment 6

This experiment was carried out in parallel after some failures in order to check the final hypothesis that it is not possible to harmonize the current dataset. It was designed as a quick test by just loading different images while using the same approach as in Experiment 5. Therefore, there is no optimization for perfect results. It can be seen by the loss values of the discriminator during training (being very low) and the training visual results, that the epoch is not optimal. Nevertheless, histograms match, cross correlation and SSIM are higher for A-fake A and B-fake B and the opposite for NRMSE.

Our results warrant some limitations of the outcomes presented in the literature. The images that the authors of previous papers have tried to harmonize before are all clearly very different. This allows for a visual assessment of the performance of the model. In (Modanwal et al., 2020), a standard deviation of gray values is used as validation metric to compare real and fake images. Given the histograms resulting from the experiments of this project, it would result in the current model being definitely successful. In (Xiang et al., 2018), a high value of SSIM appears to be presented between output (fake A or fake B) and reference image (real A or a real B respectively) without comparison to the SSIM of other image combinations. Again, by this validation method, the results of our project would also seem good.
In all these publications, images that undergo large transformations toward the target image seem to make thorough numerical validation seem less important. In the last experiment, where images are very different looking, being T1 and T2 images, visual and statistical results match all the standards set to prove a successful harmonization. This is an indication that the CycleGAN approach does work properly. Therefore, the problem lays within the dataset, which might contain too similar images, or with the combination of the CycleGAN and this dataset. Also, given the strong similarity between the images from the two scanners, the loss function might be too general, and its optimization does not suffice.

4.2. Recommendations and Future work

The time and resources limitations of the project did not allow for more work on the desired harmonization. A series of suggestions are made for any possible future research on the topic.

- For an easier statistical study of results, images with perfect anatomical match should be acquired. Measures like the ones used in this project, SSIM and RMSE, depend on a pixel to pixel correspondence of the structures in the images. With a better match between them, errors can be reduced. The possibility of a nonrigid registration is also there. However, the interpolation needed for such task might affect the images in an unknown way, probably making the harmonization less good on untouched images.
- A more in-depth study of the classifier that distinguishes between the scanners and how it does so could be the key to designing a loss function for the CycleGAN that finds the fine differences.

Although results are not what could have been expected by the straightforward application of the CycleGAN in the literature and their seemingly good results, it is important to bear in mind the following. Although artificial intelligence can be a very useful tool and has been proven valuable in many situations, medical images are a very delicate source of information that should not be altered lightly. As such, more rigorous validation techniques should be applied to check whether the generated images using GANs as described in the literature are truly clinically useful.

5 Bibliography

Activation Functions—ML Glossary documentation. (n.d.). Retrieved October 8, 2020, from https://ml-cheatsheet.readthedocs.io/en/latest/activation_functions.html

AlBadawy, E. A., Saha, A., & Mazurowski, M. A. (2018). Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical Physics*, *45*(3), 1150–1158. https://doi.org/10.1002/mp.12752

Andrew J. Taylor, Salerno, M., Dharmakumar, R., & Jerosch-Herold, M. (2016). *T1 Mapping | Elsevier Enhanced Reader*. https://doi.org/10.1016/j.jcmg.2015.11.005

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., & Yang, B. (2020). MedGAN: Medical image translation using GANs. *Computerized Medical Imaging and Graphics*, *79*, 101684. https://doi.org/10.1016/j.compmedimag.2019.101684

Bellman, R. E. (2015). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Ben-Cohen, A., Klang, E., Raskin, S. P., Soffer, S., Ben-Haim, S., Konen, E., Amitai, M. M., & Greenspan, H. (2018). Cross-Modality Synthesis from CT to PET using FCN and GAN Networks for Improved Automated Lesion Detection. *ArXiv:1802.07846 [Cs]*. http://arxiv.org/abs/1802.07846

Bergeest, J.-P., & Jäger, F. (2008). A Comparison of Five Methods for Signal Intensity Standardization in MRI. In T. Tolxdorff, J. Braun, T. M. Deserno, A. Horsch, H. Handels, & H.-P. Meinzer (Eds.), *Bildverarbeitung für die Medizin 2008* (pp. 36–40). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78640-5_8

Brownlee, J. (2019a, August 1). How to Develop a Pix2Pix GAN for Image-to-Image Translation. *Machine Learning Mastery*. https://machinelearningmastery.com/how-to-develop-a-pix2pix-gan-for-image-to-image-translation/

Brownlee, J. (2019b, August 8). How to Develop a CycleGAN for Image-to-Image Translation with Keras. *Machine Learning Mastery*. https://machinelearningmastery.com/cyclegan-tutorial-with-keras/

Brunner, P., & Ernst, R. R. (1979). Sensitivity and performance time in NMR imaging. *Journal of Magnetic Resonance (1969)*, *33*(1), 83–106. https://doi.org/10.1016/0022-2364(79)90192-6

Chantal M.W.Tax. (n.d.). *Cross-scanner and cross-protocol multi-shell diffusion MRI data harmonization: Algorithms and results | Elsevier Enhanced Reader*. https://doi.org/10.1016/j.neuroimage.2020.117128

Chavhan, G. B., Babyn, P. S., Thomas, B., Shroff, M. M., & Haacke, E. M. (2009). Principles, Techniques, and Applications of T2*-based MR Imaging and Its Special Applications. *RadioGraphics*, *29*(5), 1433–1449. https://doi.org/10.1148/rg.295095034

Chen, J., Chen, J., Chao, H., & Yang, M. (2018). Image Blind Denoising with Generative Adversarial Network Based Noise Modeling. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3155–3164. https://doi.org/10.1109/CVPR.2018.00333

Cohen, J. P., Luck, M., & Honari, S. (2018). Distribution Matching Losses Can Hallucinate Features in Medical Image Translation. *ArXiv:1805.08841 [Cs]*. http://arxiv.org/abs/1805.08841

Dar, S. U. H., Yurt, M., Karacan, L., Erdem, A., Erdem, E., & Çukur, T. (2018). Image Synthesis in Multi-Contrast MRI with Conditional Generative Adversarial Networks. *ArXiv:1802.01221 [Cs]*. http://arxiv.org/abs/1802.01221

Deoni, S. C. L., Williams, S. C. R., Jezzard, P., Suckling, J., Murphy, D. G. M., & Jones, D. K. (2008). Standardized structural magnetic resonance imaging in multicentre studies using quantitative T1 and T2 imaging at 1.5 T. *NeuroImage*, *40*(2), 662–671. https://doi.org/10.1016/j.neuroimage.2007.11.052

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., Saidha, S., Oh, J., Pham, D. L., Calabresi, P. A., van Zijl, P. C. M., & Prince, J. L. (2019). DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic Resonance Imaging*, *64*, 160–170. https://doi.org/10.1016/j.mri.2019.05.041

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78. https://doi.org/10.1145/2347736.2347755

Dong, S., Luo, G., Wang, K., Cao, S., Mercado, A., Shmuilovich, O., Zhang, H., & Li, S. (2018). VoxelAtlasGAN: 3D Left Ventricle Segmentation on Echocardiography with Atlas Guided Generation and Voxel-to-voxel Discrimination. *ArXiv:1806.03619 [Cs]*. http://arxiv.org/abs/1806.03619

Fetty, L., Löfstedt, T., Heilemann, G., Furtado, H., Nesvacil, N., Nyholm, T., Georg, D., & Kuess, P. (2020). Investigating conditional GAN performance with different generator architectures, an ensemble model, and different MR scanners for MR-sCT conversion. *Physics in Medicine & Biology*, *65*(10), 105004. https://doi.org/10.1088/1361-6560/ab857b

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M., & Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120.

https://doi.org/10.1016/j.neuroimage.2017.11.024
Fullerton, G. D. (1987). Magnetic resonance imaging signal concepts. *RadioGraphics*, *7*(3), 579–596. https://doi.org/10.1148/radiographics.7.3.3448647
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (n.d.). *Generative Adversarial Nets*. 9.
Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G., Cai, J., & Chen, T. (2017). Recent Advances in Convolutional Neural Networks. *ArXiv:1512.07108 [Cs]*. http://arxiv.org/abs/1512.07108
Hagiwara, A., Hori, M., Cohen-Adad, J., Nakazawa, M., Suzuki, Y., Kasahara, A., Horita, M., Haruyama, T., Andica, C., Maekawa, T., Kamagata, K., Kumamaru, K. K., Abe, O., & Aoki, S. (2019). Linearity, Bias, Intrascanner Repeatability, and Interscanner Reproducibility of Quantitative Multidynamic Multiecho Sequence for Rapid Simultaneous Relaxometry at 3 T: A Validation Study With a Standardized Phantom and Healthy Controls. *Investigative Radiology*, *54*(1), 39–47. https://doi.org/10.1097/RLI.0000000000000510
Hagiwara, A., Warntjes, M., Hori, M., Andica, C., Nakazawa, M., Kumamaru, K. K., Abe, O., & Aoki, S. (2017). SyMRI of the Brain: Rapid Quantification of Relaxation Rates and Proton Density, With Synthetic MRI, Automatic Brain Segmentation, and Myelin Measurement. *Investigative Radiology*, *52*(10), 647–657. https://doi.org/10.1097/RLI.0000000000000365
Harvard University. (2017). *Diffusion MRI Data Harmonisation (2017)*. https://projects.iq.harvard.edu/cdmri2017/challenge
Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv:1502.03167 [Cs]*. http://arxiv.org/abs/1502.03167
Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2018). Image-to-Image Translation with Conditional Adversarial Networks. *ArXiv:1611.07004 [Cs]*. http://arxiv.org/abs/1611.07004
Jackson, E. F., Ginsberg, L. E., Schomer, D. F., & Leeds, N. E. (1997). A review of MRI pulse sequences and techniques in neuroimaging. *Surgical Neurology*, *47*(2), 185–199. https://doi.org/10.1016/s0090-3019(96)00375-8
Jagtap, R. (2020, June 25). *A Comprehensive Guide to Generative Adversarial Networks (GANs)*. Medium. https://towardsdatascience.com/a-comprehensive-guide-to-generative-adversarial-networks-gans-fcfe65d1cfe4
Jiang, D., Liu, P., Li, Y., Mao, D., Xu, C., & Lu, H. (2018). Cross-vendor harmonization of T_2-relaxation-under-spin-tagging (TRUST) MRI for the assessment of cerebral venous oxygenation: Cross-Vendor Harmonization of TRUST MRI. *Magnetic Resonance in Medicine*, *80*(3), 1125–1131. https://doi.org/10.1002/mrm.27080
Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., & Song, M. (2018). Neural Style Transfer: A Review. *ArXiv:1705.04058 [Cs, Eess, Stat]*. http://arxiv.org/abs/1705.04058
Joseph P. Hornak. (1996). *The Basics of MRI*. https://www.cis.rit.edu/htbooks/mri/inside.htm
Kandel, I., & Castelli, M. (2020). The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*, *6*(4), 312–315. https://doi.org/10.1016/j.icte.2020.04.010
Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., Bae, H., & Kim, N. (2019). Deep Learning in Medical Imaging. *Neurospine*, *16*(4), 657–668. https://doi.org/10.14245/ns.1938396.198
Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017). Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. *ArXiv:1703.05192 [Cs]*. http://arxiv.org/abs/1703.05192
Kingma, D. P., & Welling, M. (2019). An Introduction to Variational Autoencoders. *ArXiv:1906.02691 [Cs, Stat]*. https://doi.org/10.1561/2200000056
Kransdorf, M. J., & Murphey, M. D. (2000). Radiologic Evaluation of Soft-Tissue Masses. *American Journal of Roentgenology*, *175*(3), 575–587. https://doi.org/10.2214/ajr.175.3.1750575
Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A. D., Bogovic, J. A., Hua, J., Chen, M., Jarso, S., Smith, S. A., Joel, S., Mori, S., Pekar, J. J., Barker, P. B., Prince, J. L., & van Zijl, P. C. M. (2011). Multi-parametric neuroimaging reproducibility: A 3-T resource study. *NeuroImage*, *54*(4), 2854–2866. https://doi.org/10.1016/j.neuroimage.2010.11.047
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539
Liao, H., Huo, Z., Sehnert, W. J., Zhou, S. K., & Luo, J. (2018). Adversarial Sparse-View CBCT Artifact Reduction. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, & G. Fichtinger (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (pp. 154–162). Springer International Publishing. https://doi.org/10.1007/978-3-030-00928-1_18
Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets. *ArXiv:1411.1784 [Cs, Stat]*. http://arxiv.org/abs/1411.1784
Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C. E., Morey, R.

A., Flashman, L. A., George, M. S., McAllister, T. W., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R. D., Coleman, M. J., Kubicki, M., Westin, C. F., … Rathi, Y. (2016). Inter-site and inter-scanner diffusion MRI data harmonization. *NeuroImage*, *135*, 311–323. https://doi.org/10.1016/j.neuroimage.2016.04.041
Mirzaalian, Hengameh, Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Karmacharya, S., Grant, G., Marx, C. E., Morey, R. A., Flashman, L. A., George, M. S., McAllister, T. W., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R. D., Coleman, M. J., Kubicki, M., … Rathi, Y. (2018). Multi-site harmonization of diffusion MRI data in a registration framework. *Brain Imaging and Behavior*, *12*(1), 284–295. https://doi.org/10.1007/s11682-016-9670-y
Modanwal, G., Vellal, A., Buda, M., & Mazurowski, M. (2020). *MRI image harmonization using cycle-consistent generative adversarial network*. https://doi.org/10.1117/12.2551301
Moyer, D., Ver Steeg, G., Tax, C. M. W., & Thompson, P. M. (2020). Scanner invariant representations for diffusion MRI harmonization. *Magnetic Resonance in Medicine*, *84*(4), 2174–2189. https://doi.org/10.1002/mrm.28243
Murphy, M. (2011). *Parallelism, Patterns, and Performance in Iterative MRI Reconstruction*.
Pierpaoli, C. (2010). Quantitative Brain MRI. *Topics in Magnetic Resonance Imaging : TMRI*, *21*(2), 63. https://doi.org/10.1097/RMR.0b013e31821e56f8
Pinto, M. S., Paolella, R., Billiet, T., Van Dyck, P., Guns, P.-J., Jeurissen, B., Ribbens, A., den Dekker, A. J., & Sijbers, J. (2020). Harmonization of Brain Diffusion MRI: Concepts and Methods. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00396
Qayyum, A., Anwar, S., Majid, M., Awais, M., & Alnowami, M. (2018). Medical Image Analysis using Convolutional Neural Networks: A Review. *Journal of Medical Systems*. https://doi.org/10.1007/s10916-018-1088-1
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv:1505.04597 [Cs]*. http://arxiv.org/abs/1505.04597
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7
Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556 [Cs]*. http://arxiv.org/abs/1409.1556
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.
SyntheticMR. (n.d.). Retrieved September 23, 2020, from https://syntheticmr.com/products/symri-image/
Tax, C. MW., Grussu, F., Kaden, E., Ning, L., Rudrapatna, U., John Evans, C., St-Jean, S., Leemans, A., Koppers, S., Merhof, D., Ghosh, A., Tanno, R., Alexander, D. C., Zappalà, S., Charron, C., Kusmia, S., Linden, D. EJ., Jones, D. K., & Veraart, J. (2019). Cross-scanner and cross-protocol diffusion MRI data harmonisation: A benchmark database and evaluation of algorithms. *NeuroImage*, *195*, 285–299. https://doi.org/10.1016/j.neuroimage.2019.01.077
Vemulapalli, R., Nguyen, H. V., & Zhou, S. K. (2015). Unsupervised Cross-Modal Synthesis of Subject-Specific Scans. *2015 IEEE International Conference on Computer Vision (ICCV)*, 630–638. https://doi.org/10.1109/ICCV.2015.79
Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., Duncan, J. S., Richardson, M. P., & Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. *NeuroImage*, *51*(4), 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046
Wang, J., Zhao, Y., Noble, J. H., & Dawant, B. M. (2018). Conditional Generative Adversarial Networks for Metal Artifact Reduction in CT Images of the Ear. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, & G. Fichtinger (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (pp. 3–11). Springer International Publishing. https://doi.org/10.1007/978-3-030-00928-1_1
Warntjes, J. B. M., Leinhard, O. D., West, J., & Lundberg, P. (2008). Rapid magnetic resonance quantification on the brain: Optimization for clinical usage. *Magnetic Resonance in Medicine*, *60*(2), 320–329. https://doi.org/10.1002/mrm.21635
Xiang, L., li, Y., Lin, W., & Wang, Q. (2018). *Unpaired Deep Cross-Modality Synthesis with Fast Training: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings* (Vol. 11045, pp. 155–164). https://doi.org/10.1007/978-3-030-00889-5_18
Yang, Q., Li, N., Zhao, Z., Fan, X., Chang, E. I.-C., & Xu, Y. (2018). MRI Cross-Modality NeuroImage-to-NeuroImage Translation. *ArXiv:1801.06940 [Cs]*. http://arxiv.org/abs/1801.06940
Yann Lecun, Patrick Haffner, & Y. Bengio. (2000). *Object Recognition with Gradient-Based Learning*. https://www.researchgate.net/publication/2816141_Object_Recognition_with_Gradient-Based_Learning
Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, *58*, 101552. https://doi.org/10.1016/j.media.2019.101552

Yves Chauvin & David E. Rumelhart. (1995). *Backpropagation: Theory, Architectures, and Applications*. Psychology Press.
Zhao, H., Li, H., & Cheng, L. (2017). Synthesizing Filamentary Structured Images with GANs. *ArXiv:1706.02185 [Cs]*. http://arxiv.org/abs/1706.02185
Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2020). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *ArXiv:1703.10593 [Cs]*. http://arxiv.org/abs/1703.10593

Annex A – Complete Literature Review

1. INTRODUCTION

Nuclear magnetic resonance (NMR) is a frequently used medical imaging technique, that allows for a very high contrast between tissues. A crucial advantage over x-ray imaging is that it exploits magnetic properties of the nuclei of atoms instead of using ionizing radiation (Brunner & Ernst, 1979)(Kransdorf & Murphey, 2000). The mechanisms on which the technique is based are complex and a full review of them is out of the scope of this study. However, this literature study provides a simplified summary to achieve a basic understanding of the project rationale.

A. MAGNETIC RESONANCE IMAGING (MRI)

Hydrogen atoms, present in water and fat, constitute 63% of the body content. They have spin, which can be conceived as a magnetic dipole that precesses around an axis (see Figure 1). To acquire an image, first, an external magnetic field is applied to align the spins in a specific direction. Subsequently, a radiofrequency (RF) pulse is used to strategically and locally alter this alignment. After stopping said pulse, the recovery to their resting state generates the signal that is recorded to produce the resulting images (Joseph P. Hornak, 1996).

Two of the most common types of images we can find in clinical applications are called T1-weighted and T2-weighted images respectively. T1 weighted images rely on the time it takes for the spins to realign in the direction of the B0 field (i.e. the main static magnetic field of the (MR) scanner). In other words, the amount of magnetization after some time has elapsed and longitudinal magnetization, M_Z , is recovering (Andrew J. Taylor et al., 2016). T2 weighted images, on the other hand, represent the decay of transverse magnetization, $M_{x,y}$, at some chosen moment while the magnetization is decaying (Chavhan et al., 2009). Figure 1 helps understand the two modalities.

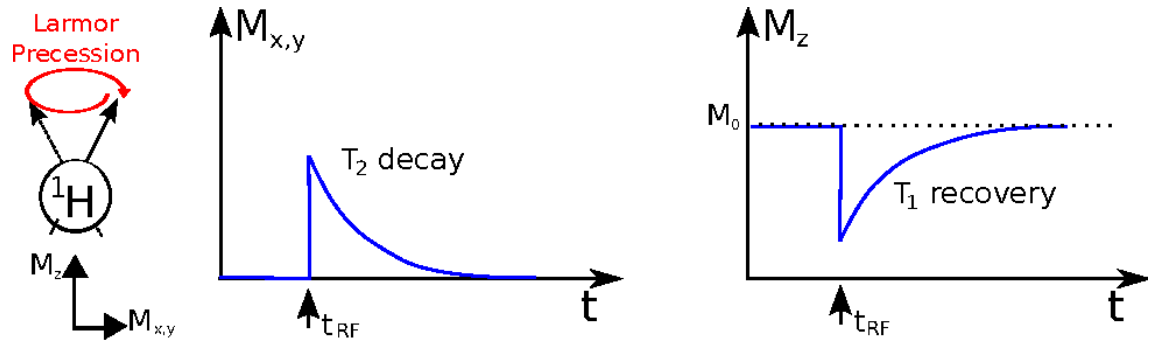


Figure 1. From left to right. A) Schematics of the precession of a hydrogen atom. Definition of the system of reference for axial magnetization, M_z (direction of B0), and transverse magnetization $M_{x,y}$, which appears after a disruptive RF pulse is applied. B) Plot of T2 decay after a RF disruption at time t_{RF}. C) Plot of T1 recovery after RF disruption at time t_{RF}. Figure from (Murphy, 2011)

The contrast of weighted MRI acquisition is affected by factors that are intrinsic to the tissue but also factors that are experiment-dependent (Pierpaoli, 2010). The versatility of MRI is because different system parameter settings will highlight different tissue properties. An important issue, however, is that the signal intensity depends on things such as the strength or homogeneity of the static magnetic fields like B0, the sequence of pulses used to acquire the signal, etc (Fullerton, 1987) (Jackson et al., 1997). Consequently, maps obtained with most of the weighted MRI sequences provide signal values with no direct physical or chemical meaning and that cannot be compared with other tissues or patients (Bergeest & Jäger, 2008)(Pierpaoli, 2010).

B. QUANTITATIVE MRI

Some techniques have been developed with the aim of transforming such qualitative techniques into quantitative approaches. Quantitative MRI is the procurement of maps with meaningful physical or chemical

characteristics that are reproducible and can be compared between subjects and experiments (Pierpaoli, 2010). There are different ways to obtain these maps.
Synthetic MR (*SyntheticMR*, n.d.) has commercialized a software already in clinical use, which can create quantitative T1 and T2 maps. It does so by fitting a model of the relaxation to several, quickly acquired, weighted images. While doing so, system imperfections such as magnetic field inhomogeneities are taken into account. This would allow for detection of damaged tissue in, for example, multiple sclerosis patients by a physically sound quantification of tissue properties (Warntjes et al., 2008).
In (Warntjes et al., 2008) it is claimed that this method should, in principle, eliminate all the scanner dependencies caused by system parameters variations, different software versions, hardware, etc. A limitation of this work, however, is that all experiments are done using the same scanner.
Similarly, in (Hagiwara et al., 2017), the authors experiment with different vendors (Siemens and Philips) and different coils (1.5T and 3T), claiming they obtain the comparable results. However no quantitative data is provided.
Later studies provide further insight in the variability from which quantitative acquisition techniques can suffer. In (Deoni et al., 2008) the authors reported 6.5% variability in T1 time and 8% variability in T2 time measured on scanner from two different vendors. (Bauer et al., 2010) tests on T2 maps from 3 different vendors, finding a 20% variability. The most optimictic study, by (Hagiwara et al., 2019), reports a maximum time difference of 3.15% for T1 and 5.6% for T2.

C. PROBLEM DEFINITION

Artificial intelligence-based image analysis techniques are known to be sensitive to these variabilities. For instance, a study by (AlBadawy et al., 2018) concludes that training models with data from different institutions produces a dramatic deterioration in model performance. Their main hypothesis is that the differences in scanners and their parameters are the main reason behind this. Eliminating these variances could enhance numerous amount of post-acquisition processes.
This literature study aims towards reviewing the state of the art on inter-scanner harmonization deep learning techniques with the objective of finding the most suited technique to bridge the gap between quantitative maps from different scanners. It will do so in 4 chapters. Chapter 1 introduces the issue of lack of harmonization in quantitative MR. Chapter 2 reviews the basics concepts of Machine Learning necessary to understand the literature. Chapter 3 summarizes the main findings in state of the art in MRI (and sometimes other imaging modalities too) inter scanner harmonization deep learning techniques. Chapter 4 provides some conclusions.

2. MACHINE LEARNING

This section revisits the basic concepts of Machine Learning to understand the advanced techniques found in the literature.

A. ARTIFICIAL INTELLIGENCE

The grand challenge of Artificial Intelligence consists on being able to solve problems that are intuitive for humans but hard to describe formally (Goodfellow et al., 2016). In particular, machine learning is the capability of an algorithm to detect a pattern by generalizing from given example data (Domingos, 2012). Within this field, Deep Learning is the specific use of neural networks to solve such matter. These are structures that do not need a manual extraction of the features; instead, they receive raw data and learn relevant features automatically. The learning process is achieved by building complex constructs by a combination of simpler ones (LeCun et al., 2015).

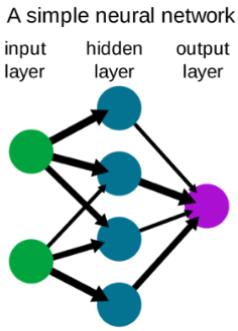


Figure 2. A simple neural network

Neural networks are a succession of layers composed by nodes as depicted in Figure 2. Each of these nodes applies parameters defining a linear function f(x)=ax+b, where x is the raw input data, a is the weights that linearly combine the inputs and b represents a potential offset. The most common models are deep

feedforward networks, where information flows from input x to output f(x)=y without any feedback connection. Intermediate computations define the total function that maps the input into the final output F(X)=Y. As networks can have many layers, each network generates a function that is connected by the chain rule to the next as follows: $F(X)=f^n(….f^2(f^1(x)))$. Where 1 is the first layer to which the input is introduced, and n is the last one. After each layer, a nonlinear activation is usually applied, enabling more complex (nonlinear) functions to be learnt.

Training of neural networks is an optimization problem to find the most suitable parameters. Backpropagation is the most common way to do this by calculating the gradient of the loss function of the network. This loss function quantifies the difference between the final prediction of the network and the true measure of the (training) sample. Backpropagation works such that it iterates backwards in the network using the chain rule to get the gradient of each layer at a time (Yves Chauvin & David E. Rumelhart, 1995). This allows for gradient descent optimization eventually minimizing the loss function across different training samples.

B. NEURAL STYLE TRANSFER

Neural Style Transfer (NST) consists on applying the style of one image to the content of another. A common use is to reproduce the artistic style of a famous painter into, for example, photographs (Jing et al., 2018) as can be seen in Figure 3. Such algorithms are originally based on Convolutional Neural Networks (CNNs)(Simonyan & Zisserman, 2015).

Figure 3. NST from Van Gogh's style applied to a picture of Persepolis

CNNs are the most common type of networks used for image processing. This is so because the number of trainable parameters is contained in convolution kernels and, as such, their number is independent on the size of the image, therefore bypassing the curse of dimensionality (Yann Lecun et al., 2000) The curse of dimensionality states that the training set has to increase exponentially with respect to the number of dimensions the model possesses (Bellman, 2015). In the case of neural networks, the number of dimensions scales with the number of inputs. In traditional multi-layer networks, dimensionality would increase if we input bigger images because every pixel is a separate input. This would, therefore, require a bigger set of images or, paradoxically, bigger images, to do an accurate prediction.

The main building blocks of CNNs, which can be seen in Figure 4, are:

- **Convolution layer.** Convolution is an operation that combines two functions (image and kernel) and essentially outputs how the shape of one is modified by the other. It can be used to detect certain features, i.e. the first derivative, that are repeated over the whole image. In each of these layers there is an a priori fixed number of kernels typically of sizes 2x2,3x3 or 4x4 pixels that act as the "revealing" function by being slid across the entire image. Thus, the weights of such kernels are also fixed in number and trained to find the most significant feature maps for whichever purpose the network has (classification, segmentation, etc.). Initially, simpler ones such as edges can be detected. More complex features can be found by applying more layers.

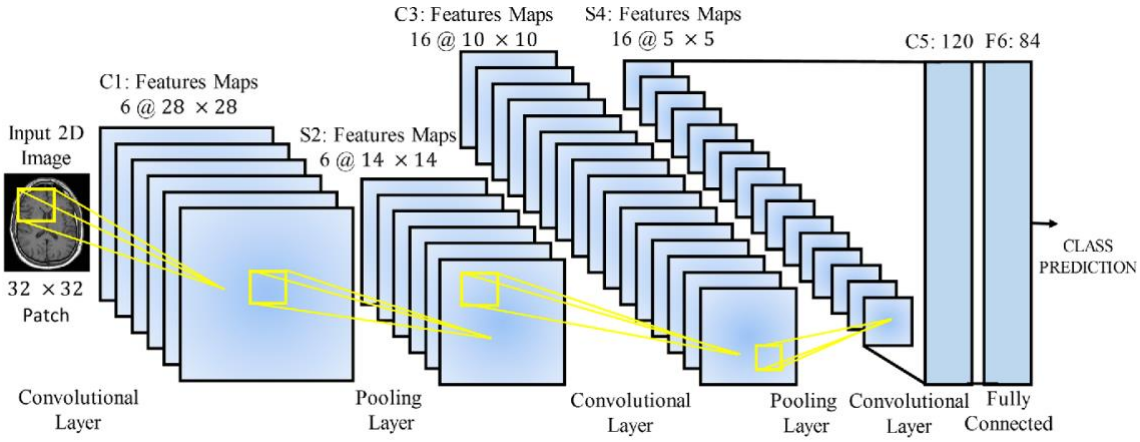


Figure 4. Example of CNN architecture. Image from (Qayyum et al., 2018)

- **Activation.** An element-wise, nonlinear activation is typically applied to the convolution results. This allows to convert linear functions into nonlinear ones. Thanks to this, more complex features can be learnt. The most commonly used activation functions are sigmoid, tanh or ReLU. Definitions and the use of each of them can be found in (*Activation Functions — ML Glossary Documentation*, n.d.).
- **Pooling.** A subsampling layer is often applied (unless padding is used, a technique aimed at keeping the spatial resolution constant during convolution) to reduce the space, save computational time, etc. The main idea behind using such layer is that the exact location of features is not as important as its relative position with respect to other features. As such the representation is kept invariant after this layer (Goodfellow et al., 2016). Average or max pooling are the two most common pooling methods (Gu et al., 2017).

To stabilize training and to regularize the network there are some common techniques that are often applied. Regularization targets overfitting prevention. This overfitting consists on losing the generalization properties of a model by overtraining and getting the model to learn too specific characteristics of the training set. As a consequence, this produces a very low training error but a very high test error.

- **Dropout.** It is a technique to avoid overfitting by which a large number of different network architectures are simulated by randomly switching off some nodes during training. This introduces noise in the process, forcing nodes to have evenly distributed importance in feature extraction (Srivastava et al., 2014).
- **Batch normalization.** Consists on standardizing the input for each layer with respect to a batch (group of training samples used at the same time). It corrects for the internal covariate shift (change in the distribution of inputs to each layer due to the change of parameters during training). It sometimes eliminates the need for adding dropout (Ioffe & Szegedy, 2015).

Many types of networks are built to do style transfer by using CNNs as building blocks. A review of deep learning techniques on medical imaging (M. Kim et al., 2019) divides image-to-image translation techniques into those using Generative Adversarial Networks (GANs), the most common, and those using other techniques. Focus will be set in GANs due to their widespread presence in the literature. According to the review by (Yi et al., 2019), the number of GAN related articles has risen to double from 2017 and 2018 and a 43% of those are on MR as compared to other imaging modalities.

C. GENERATIVE ADVERSARIAL NETWORKS (GANs)

GANs are based on the simultaneous training of two models. There is a Generator, G, which tries to approach the data distribution of the training set from a random noise input, and a Discriminator, D, generates a value expressing the probability that two inputs are drawn from the same distribution (e.g. images from the same scanner). Practically, D is trained such that it has optimal performance in distinguishing between real and fake or generated images (Goodfellow et al., n.d.). Both networks are trained simultaneously, in competition to perform better than the other and using each other's information. Convergence is achieved when the discriminator cannot tell the difference between real and generated samples. To put a practical example, if a

GAN is trained with images of faces, it will learn the characteristics of a face and generate a randomized combination of those characteristics to build a fake face. A schematic can be seen in Figure 5.
Conditional GANs (cGANs) are a version of GANs in which we input some condition to both the Generator and the Discriminator in order to guide it into what distribution we want it to generate (Mirza & Osindero, 2014). This could be, for example, to add the label of smiling person to the training of people's faces to always generate a smiley face image even though training set contains all sorts of facial expressions.

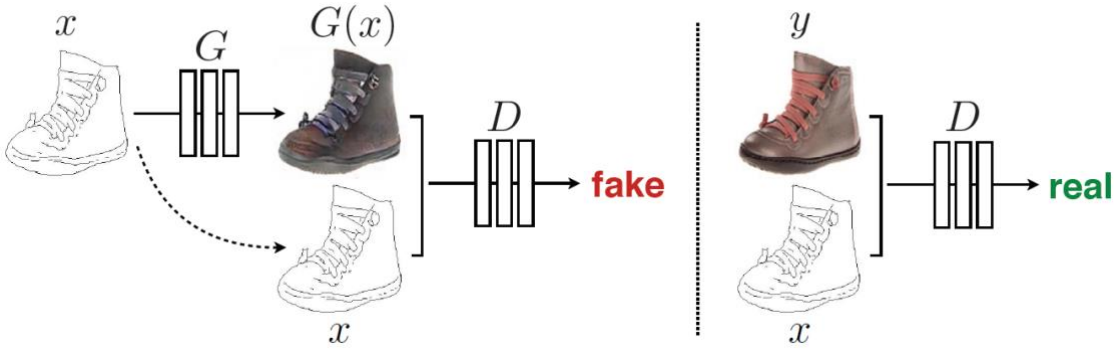


Figure 5. Schematic of Pix2Pix algorithm. Picture taken from (Isola et al., 2018)

Based on this concept, several algorithms were proposed such as Pix2Pix (Isola et al., 2018), Cycle-GAN (Zhu et al., 2020), StarGAN (Chen et al., 2018) or MedGAN (Armanious et al., 2020), Disco-GANs (T. Kim et al., 2017), Fila-sGAN (Zhao et al., 2017), among others. They can be used for many purposes in medical imaging like artifact elimination (Wang et al., 2018)(Liao et al., 2018), segmentation (Dong et al., 2018) or, the most relevant application for this project, image translation between scanners (Ben-Cohen et al., 2018)(Yang et al., 2018)(Dar et al., 2018). A brief summary is done for two of the most common GANs (Jagtap, 2020): Pix2Pix and CycleGAN, for paired and unpaired image transfer respectively.

- **Pix2Pix** is a network used for paired image translation. This means that there has to be a pixel to pixel correspondence between the pairs of images necessary to train the network because the training is designed to use spatially dependent calculations. The objective of the network is to optimize Equation 1. This mixes that of normal cGANs (conditional GANs because we use a reference image as input for the generator), equation 2, plus an L1 loss, equation 3.

$$G^* = \arg min_G max_D \, \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \qquad \text{Eq. 8}$$

Where

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log (1 - D\big(x, G(x, z)\big)] \qquad \text{Eq. 2}$$

And

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\| \, y - G(x, z) \, \|_1] \qquad \text{Eq. 3}$$

Here x and y are the paired input and reference image respectively and z is the noise that is input in the generator together with x for randomicity. G and D the functions defining the Generator and Discriminator respectively. What the cGAN describes formally in the first part of equation 2 is the probability that the discriminator is able to distinguish between real and fake images. For that, D is trained to distinguish between the real input, x, and the real reference, y. In the second part of equation 2, it is defined the probability of fooling the Discriminator into classifying the generated image G(x,z) as a reference image, y. What the L1-norm provides, through equation 3, is a pixel to pixel comparison of generated and reference image. Here is where pixel to pixel pairing of x and y is most necessary (paired data). This helps keep the output similar to the input for situations in which color, anatomy or other characteristics want to be kept. It also encourages less blurring. Both terms are combined in a weighted sum in equation 1 while λ gives more or less importance to equation 3 as desired. In this equation, the main objective of the training is set. As such, two measures need to be minimized. First, the discriminator is maximized to be good at distinguishing the real x and y. Then, the generator is minimized so that the maximum probability of the discriminator confusing G(x,z) with y is achieved.

- **CycleGAN.** This technique is able to do the image translation with unpaired images. It does so by substituting the L1 loss by a cycle consistency loss. For that purpose, it requires a set of two discriminators (D_X, D_Y) and two generators (G, F) that work the same as in GANs and do the mapping from one image to the other and vice versa. What is formally described in equation 6, the cycle consistency loss, is that whatever is input into the first generator has to be able to come back as equal as possible after the backward mapping of the second generator.

$$G^*, F^* = \arg min_{G,F}\, max_{D_X D_Y} \mathcal{L}(G, F, D_X, D_Y) \quad \text{Eq. 4}$$

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F) \quad \text{Eq. 5}$$

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\| F\big(G(x)\big) - x \|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\| G\big(F(y)\big) - y \|_1] \quad \text{Eq. 6}$$

The full objective is described in equation 5 and the aim is to solve equation 6.
Equation 5 uses the loss function defined in equation 2 of Pix2Pix albeit twice for a mapping from x to y and vice versa plus the cyclic loss instead of the L1 loss. Again, λ can be found as a weight to control the importance of the cyclic loss during training.
Finding the equilibrium for such complex optimization requires a large amount of data. The CycleGAN technique is useful when no paired data is available. A sketch of half of the training process of a CycleGAN is shown in Figure 6. The same scheme is trained with input B on the same, but switched, Generators and Discriminators in parallel.

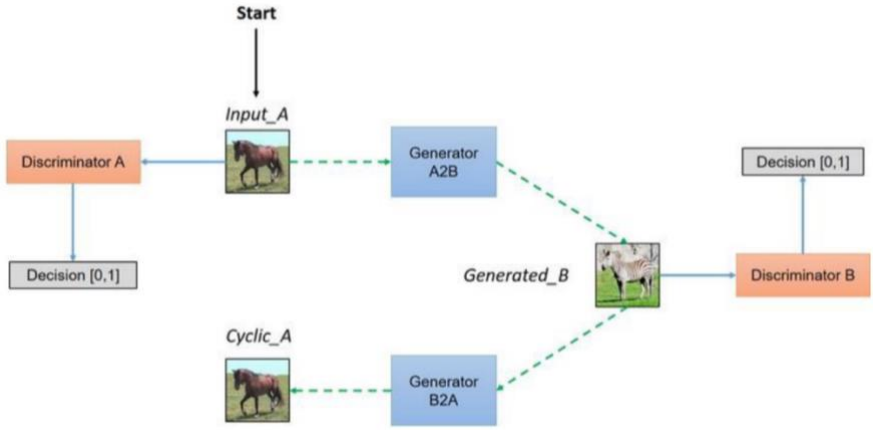


Figure 6. One-way CycleGAN schematic. Picture taken from (Zhu et al., 2020)

3. STATE OF THE ART

A literature search was conducted to find out which deep learning tools have been used for the purpose of reducing variabilities between MRI scanners. The results that were found are meant to be used in the design of the best possible solution for quantitative MRI harmonization.
First, it is first important to mention that the available literature on cross scanner harmonization is not extensive, especially with the focus set only on Deep Learning. Model based techniques were largely ignored. Examples of the latter techniques include mathematical or acquisition based approaches such as (Fortin et al., 2018)(Jiang et al., 2018)(Vemulapalli et al., 2015)(Hengameh Mirzaalian et al., 2018)(H. Mirzaalian et al., 2016).
The most explored topic, concerning both mathematical and machine learning approaches, appears to be diffusion MRI data harmonization. This could be due to a challenge (Harvard University, 2017) organized by the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI 2017) to encourage the investigation of the optimal homogenization approach. Diffusion tensor MRI is quantitative technique, but its scans show variability caused by differences in applied acquisition protocols

amongst other factors (Vollmar et al., 2010)(Landman et al., 2011). Some of the deep learning approaches suggested in the literature are based on the spherical properties that diffusion MRI possesses (Tax et al., 2019) (Chantal M.W.Tax, n.d.) as reviewed in (Pinto et al., 2020).
A different approach was suggested by (Moyer et al., 2020) for diffusion MRI scanner is dependency elimination. This procedure consists on learning what would make it possible to map the image taken from one scanner into one that looks as if it came from another scanner. This is then used to subtract from each image the characteristics of one or the other scanner. The outcome of this would be the creation of scanner-independent images, a step forward to the original idea of this project to induce the dependency of only one scanner. Furthermore, variational autoencoders (Kingma & Welling, 2019) are used for the same purpose. Variational autoencoders are the other possible type of generative models and they are less used due to their high degree of complexity.
Some relevant examples (Xiang et al., 2018)(Modanwal et al., 2020)(Dewey et al., 2019) can be found in the literature for techniques different to diffusion MRI.
In (Modanwal et al., 2020), targeted harmonizing breast tissue from a Siemens to a GE Healthcare scanner with Dynamic contrast-enhanced (DCE) MRI. To do so, they used a Cycle GAN discriminator with a smaller field of view than usual to help maintain anatomy. To check the success of the model, the authors do a qualitative assessment of the images by the eye and also quantitatively based on the mean intensity value distribution before and after harmonization. Results show good harmonization performance and visually convincing images while there are clearly observable differences between the images produced by each scanner (see Figure 7).

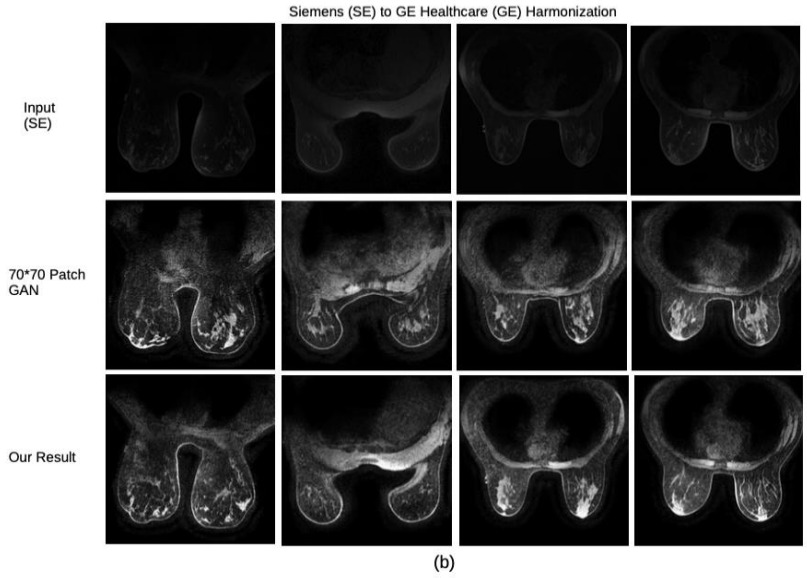


Figure 7. Proposed Image Harmonization results for Siemens to GE Healthcare in (Modanwal et al., 2020)

In (Dewey et al., 2019), the authors aimed at harmonizing images of different modalities between protocols and scanners. Prior to using any deep learning tools, supervised pre-processing algorithms were applied to homogenize and align the images. For the harmonization purpose a U-net is used, which is a standardized CNN originally designed for image segmentation (Ronneberger et al., 2015). After a first harmonization step, a second one is required for noise reduction. The validation methods applied were Mean Structure Similarity Index (SSIM) and Mean Absolute Error (MAE).
Another paper (Xiang et al., 2018) attempted to harmonize data from a 3T and a 7T scanner. To do so, a CycleGan with a modified generator for faster performance was used. A structural dissimilarity loss was added to the loss function that was optimized during training of the network. This derived from the objective of further preserving anatomical details further. Validation was performed with three statistical measures: peak signal to noise ratio (PSNR), SSIM and Normalized Mean Square Error (NMSE). Xiang et al concluded that satisfactory harmonization is achieved through these statistical measures.

4. CONCLUSIONS

The aim of this literature study was to find the best possible approach to quantitative MR inter-scanner harmonization using deep learning approaches. In particular, the goal was to train a network such that, an image from a specific patient can be acquired in one scanner and this can be transformed as if the patient had been scanned in another scanner.

Only deep learning methods were considered as that currently appears to be the most promising approach. Therefore, all the mathematical based approaches were discarded.

Since diffusion MRI methods are often based on spherical harmonics models, due to the nature of the acquisition technique, those results are not appropriate for this project either.

The method suggested by (Moyer et al., 2020) seems like an ideal approach. Nevertheless, this technique is very new and has not been explored in detail nor compared with further studies.

Furthermore, the CycleGAN approach appears as a fitting option both our data (which is unpaired) and appears to yield good performance. Success seems to be achieved with many combinations of generator and discriminator, which should be investigated experimentally or through other sources. To do so, measures like the SSIM, NMSE or MAE can be used to verify if harmonization has been achieved as desired.

5. BIBLIOGRAPHY

Activation Functions—ML Glossary documentation. (n.d.). Retrieved October 8, 2020, from https://ml-cheatsheet.readthedocs.io/en/latest/activation_functions.html

AlBadawy, E. A., Saha, A., & Mazurowski, M. A. (2018). Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical Physics*, *45*(3), 1150–1158. https://doi.org/10.1002/mp.12752

Andrew J. Taylor, Salerno, M., Dharmakumar, R., & Jerosch-Herold, M. (2016). *T1 Mapping | Elsevier Enhanced Reader*. https://doi.org/10.1016/j.jcmg.2015.11.005

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., & Yang, B. (2020). MedGAN: Medical image translation using GANs. *Computerized Medical Imaging and Graphics*, *79*, 101684. https://doi.org/10.1016/j.compmedimag.2019.101684

Bellman, R. E. (2015). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Ben-Cohen, A., Klang, E., Raskin, S. P., Soffer, S., Ben-Haim, S., Konen, E., Amitai, M. M., & Greenspan, H. (2018). Cross-Modality Synthesis from CT to PET using FCN and GAN Networks for Improved Automated Lesion Detection. *ArXiv:1802.07846 [Cs]*. http://arxiv.org/abs/1802.07846

Bergeest, J.-P., & Jäger, F. (2008). A Comparison of Five Methods for Signal Intensity Standardization in MRI. In T. Tolxdorff, J. Braun, T. M. Deserno, A. Horsch, H. Handels, & H.-P. Meinzer (Eds.), *Bildverarbeitung für die Medizin 2008* (pp. 36–40). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78640-5_8

Brownlee, J. (2019a, August 1). How to Develop a Pix2Pix GAN for Image-to-Image Translation. *Machine Learning Mastery*. https://machinelearningmastery.com/how-to-develop-a-pix2pix-gan-for-image-to-image-translation/

Brownlee, J. (2019b, August 8). How to Develop a CycleGAN for Image-to-Image Translation with Keras. *Machine Learning Mastery*. https://machinelearningmastery.com/cyclegan-tutorial-with-keras/

Brunner, P., & Ernst, R. R. (1979). Sensitivity and performance time in NMR imaging. *Journal of Magnetic Resonance (1969)*, *33*(1), 83–106. https://doi.org/10.1016/0022-2364(79)90192-6

Chantal M.W.Tax. (n.d.). *Cross-scanner and cross-protocol multi-shell diffusion MRI data harmonization: Algorithms and results | Elsevier Enhanced Reader*. https://doi.org/10.1016/j.neuroimage.2020.117128

Chavhan, G. B., Babyn, P. S., Thomas, B., Shroff, M. M., & Haacke, E. M. (2009). Principles, Techniques, and Applications of T2*-based MR Imaging and Its Special Applications. *RadioGraphics*, *29*(5), 1433–1449. https://doi.org/10.1148/rg.295095034

Chen, J., Chen, J., Chao, H., & Yang, M. (2018). Image Blind Denoising with Generative Adversarial Network Based Noise Modeling. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3155–3164. https://doi.org/10.1109/CVPR.2018.00333

Cohen, J. P., Luck, M., & Honari, S. (2018). Distribution Matching Losses Can Hallucinate Features in Medical Image Translation. *ArXiv:1805.08841 [Cs]*. http://arxiv.org/abs/1805.08841

Dar, S. U. H., Yurt, M., Karacan, L., Erdem, A., Erdem, E., & Çukur, T. (2018). Image Synthesis in Multi-Contrast MRI with Conditional Generative Adversarial Networks. *ArXiv:1802.01221 [Cs]*. http://arxiv.org/abs/1802.01221

Deoni, S. C. L., Williams, S. C. R., Jezzard, P., Suckling, J., Murphy, D. G. M., & Jones, D. K. (2008). Standardized structural magnetic resonance imaging in multicentre studies using quantitative T1 and T2 imaging at 1.5 T. *NeuroImage*, *40*(2), 662–671. https://doi.org/10.1016/j.neuroimage.2007.11.052

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., Saidha, S., Oh, J., Pham, D. L.,

Calabresi, P. A., van Zijl, P. C. M., & Prince, J. L. (2019). DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic Resonance Imaging*, *64*, 160–170. https://doi.org/10.1016/j.mri.2019.05.041
Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78. https://doi.org/10.1145/2347736.2347755
Dong, S., Luo, G., Wang, K., Cao, S., Mercado, A., Shmuilovich, O., Zhang, H., & Li, S. (2018). VoxelAtlasGAN: 3D Left Ventricle Segmentation on Echocardiography with Atlas Guided Generation and Voxel-to-voxel Discrimination. *ArXiv:1806.03619 [Cs]*. http://arxiv.org/abs/1806.03619
Fetty, L., Löfstedt, T., Heilemann, G., Furtado, H., Nesvacil, N., Nyholm, T., Georg, D., & Kuess, P. (2020). Investigating conditional GAN performance with different generator architectures, an ensemble model, and different MR scanners for MR-sCT conversion. *Physics in Medicine & Biology*, *65*(10), 105004. https://doi.org/10.1088/1361-6560/ab857b
Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M., & Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024
Fullerton, G. D. (1987). Magnetic resonance imaging signal concepts. *RadioGraphics*, *7*(3), 579–596. https://doi.org/10.1148/radiographics.7.3.3448647
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (n.d.). *Generative Adversarial Nets*. 9.
Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G., Cai, J., & Chen, T. (2017). Recent Advances in Convolutional Neural Networks. *ArXiv:1512.07108 [Cs]*. http://arxiv.org/abs/1512.07108
Hagiwara, A., Hori, M., Cohen-Adad, J., Nakazawa, M., Suzuki, Y., Kasahara, A., Horita, M., Haruyama, T., Andica, C., Maekawa, T., Kamagata, K., Kumamaru, K. K., Abe, O., & Aoki, S. (2019). Linearity, Bias, Intrascanner Repeatability, and Interscanner Reproducibility of Quantitative Multidynamic Multiecho Sequence for Rapid Simultaneous Relaxometry at 3 T: A Validation Study With a Standardized Phantom and Healthy Controls. *Investigative Radiology*, *54*(1), 39–47. https://doi.org/10.1097/RLI.0000000000000510
Hagiwara, A., Warntjes, M., Hori, M., Andica, C., Nakazawa, M., Kumamaru, K. K., Abe, O., & Aoki, S. (2017). SyMRI of the Brain: Rapid Quantification of Relaxation Rates and Proton Density, With Synthetic MRI, Automatic Brain Segmentation, and Myelin Measurement. *Investigative Radiology*, *52*(10), 647–657. https://doi.org/10.1097/RLI.0000000000000365
Harvard University. (2017). *Diffusion MRI Data Harmonisation (2017)*. https://projects.iq.harvard.edu/cdmri2017/challenge
Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv:1502.03167 [Cs]*. http://arxiv.org/abs/1502.03167
Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2018). Image-to-Image Translation with Conditional Adversarial Networks. *ArXiv:1611.07004 [Cs]*. http://arxiv.org/abs/1611.07004
Jackson, E. F., Ginsberg, L. E., Schomer, D. F., & Leeds, N. E. (1997). A review of MRI pulse sequences and techniques in neuroimaging. *Surgical Neurology*, *47*(2), 185–199. https://doi.org/10.1016/s0090-3019(96)00375-8
Jagtap, R. (2020, June 25). *A Comprehensive Guide to Generative Adversarial Networks (GANs)*. Medium. https://towardsdatascience.com/a-comprehensive-guide-to-generative-adversarial-networks-gans-fcfe65d1cfe4
Jiang, D., Liu, P., Li, Y., Mao, D., Xu, C., & Lu, H. (2018). Cross-vendor harmonization of T_2-relaxation-under-spin-tagging (TRUST) MRI for the assessment of cerebral venous oxygenation: Cross-Vendor Harmonization of TRUST MRI. *Magnetic Resonance in Medicine*, *80*(3), 1125–1131. https://doi.org/10.1002/mrm.27080
Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., & Song, M. (2018). Neural Style Transfer: A Review. *ArXiv:1705.04058 [Cs, Eess, Stat]*. http://arxiv.org/abs/1705.04058
Joseph P. Hornak. (1996). *The Basics of MRI*. https://www.cis.rit.edu/htbooks/mri/inside.htm
Kandel, I., & Castelli, M. (2020). The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*, *6*(4), 312–315. https://doi.org/10.1016/j.icte.2020.04.010
Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., Bae, H., & Kim, N. (2019). Deep Learning in Medical Imaging. *Neurospine*, *16*(4), 657–668. https://doi.org/10.14245/ns.1938396.198
Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017). Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. *ArXiv:1703.05192 [Cs]*. http://arxiv.org/abs/1703.05192
Kingma, D. P., & Welling, M. (2019). An Introduction to Variational Autoencoders. *ArXiv:1906.02691 [Cs, Stat]*. https://doi.org/10.1561/2200000056

Kransdorf, M. J., & Murphey, M. D. (2000). Radiologic Evaluation of Soft-Tissue Masses. *American Journal of Roentgenology*, *175*(3), 575–587. https://doi.org/10.2214/ajr.175.3.1750575
Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A. D., Bogovic, J. A., Hua, J., Chen, M., Jarso, S., Smith, S. A., Joel, S., Mori, S., Pekar, J. J., Barker, P. B., Prince, J. L., & van Zijl, P. C. M. (2011). Multi-parametric neuroimaging reproducibility: A 3-T resource study. *NeuroImage*, *54*(4), 2854–2866. https://doi.org/10.1016/j.neuroimage.2010.11.047
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539
Liao, H., Huo, Z., Sehnert, W. J., Zhou, S. K., & Luo, J. (2018). Adversarial Sparse-View CBCT Artifact Reduction. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, & G. Fichtinger (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (pp. 154–162). Springer International Publishing. https://doi.org/10.1007/978-3-030-00928-1_18
Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets. *ArXiv:1411.1784 [Cs, Stat]*. http://arxiv.org/abs/1411.1784
Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C. E., Morey, R. A., Flashman, L. A., George, M. S., McAllister, T. W., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R. D., Coleman, M. J., Kubicki, M., Westin, C. F., … Rathi, Y. (2016). Inter-site and inter-scanner diffusion MRI data harmonization. *NeuroImage*, *135*, 311–323. https://doi.org/10.1016/j.neuroimage.2016.04.041
Mirzaalian, Hengameh, Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Karmacharya, S., Grant, G., Marx, C. E., Morey, R. A., Flashman, L. A., George, M. S., McAllister, T. W., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R. D., Coleman, M. J., Kubicki, M., … Rathi, Y. (2018). Multi-site harmonization of diffusion MRI data in a registration framework. *Brain Imaging and Behavior*, *12*(1), 284–295. https://doi.org/10.1007/s11682-016-9670-y
Modanwal, G., Vellal, A., Buda, M., & Mazurowski, M. (2020). *MRI image harmonization using cycle-consistent generative adversarial network*. https://doi.org/10.1117/12.2551301
Moyer, D., Ver Steeg, G., Tax, C. M. W., & Thompson, P. M. (2020). Scanner invariant representations for diffusion MRI harmonization. *Magnetic Resonance in Medicine*, *84*(4), 2174–2189. https://doi.org/10.1002/mrm.28243
Murphy, M. (2011). *Parallelism, Patterns, and Performance in Iterative MRI Reconstruction*.
Pierpaoli, C. (2010). Quantitative Brain MRI. *Topics in Magnetic Resonance Imaging : TMRI*, *21*(2), 63. https://doi.org/10.1097/RMR.0b013e31821e56f8
Pinto, M. S., Paolella, R., Billiet, T., Van Dyck, P., Guns, P.-J., Jeurissen, B., Ribbens, A., den Dekker, A. J., & Sijbers, J. (2020). Harmonization of Brain Diffusion MRI: Concepts and Methods. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00396
Qayyum, A., Anwar, S., Majid, M., Awais, M., & Alnowami, M. (2018). Medical Image Analysis using Convolutional Neural Networks: A Review. *Journal of Medical Systems*. https://doi.org/10.1007/s10916-018-1088-1
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv:1505.04597 [Cs]*. http://arxiv.org/abs/1505.04597
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7
Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556 [Cs]*. http://arxiv.org/abs/1409.1556
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.
SyntheticMR. (n.d.). Retrieved September 23, 2020, from https://syntheticmr.com/products/symri-image/
Tax, C. MW., Grussu, F., Kaden, E., Ning, L., Rudrapatna, U., John Evans, C., St-Jean, S., Leemans, A., Koppers, S., Merhof, D., Ghosh, A., Tanno, R., Alexander, D. C., Zappalà, S., Charron, C., Kusmia, S., Linden, D. EJ., Jones, D. K., & Veraart, J. (2019). Cross-scanner and cross-protocol diffusion MRI data harmonisation: A benchmark database and evaluation of algorithms. *NeuroImage*, *195*, 285–299. https://doi.org/10.1016/j.neuroimage.2019.01.077
Vemulapalli, R., Nguyen, H. V., & Zhou, S. K. (2015). Unsupervised Cross-Modal Synthesis of Subject-Specific Scans. *2015 IEEE International Conference on Computer Vision (ICCV)*, 630–638. https://doi.org/10.1109/ICCV.2015.79
Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., Duncan, J. S., Richardson, M. P., & Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. *NeuroImage*, *51*(4), 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046
Wang, J., Zhao, Y., Noble, J. H., & Dawant, B. M. (2018). Conditional Generative Adversarial Networks for Metal Artifact Reduction in CT Images of the Ear. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, & G.

Fichtinger (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (pp. 3–11). Springer International Publishing. https://doi.org/10.1007/978-3-030-00928-1_1
Warntjes, J. B. M., Leinhard, O. D., West, J., & Lundberg, P. (2008). Rapid magnetic resonance quantification on the brain: Optimization for clinical usage. *Magnetic Resonance in Medicine*, *60*(2), 320–329. https://doi.org/10.1002/mrm.21635
Xiang, L., li, Y., Lin, W., & Wang, Q. (2018). *Unpaired Deep Cross-Modality Synthesis with Fast Training: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings* (Vol. 11045, pp. 155–164). https://doi.org/10.1007/978-3-030-00889-5_18
Yang, Q., Li, N., Zhao, Z., Fan, X., Chang, E. I.-C., & Xu, Y. (2018). MRI Cross-Modality NeuroImage-to-NeuroImage Translation. *ArXiv:1801.06940 [Cs]*. http://arxiv.org/abs/1801.06940
Yann Lecun, Patrick Haffner, & Y. Bengio. (2000). *Object Recognition with Gradient-Based Learning*. https://www.researchgate.net/publication/2816141_Object_Recognition_with_Gradient-Based_Learning
Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, *58*, 101552. https://doi.org/10.1016/j.media.2019.101552
Yves Chauvin & David E. Rumelhart. (1995). *Backpropagation: Theory, Architectures, and Applications*. Psychology Press.
Zhao, H., Li, H., & Cheng, L. (2017). Synthesizing Filamentary Structured Images with GANs. *ArXiv:1706.02185 [Cs]*. http://arxiv.org/abs/1706.02185
Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2020). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *ArXiv:1703.10593 [Cs]*. http://arxiv.org/abs/1703.10593

Annex B – Extra results

Experiment 1

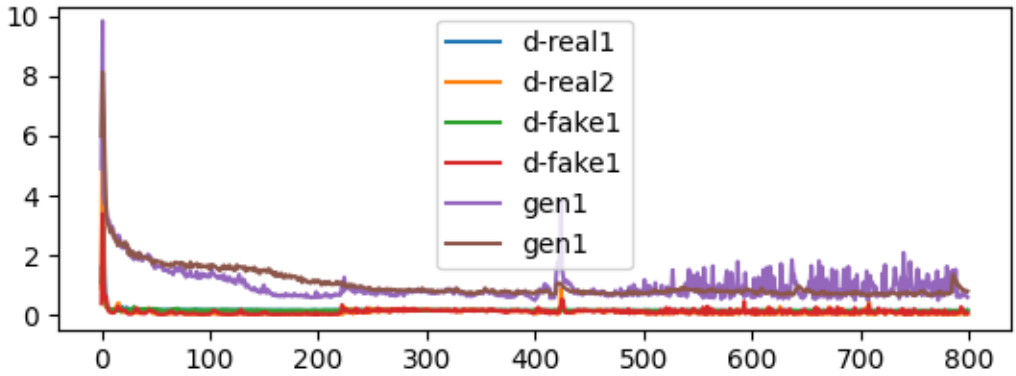


Figure B1. Training loss values as a function of epochs in Experiment 1

Experiment 2

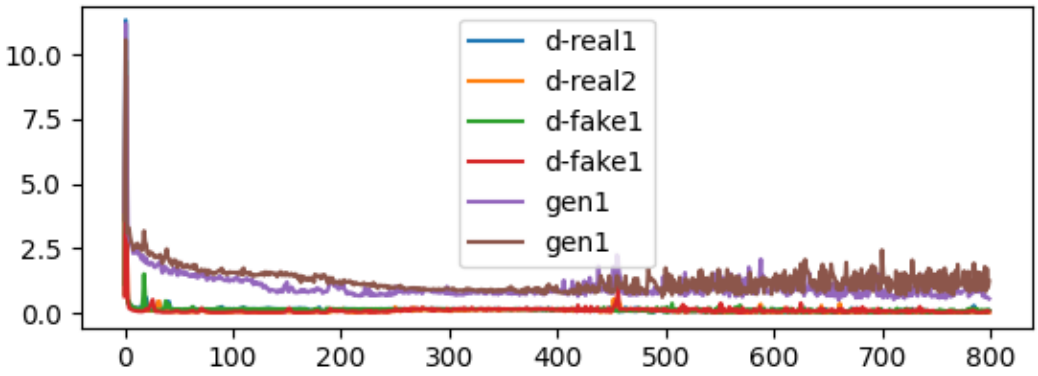


Figure B2. Training loss values as a function of epochs in Experiment 2

Experiment 3

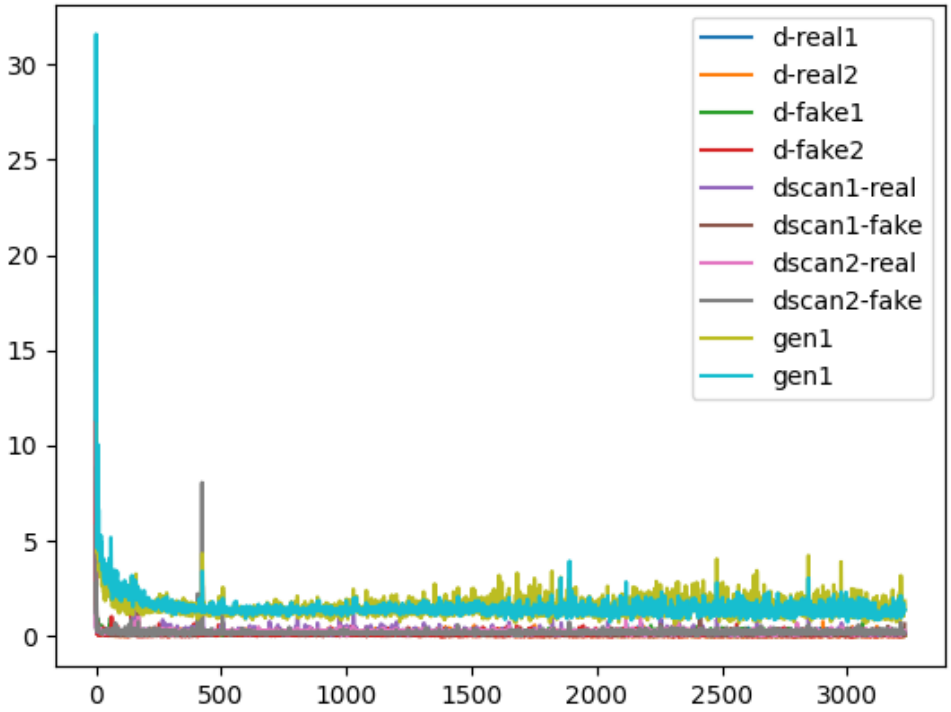


Figure B3. Training loss values as a function of epochs in Experiment 3

Experiment 4

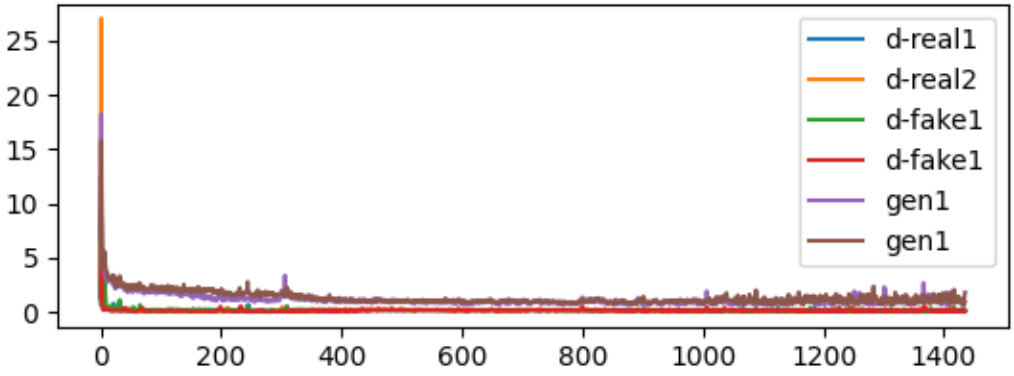


Figure B4. Training loss values as a function of epochs in Experiment 4

Experiment 5

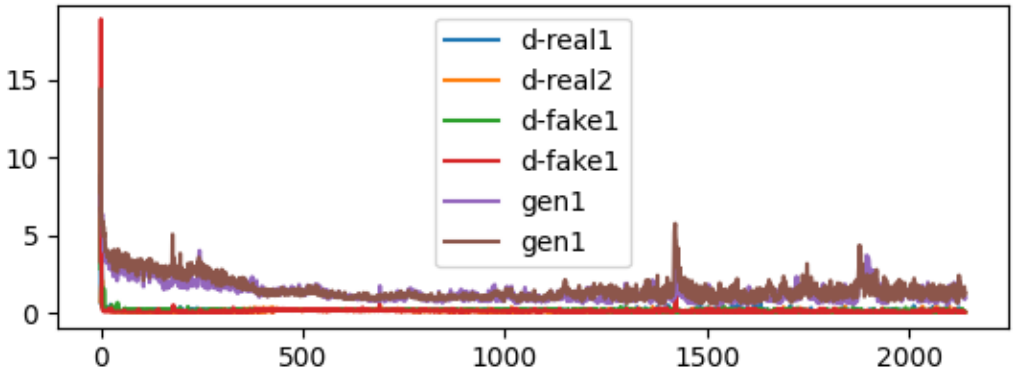


Figure B5. Training loss values as a function of epochs in Experiment 5

Experiment 6

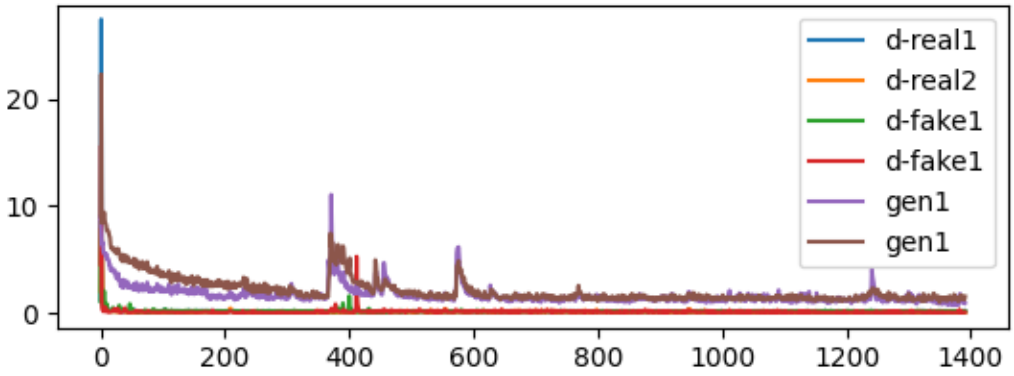


Figure B6. Training loss values as a function of epochs in Experiment 6