Deep Learning for Post-Contrast T1-Weighted Brain MRI Synthesis

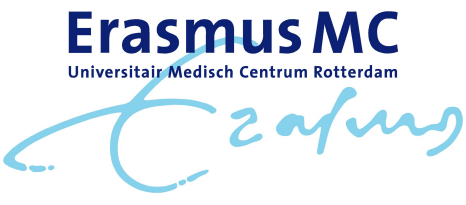
MSc Thesis Biomedical Engineering - BM51035
Medical Physics

by

Ruben C. van Oosterhoudt

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended on Monday February 27, 2023 at 9:00 AM.

Chair:	Dr. F.M. Vos,	Erasmus MC, TU Delft
Supervisors:	Prof. J.A. Hernandeze-Tamanes,	Erasmus MC, TU Delft
	Dr. S.R. van der Voort,	Erasmus MC
Independent committee member:	Dr. J.F. Veenland,	Erasmus MC



Deep Learning for Post-Contrast T1-Weighted Brain MRI Synthesis

R.C. van Oosterhoudt[1], S.R. van der Voort[2], J.A. Hernández-Tamames[2, 3], and F.M. Vos[2, 3]

[1] *Master student Biomedical Engineering, Technical University Delft, The Netherlands*
[2] *Department of Radiology and Nuclear medicine, Erasmus MC, University Medical Center, Rotterdam, the Netherlands*
[3] *Department of Imaging Physics, Delft University of Technology, Delft, The Netherlands*

Abstract

Introduction: Magnetic Resonance Imaging is a commonly used technique for the initial diagnosis of gliomas. T1, T2, T2-FLAIR, and post-contrast T1 with gadolinium-based contrast agents (GBCAs) can show tumor characteristics. However, using this contrast agent poses a risk to patients with kidney failures, has environmental impact, and increases cost. To address these issues, we aimed to evaluate the potential of deep learning in generating synthetic post-contrast T1 images without using contrast agents.

Method: The project investigates the potential of using deep learning (DL) to generate synthetic post-contrast T1 images based on T1, T2, and T2-FLAIR provided by the Erasmus Glioma Database. Exploring different model architectures, loss functions, and input sizes to discover the optimal approach.

Results: Results show that individual loss functions, input size, and model complexity slightly impact the accuracy of synthetic post-contrast T1 images. Combining loss functions, however, was the most promising approach for image generation. Models trained with $\mathcal{L}_M$ could generate low detail enhancement. Resulting in 0.0478 ± 0.0076, 0.0139 ± 0.0036, and 0.879 ± 0.024 for MAE, MSE, and SSIM, respectively.

Conclusion: The study's findings indicate that DL is promising for generating synthetic post-contrast T1 images without using GBCAs. However, further research is required to generate realistic synthetic post-contrast T1 images. The study, however, provides a basis for future work and highlights the importance of reducing the use of GBCAs in clinical practice.

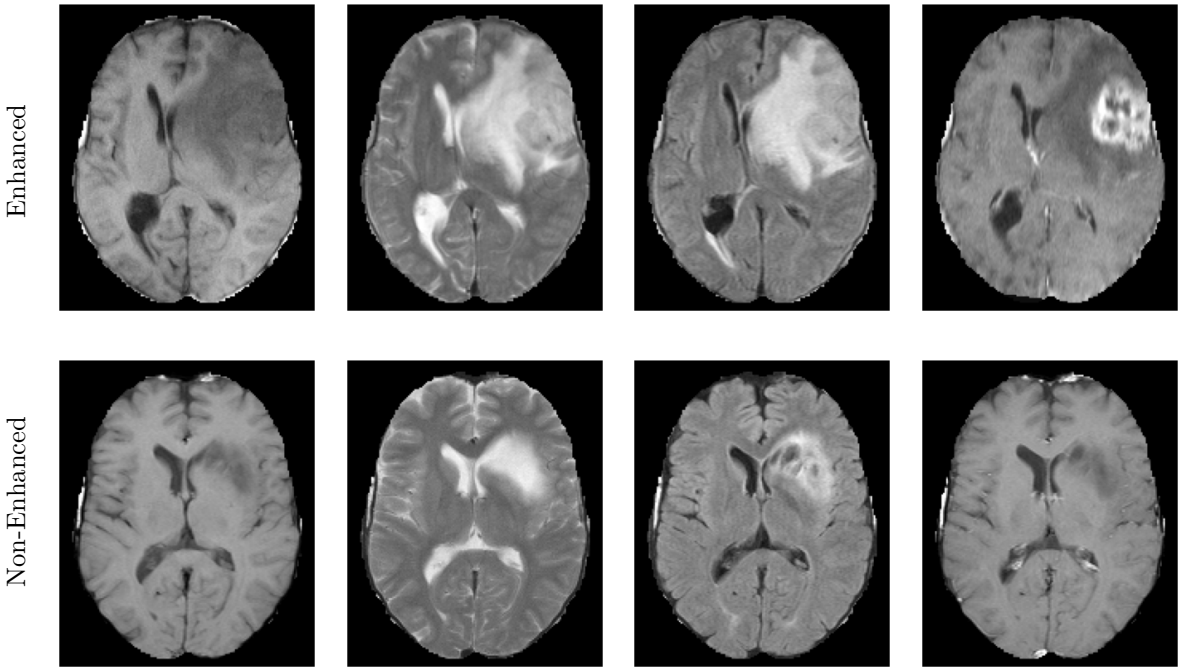
Index terms— U-Net, gliomas, post-contrast T1, synthetic, enhancement, GBCAs. MRI, deep learning

1 Introduction

In 2019, almost 120,000 new cancer cases were diagnosed in the Netherlands, of which 1.2% concerned a form of brain cancer [1]. Brain cancer patients have a relatively low survival, ranging from less than a year to 10+ years [2, 3]. Conventionally, the applied treatment is based on disease characteristics [4]. In the case of gliomas (a type of brain cancer), treatments are based on characteristics such as detected genetic mutations and pathological inspection. The same characteristics are also used to classify gliomas into grades II to IV, which standard was set by the World Health Organization (WHO) [5], with higher grades corresponding to more aggressive tumors and lower survival rates. In literature, it has become a convention that grades II and III are also classified as low-grade gliomas (LGG) and IV as high-grade gliomas (HGG). An initial distinction between LGG and HGG could be determined based on MR imaging.

Magnetic Resonance Imaging (MRI) is used for initial glioma diagnosis by acquiring T1, T2, T2-FLAIR, and post-contrast T1 images (cT1), shown in figure 1. The cT1 modality is acquired after gadolinium injection and is known to be particularly useful for assessing the aggressiveness of gliomas [6, 7, 8]. HGG gliomas grow more aggressively, compared to LGG, causing blood-brain barrier leakage [9]. As such gadolinium-based contrast agents (GBCAs) can transcend the BBB and deposit in the extravascular space, resulting in an enhancement in the cT1 images. However, in rare cases, LGG can also show enhancement on cT1 images. That is why a biopsy is essential for accurate diagnosis. While gadolinium is essential during initial diagnosis, clinical doctors aim to reduce or even replace gadolinium due to observed side effects.

Importantly, GBCAs were shown to impact both patients and the environment negatively. Recent studies show that the linear form of GBCAs can deposit in the skin, bone, liver, and other organs, although the implications of this deposition remain to be investigated [10, 11, 12]. For patients with kidney failure, gadolinium can, in rare events, cause nephrogenic systemic fibrosis [13]. Another problem of GBCAs is the environmental impact [14, 15]. Approximately 95-98% of GB-



(a) T1 (b) T2 (c) T2-FLAIR (d) cT1

Figure 1: Different modalities, each containing specific information, utilized for brain tumor diagnostics. a) T1 shows the general structure, b) T2 shows fluids, c) T2-FLAIR shows inflammation in tissues, and d) cT1 shows leakage of blood vessels caused by tumor aggression, which is the most significant difference between enhanced and non-enhanced.

CAs are excreted in the sewerage system within 24h[14]. Furthermore, several studies have shown that GBCA deposits, via the sewerage, in the water, soil, and living organisms [15]. Just as the deposition in the body, the implications of this remains unknown. Another problem of gadolinium is financially related. Taking an extra cT1 comes with a higher cost, due to the additional scan time and material use of GBCAs. In order to avoid inducing kidney failure, to reduce gadolinium deposition in the body, to reduce the discharge of GBCAs in the environment, and to reduce the cost, an alternative to cT1 with GBCAs is highly preferred.

Enhancing images with deep learning (DL) is a promising approach to reduce or even avoid the intake of GBCAs. In the past, DL has been utilized to generate different modalities. For example, Bahrami et al. proposed a method that translates 3T images to 7T-like images [16]. Furthermore, Conte et al. focused more on translation between modalities by generating T1 and T2-FLAIR images based on cT1 and T2 with two generative adversarial networks (GANs) [17]. Similar to our work (see below), Gong et al. proposed a U-Net-based model to reduce the injection volume by 90% [18]. The model could generate full-dose cT1 images based on T1 and cT1 data with only 10% of the original injection dose. To avoid using GBCAs, Chen et al. proposed a fully connected network (FCN) [19]. The network generated cT1 images based on T1, T2, and apparent diffusion coefficient maps (ADC). The inclusion of ADC proved to have a slight advantage over only utilizing T1 and T2.

Further research is needed, however, to explore the possibility of completely excluding gadolinium. As such, we propose to apply different U-Net-based models that generate synthetic cT1 (scT1) images based on T1, T2, and T2-FLAIR images. We do so as the latter images are conventionally already at the basis of a comprehensive assessment of low-grade glioma.

2 Methods

This project investigated the potential of using DL to generate scT1 images based on T1, T2, and T2-FLAIR from the EGD dataset (2.1). To do so, we explored using different loss functions, model architectures, and input sizes. Different methodologies were studied regarding data processing (2.2),

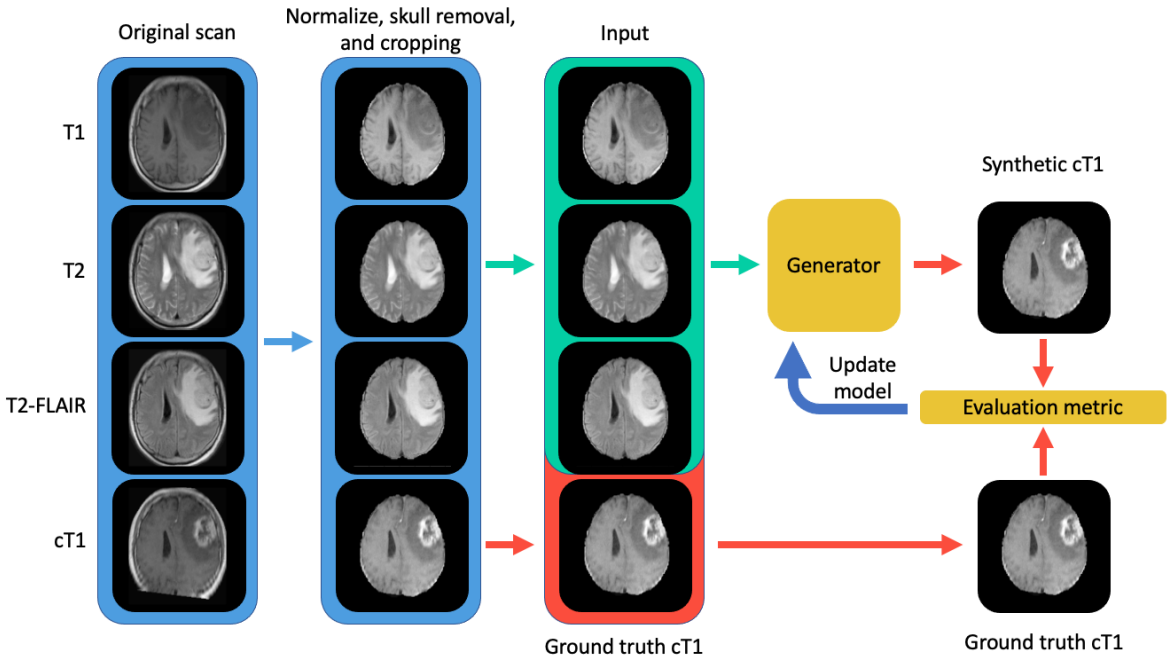


Figure 2: Detailed workflow from the original data to the generation of scT1 images. The original data is already registered to the MNI152 atlas. Normalization was applied to remove variations in intensity. Removal of the skull and cropping was applied to remove non-relevant voxels. Then only T1, T2, and T2-FLAIR are fed to the model to generate scT1 images. The scT1 images are then compared to the orginal cT1 images to evaluate and train the model.

data augmentation (2.3), model architecture (2.4), optimization (2.5), loss functions (2.6), evaluation metrics (2.7), and radiomics (2.8). Our data workflow is visually represented in figure 2.

2.1 Data

We used T1, T2, and T2-FLAIR images, from the Erasmus Glioma Database (EGD), that were acquired during routine clinical practice. The EGD is a database collected retrospectively from patients treated for gliomas at the Erasmus MC between 2008 and 2018. It consisted of 774 patients (281 female, 492 male, 1 unknown) ranging from 19 to 86 years of age [20]. T1, T2, T2-FLAIR, and cT1 images were acquired preoperatively for every patient. The images were obtained with scanners from four vendors: Siemens (347), Phillips (254), GE (172), and Toshiba (1). The field strength varied as follows: 3T (83), 1.5T (571), 1T (110), 0.5 (6), and unknown (4). While the patients were treated at the Erasmus MC, a fraction of the images were acquired at other institutes. After image acquisition, a biopsy or resection was performed to determine gene mutation and codeletion for precise diagnosis. A pathological assessment of the biopsy or resection specimen determined the tumor grade based on the WHO 2016 grading, resulting in: II (135), III (79), IV (502), and unknown (58) [5]. A tumor segmentation was available for each image, either manual (374) or automatic (400).

2.2 Data pre-processing

Pre-processing involved preparing the data for learning by registration, cropping, filtering, and cleaning it in various ways to reduce non-relevant variations between images. Registration is applied to remove variations in spatial coordinates and resolution between acquisitions and patients. All images were registered to the MNI152 brain atlas, with a size of $197 \times 233 \times 189$ with voxels of 1 mm^3 [21]. To remove non-relevant voxels, the skull was removed and then cropped to $160 \times 192 \times 160$ voxels. Subsequently, intensity differences between images were compensated by applying Z-score normalization, with a 4σ cut-off, and scaled between [-1, 1]. To further improve data quality, 313 images with artifacts were removed after a visual assessment of all data. Such artifacts in images consisted of low signal-to-noise (SNR), insufficient field-of-view (FOV), Gibbs artifacts, and signs of movement.

Based on the remaining 461 images, we imposed an even 50% split between enhancing and non-enhancing tumors, so that 298 images remained.

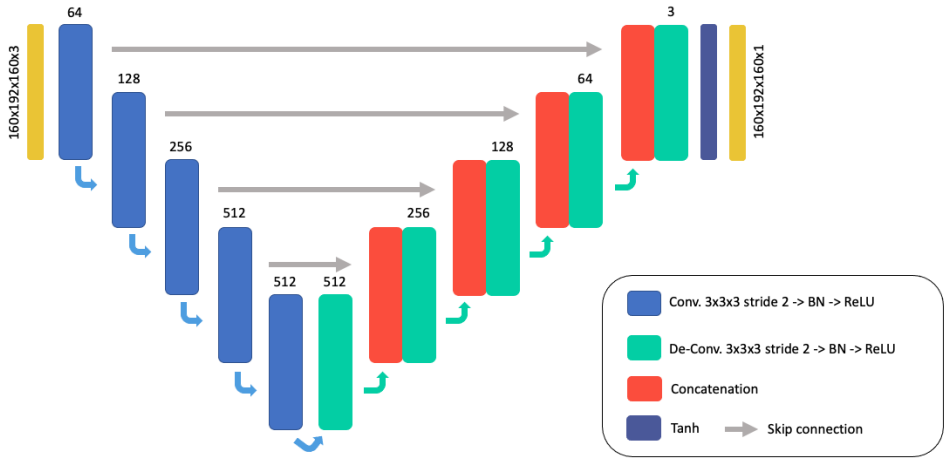


Figure 3: Overview of the U-Net architecture based on the pix2pix generator interpretation of TensorFlow. T1, T2, and T2-FLAIR were utilized as input for the model to generate synthetic cT1 images. Further referred to as our baseline model or model 0.

This even distribution should prevent biases in the model prediction. Cohen et al. highlighted the issue of hallucinating tumors in their generated T1 and FLAIR images due to this imbalance [22]. In this way, the model might generate tumors that are not there. The resulting balanced dataset was subsequently partitioned into training, testing, and evaluation datasets of 208, 45, and 45 images, respectively.

To focus our attention on the tumor, we implemented patching. This patching involved reducing the data into smaller regions, or so-called patches, centered around the tumor segmentation's center of mass (COM). We took this approach for two reasons. First, it facilitated a reduction of voxels during training, so that the computation time was reduced. Second, non-relevant voxels were discarded, and a focus on the tumor and its surroundings was created. As such, centering around the tumor and decreasing the volume would increase the influence of enhancing voxel during training. Based on this approach, three additional datasets were created by making patches of sizes 64^3, 96^3, and 128^3 voxels. The influence of patch-size on performance was explicitly studied in several experiments.

2.3 Data augmentation

Due to removing images with artifacts and limiting the influence of biases during training, the dataset contained 298 samples. Data augmentation was performed by flipping the brain along its sagittal axis to increase the number of samples. In effect, this doubled the number of samples to 596.

2.4 Model architecture

As a DL model, we focused on convolutional neural networks (CNN) because of their good performance in many image processing applications. The U-Net, a particular type of CNN, has, besides segmentation applications, proven its capabilities in image-to-image (I2I) translation [23]. I2I aims to transform images from one domain to another. In our case, this would be from T1, T2, and T2-FLAIR to a synthetic cT1 image. As a baseline U-Net model, we used a modified version of the pix2pix generator implemented by TensorFlow, as shown in figure 3 [24, 25]. Instead of using 2D convolution filters, 3D convolution filters were used to allow the model to capture 3D information. 3D models can include information of neighboring slices, which is known to increase classification accuracy [19]. Model input and output channels were changed to 3 (T1, T2, and T2-FLAIR) and 1 (scT1). Furthermore, due to the output range ([-1, 1]), the sigmoid activation function at the end was replaced by a tanh.

Four other model architectures were created to study the efficiency of variations of the baseline model ('Model 0') further, as follows.

Model 1: fine-tuned baseline

Model 1 addresses some bottlenecks in the model by varying the parameter initialization, upsampling, kernel size, and high-resolution skip connection. Parameter initialization changed from normal initialization to Xavier initialization to improve learning [26]. He-initialization was also considered but performed worse after some initial

testing [27]. Upsampling in the last layer consisted of only 3 filters. We suspected that the lack of filters caused checkerboard artifacts in the generated images. Therefore, additional filters (3 $\rightarrow$ 64) and a convolution layer with 3 filters were added after the last upsample layer to prevent this. Additionally, the filter size was adjusted from 4 to 3, decreasing learning time. Also, a full-scale skip connection was included to transfer high-resolution information over the U-Net architecture. To achieve this, an additional convolution layer was added at the start with 64 filters to match the last upsampling layer. The full architecture of the model is shown in appendix A.

Model 2: modality specific

Further building on the adaptations of model 1, model 2 includes additional modality-specific convolutional layers at the front of the model [28]. The input is split in T1, T2, and T2-FLAIR and fed to modality-specific convolution layers. Each convolution layer learns a set of modality-specific filters to extract modality-specific features rather than applying the same filters irrespective of the modality. A convolutional layer of 32 filters was added for each modality, followed by batch normalization and a ReLU activation layer. The layout of this model is presented in appendix B.

Model 3: attention-based

In model 3, based on model 2, attention layers are implemented before concatenating the upsampled information and skip-connection [29]. In these layers, the model learns to focus its attention, in our case, on the tumor to increase model accuracy. The architecture of this model is presented in appendix C.

Model 4: residual-based

For model 4, based on model 2, architecture complexity was increased to be able to learn more complex connections within the data. Increasing complexity is achieved by implementing residual layers before the current convolution and deconvolutional layers. Residual layer design is based on the work of He et al. [30]. The principle of residual layers is to learn the difference, or residual, between the input and output, rather than the full transformation. The model is presented in appendix D.

2.5 Optimization

During training, the Adam optimizer was used for all models with hyper-parameters copied from the pix2pix model (beta1=0.5, beta2=0.99, and a learning rate of 2e-4)[24, 31]. Alternative hyper-parameter configurations were tested but did not improve accuracy. A maximum of 300 epochs was set for each model. We implemented a stopping criteria to stop the model after 20 consecutive epochs without improvement.

2.6 Loss functions

It is our aim to transform from T1, T2, and T2-FLAIR to synthetic post-contrast T1 images (scT1) through our generator. Therefore, cT1 (ground truth) and scT1 are compared by a loss function that evaluates the model's performance. Three types of loss functions were evaluated: pixel-, area-wise, and a combination. The loss was determined based on either a single voxel, or an area surrounding a voxel. As such, the pixel-wise loss focuses on individual errors between images, while the area-wise approach reflects human perception.

We applied the $\mathcal{L}_1$ and $\mathcal{L}_2$ as pixel-wise loss functions, also known as the mean absolute error (MAE) and mean squared error (MSE), respectively. The loss functions are defined as:

$$\mathcal{L}(X, Y; S)_{1S} = \frac{1}{|S|} \sum_{u \in S} |G(X)_u - Y_u| \quad (1)$$

$$\mathcal{L}(X, Y; S)_{2S} = \frac{1}{|S|} \sum_{u \in S} (G(X)_u - Y_u)^2 \quad (2)$$

in which $G(X)$ corresponds to the generation of synthetic cT1 images, based on the input images X, Y are the ground truth cT1 images, and S is a binary mask. The loss functions were applied to the full images ($\mathcal{L}_1$ and $\mathcal{L}_2$), and regions restricted to the brain ($\mathcal{L}_{1B}$, $\mathcal{L}_{2B}$), the tumor ($\mathcal{L}_{1T}$ and $\mathcal{L}_{2T}$), and the enhancing part ($\mathcal{L}_{1E}$ and $\mathcal{L}_{2E}$) using masks. The brain and tumor masks were readily available from a previous study. Brain segmentation was based on the MNI152 atlas, while the tumor mask was either manually or automatically segmented. The enhancement mask was made by subtracting the registered T1 image from the cT1 image, as in the work of Chen et al. [19]. Positive differences larger than 0.3 were considered to represent enhancement. The threshold of 0.3 was empirically determined (through visual inspection). Observe that by limiting the loss functions as such, we increasingly set a focus on the region of interest (ROI).

For the area-wise loss functions, Structural Similarity Index Metric (SSIM) or $\mathcal{L}_{SSIM}$. Essentially, $\mathcal{L}_{SSIM}$ evaluated the similarity of cT1 and scT1 in contrast, structure, and illumination over a 11x11x11 area as described in [32]:

$$\mathcal{L}(X, Y; S)_{SSIM} = \frac{1}{|S|} \sum_{u=S} |1 - \text{SSIM}(G(X_u), Y_u)| \tag{3}$$

SSIM has a minimum value of 0 and a maximum of 1 when X equals Y. We subtracted SSIM from 1 to define a minimizing problem similar to the previously described loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$.

Finally, a combination loss function takes advantage of both voxel- and area-wise loss functions, combining individual voxel errors and human perception. This loss was based on the work of Chen et al. and was defined as [19]:

$$\mathcal{L}_M = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{1E} \tag{4}$$

with λ_1, λ_2, and λ_3 parameters that can be used to adjust the contribution of each individual loss function. Chen et al. initially applied an even distribution of 1 for λ_1, λ_2, and λ_3

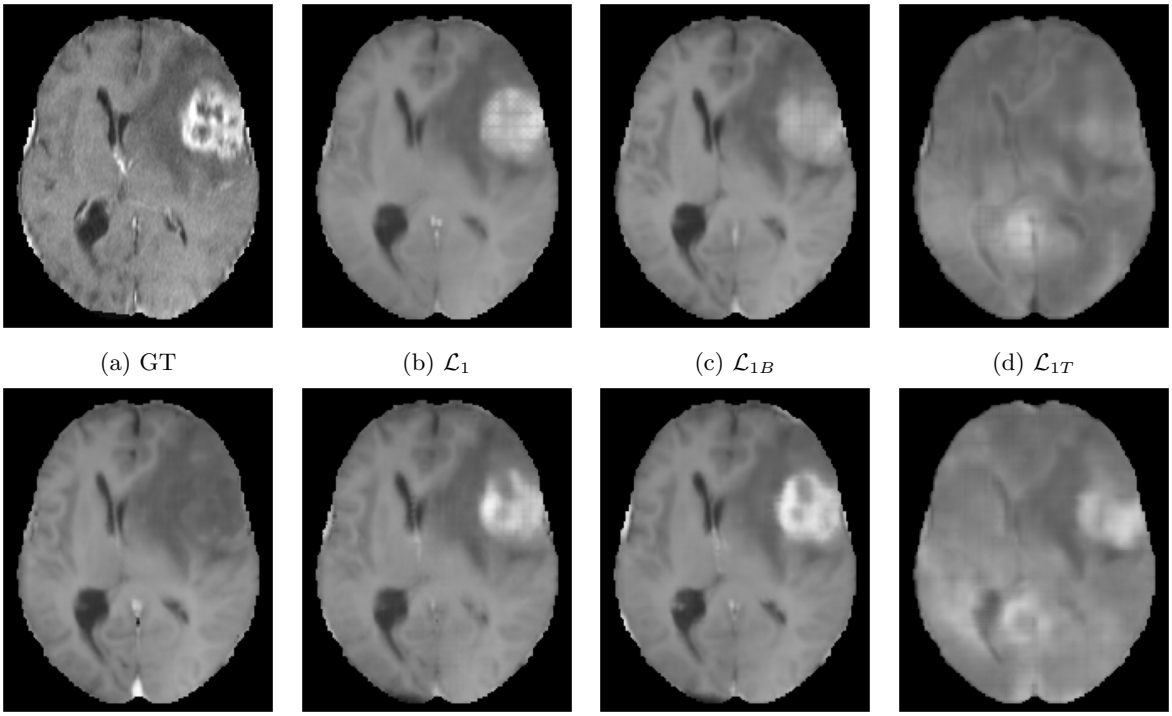
Clearly, the optimal loss functions, as with many parameters in deep learning, may be highly problem specific. To cope with this, a study was conducted to evaluate the different loss functions. Results were evaluated using evaluation metrics and visual inspection.

2.7 Evaluation metrics

For assessing the model performance, quantitative and qualitative evaluation metrics were used. Specifically, we applied the mean absolute error (MAE), mean squared error (MSE), and structural similarity index metrics (SSIM) to do so, which are commonly used metrics in image-to-image translation. In addition, similar to the loss function, a brain and tumor mask was imposed to focus these metrics. To assess the images qualitatively, I myself inspected the images regarding the presence of artifacts.

2.8 Radiomics

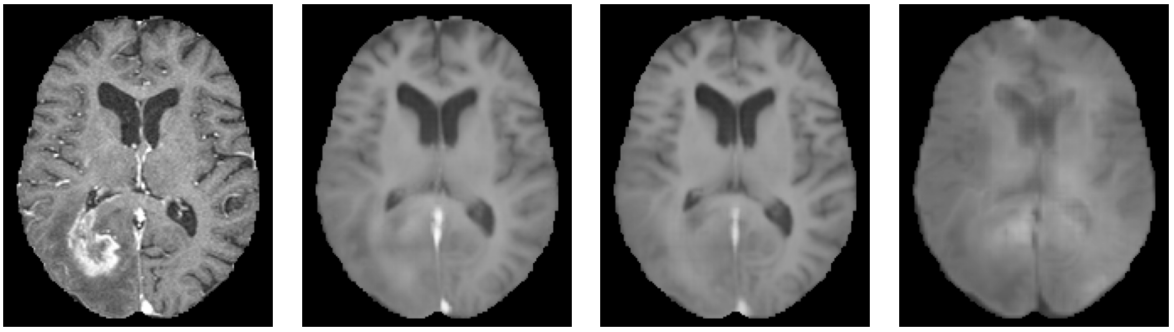
In addition to the above deep learning techniques, we implemented a radiomics approach. Radiomics analyzes quantifiable features from medical images intending to predict disease progression, treat-



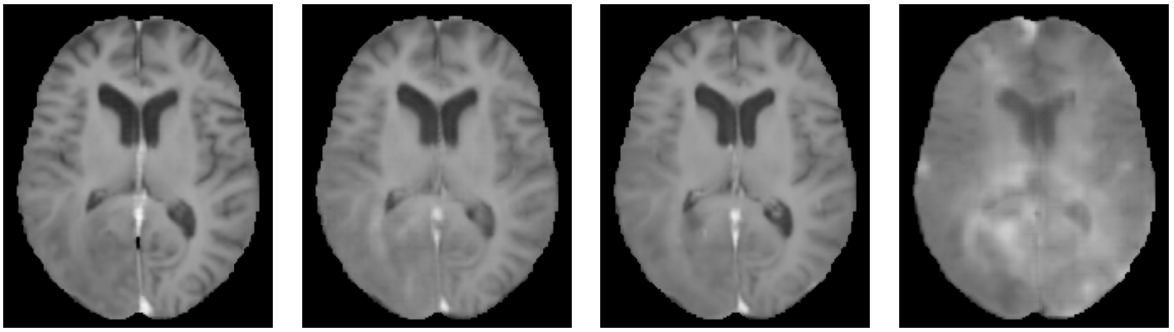
(a) GT (b) $\mathcal{L}_1$ (c) $\mathcal{L}_{1B}$ (d) $\mathcal{L}_{1T}$

(e) $\mathcal{L}_{SSIM}$ (f) $\mathcal{L}_2$ (g) $\mathcal{L}_{2B}$ (h) $\mathcal{L}_{2T}$

Figure 4: Qualitative assessment of the baseline model with different loss functions. These images were generated based on a already seen sample from the *training* dataset and compared to the ground truth a).



(a) GT (b) $\mathcal{L}_1$ (c) $\mathcal{L}_{1B}$ (d) $\mathcal{L}_{1T}$



(e) $\mathcal{L}_{SSIM}$ (f) $\mathcal{L}_2$ (g) $\mathcal{L}_{2B}$ (h) $\mathcal{L}_{2T}$

Figure 5: Qualitative assessment of the baseline model with different loss functions. These images were generated based on a new sample from the *Evaluation* dataset and compared to the ground truth a).

Table 1: Quantitative comparison between the different loss functions with the same baseline model with the evaluation dataset. Full table is in appendix E.

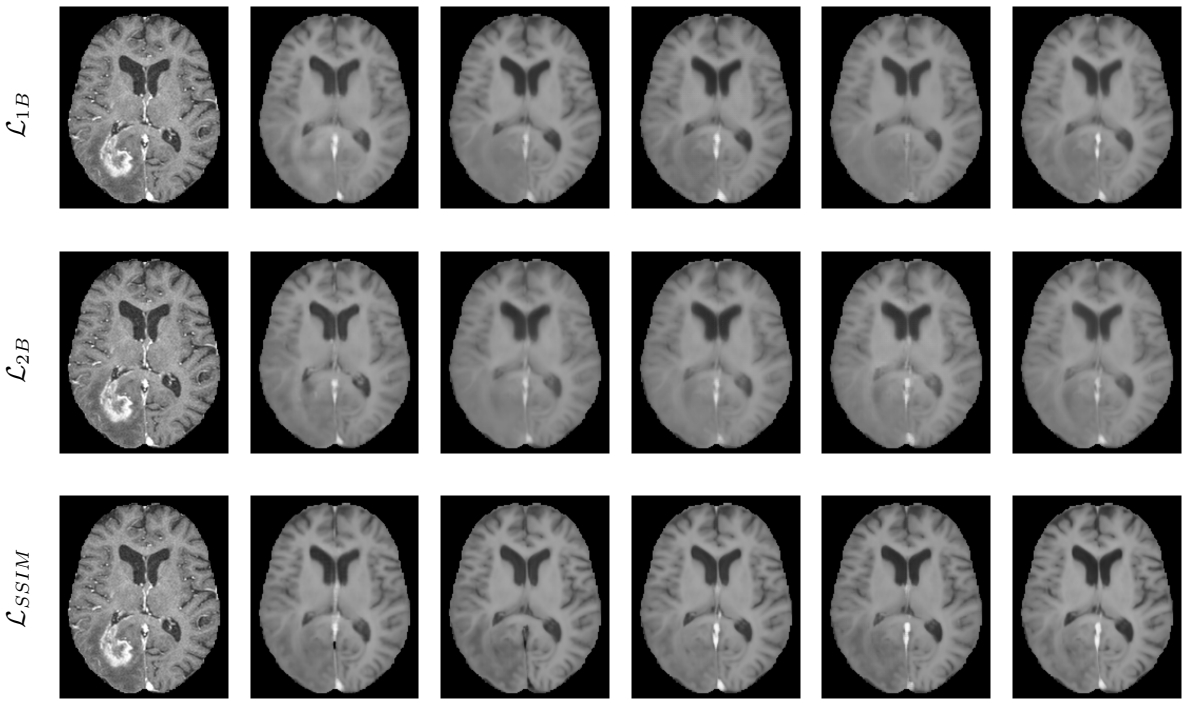
	$MAE_T\downarrow$	$MSE_T\downarrow$	$SSIM_T\uparrow$	$MAE_T\downarrow$	$MSE_T\downarrow$	$SSIM_T\uparrow$	$MAE_T\downarrow$	$MSE_T\downarrow$	$SSIM_T\uparrow$
$\mathcal{L}_1$	0,0474	0,0138	0,874	0,1207	0,0352	0,727	0,1655	0,0647	0,632
$\mathcal{L}_2$	0,0481	**0,0136**	0,872	0,1214	**0,0345**	0,724	0,1715	**0,0644**	0,624
$\mathcal{L}_{1B}$	0,0476	0,0139	0,875	0,1215	0,0356	0,729	0,1675	0,0657	0,633
$\mathcal{L}_{2B}$	0,0480	0,0137	0,875	0,1224	0,0349	0,728	0,1738	0,0661	0,622
$\mathcal{L}_{1T}$	0,0670	0,0236	0,822	0,1710	0,0603	0,618	0,1692	0,0674	0,610
$\mathcal{L}_{2T}$	0,0678	0,0239	0,822	0,1730	0,0611	0,619	0,1781	0,0662	0,609
$\mathcal{L}_{SSIM}$	**0,0470**	0,0140	**0,878**	**0,1200**	0,0358	**0,737**	**0,1621**	0,0687	**0,642**

ment, or diagnosis. Based on T1, T2, and T2-FLAIR, we aimed to merely predict if a post-contrast T1 weighted image would show enhancement, not the actual appearance. We utilized the radiomics software package WORC to achieve this [33, 34]. WORC is a standardized, modular framework that automatically optimizes the workflow for an application.

3 Results

3.1 Loss functions comparison

Initially, we investigated subjectively if the model generated enhancement within the training dataset, an example of which is shown in figure 4. The figure confirms that the model can learn enhancement based on T1, T2, and T2-FLAIR. For further model assessment, samples from the evaluation dataset were used, shown in figure 5. The figure shows that the method cannot predict enhancement in the validation dataset. The images were also objectively evaluated, of which



(a) *GT* (b) 0 (c) 1 (d) 2 (e) 3 (f) 4

Figure 6: We visually compared models 0-4 trained with either $\mathcal{L}_{1B}$, $\mathcal{L}_{2B}$, or $\mathcal{L}_{SSIM}$ to the ground truth a).

Table 2: Quantitave comparison for the different models and loss functions with a focus on the tumor region. Full table is in appendix F.

	$\mathcal{L}_{1B}$			$\mathcal{L}_{2B}$			$\mathcal{L}_{SSIM}$		
Model	MAE ↓	MSE ↓	SSIM ↑	MAE_B↓	MSE_B↓	SSIM_B↑	MAE_T↓	MSE_T↓	SSIM_T↑
0	0,167	0,066	0,633	0,174	0,066	**0,622**	0,162	0,069	0,642
1	0,164	0,064	0,631	0,173	0,064	0,616	0,165	0,071	0,644
2	**0,160**	0,068	0,636	0,173	0,064	0,619	**0,161**	**0,064**	0,648
3	0,168	0,064	0,622	0,175	0,064	0,617	0,163	0,070	0,643
4	0,164	**0,063**	**0,639**	**0,172**	0,064	0,620	0,169	0,068	**0,651**

results are presented in table 1; the full table of this assessment is in appendix E.

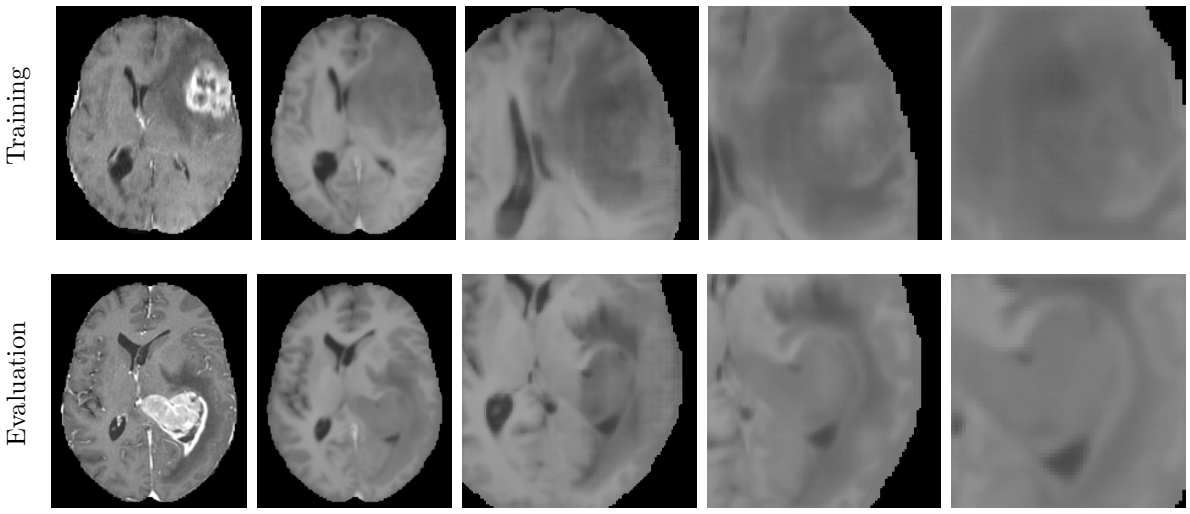
To focus our research $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_{1T}$, and $\mathcal{L}_{2T}$ were excluded from further research. $\mathcal{L}_1$ and $\mathcal{L}_2$ were omitted because $\mathcal{L}_{1B}$, and $\mathcal{L}_{2B}$ performed slightly better on visual inspection. $\mathcal{L}_{1T}$, and $\mathcal{L}_{2T}$ were not considered for further research due to overall poor performance. $\mathcal{L}_{SSIM}$ was included because it yielded high overall values in any evaluation performed; however, simultaneously, it did not generate enhancement in the images.

3.2 Model comparison

In this section, the performance of five models trained with three loss functions is reviewed based on all evaluation metrics. To set our focus, only tumor-specific loss functions are presented in table 2. Complementary to this, a full table is in appendix F. Additionally, the models were evaluated based on visual assessment, as shown in figure 6.

3.3 Input size comparison

By decreasing the input size around the tumors COM, we effectively increase the relative number of enhancing voxels. Again, as in section 2.6, we aimed to focus on enhancing voxels. As described previously, we selected patch-sizes of 128, 96, and 64 centered around the COM of the mask. The results were evaluated based on visual assessment



(a) GT (b) full (c) 128 (d) 96 (e) 64

Figure 7: Model 1 trained with $\mathcal{L}_{1B}$ for different input sizes (full, 128, 96, and 64). As example the model was fed a training and evaluation samples and compared to the ground truth a).

Table 3: Quantitative comparison between the different input sizes with the same baseline model. Full table is in appendix G

	$\mathcal{L}_{1B}$			$\mathcal{L}_{2B}$			$\mathcal{L}_{SSIM}$		
Model	$MAE_T\downarrow$	$MSE_T\downarrow$	$SSIM_T\uparrow$	$MAE_T\downarrow$	$MSE_T\downarrow$	$SSIM_T\uparrow$	$MAE_T\downarrow$	$MSE_T\downarrow$	$SSIM_T\uparrow$
full	0,1641	**0,0639**	0,6307	0,1727	0,0644	**0,6157**	**0,1651**	**0,0710**	**0,6441**
128	**0,1621**	0,0670	**0,6369**	0,1759	0,0644	0,6140	0,1710	0,0772	0,6436
96	0,1627	0,0658	0,6314	**0,1722**	**0,0628**	0,6127	0,1689	0,0763	0,6409
64	0,1718	0,0714	0,6201	0,1799	0,0671	0,6108	0,1802	0,0865	0,6214

and evaluation metrics, shown in figure 7 and table 3.

3.4 Combined loss functions

A combination of loss functions ($\mathcal{L}_M$), as defined in formula 4, was used to generate synthetic cT1 images. By combining different loss functions, the distortion, visible in figures 5d) and 5h), could be solved when including the loss of the overall brain structure. For comparability, experiments were performed in the same way as in 3.1, enabling a comparison of the best-performing loss functions ($\mathcal{L}_{1B}$, $\mathcal{L}_{2B}$ and $\mathcal{L}_{SSIM}$) to $\mathcal{L}_M$. Table 4 and figure 8 show the quantitative and qualitative assessment.

3.5 Enhancement prediction

A WORC experiment was implemented to verify that the data holds the necessary information for enhancement prediction. T1, T2, and T2-FLAIR were considered as input for WORC, with labels 0 and 1, for non-enhancing and enhancing, respectively. To simplify the problem, grade IV tumors were considered enhancing, while grades II and III were assumed to be non-enhancing. This holds for the vast majority of the cases, but there are exceptions.

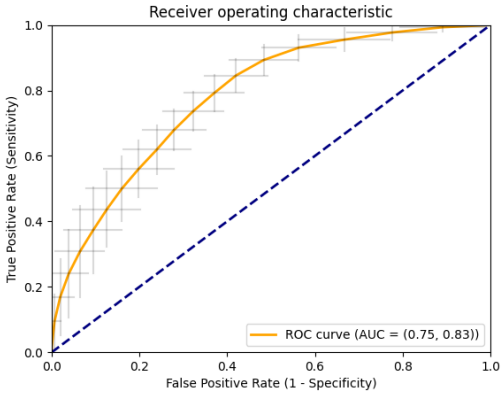
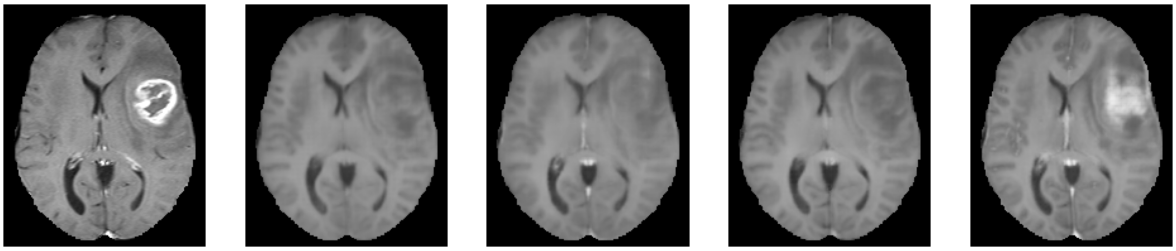


Figure 9: Receiver operating characteristic (ROC) curve of the radiomics package WORC with an average AUC of 0.79.



(a) GT (b) $\mathcal{L}_2$ (c) $\mathcal{L}_{SSIM}$ (d) $\mathcal{L}_M$ (e) $\mathcal{L}_2$

Figure 8: Qualitative comparison of model 0 trained with either $\mathcal{L}_{1B}$, $\mathcal{L}_{2B}$, $\mathcal{L}_{SSIM}$, or $\mathcal{L}_M$ to the ground truth.

Table 4: The models trained with different loss functions are evaluated with different evaluation metrics. The scores are the average of 45 evaluation samples generated by the different models.

	MAE ↓	MSE ↓	SSIM ↑	MAE_B↓	MSE_B↓	SSIM_B↑	MAE_T↓	MSE_T↓	SSIM_T↑
$\mathcal{L}_{1B}$	0,0476	0,0139	0,875	0,1215	0,0356	0,729	0,1675	0,0657	0,633
$\mathcal{L}_{2B}$	0,0480	**0,0137**	0,875	0,1224	**0,0349**	0,728	0,1738	**0,0661**	0,622
$\mathcal{L}_{SSIM}$	**0,0470**	0,0140	0,878	**0,1200**	0,0358	0,737	**0,1621**	0,0687	**0,642**
$\mathcal{L}_M$	0,0478	0,0139	**0,879**	0,1220	0,0356	**0,740**	0,1716	0,0696	0,630

4 Discussion

This study investigated whether deep learning can simulate contrast weighted T1 images based on (non-contrast weighted) T1, T2, and T2-FLAIR images. Various models, loss functions, and patch-sizes were employed to explore their effects. The findings and potential issues related to these methods are discussed in this chapter.

4.1 Loss focus

The conventional loss functions ($\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_{SSIM}$) appeared to lack focus on enhancement, as seen in figures 4 and 5. Essentially, the enhancing voxels are outnumbered by the healthy voxels, leading to an uneven contribution of inputs to the loss function. As a result, the models particularly learns from healthy voxels. Importantly, a model that learns mostly from healthy voxels appears to generate a healthy brain, as was previously also observed in the work of Cohen et al. [22]. By implementing brain and tumor-specific loss functions ($\mathcal{L}_{1B}$, $\mathcal{L}_{1T}$, $\mathcal{L}_{2B}$, and $\mathcal{L}_{2T}$) the weight of the enhancing voxels implicitly increased. In this way we enhanced the focus on diseased tissues, potentially improving accuracy in these regions. A model was trained with the different loss functions to test this hypothesis and was evaluated based on visual assessment and evaluation metrics.

Initially, figure 4 confirmed the capabilities of generating enhancement based on the training samples. Additionally, the experiment using the WORC environment confirmed that information on enhancement is contained within the input images, see figure 9: the mean AUC (0.79) is significantly different from random guessing. However, the model could not generate enhancement in the tumor based on evaluation samples. The only generated enhancement was in the great longitudinal fissure (fissure between both hemispheres), as seen in figure 5. Compared to tumor enhancement, the fissure enhancement is visible in most cases, so that it can be efficiently learned by the networks. In a few cases, however, this enhancement was not visible in the true contrast enhanced images. In those cases the predicted enhancement of the fissure should be considered a hallucination.

Both qualitative and quantitative assessments of the synthetic cT1 images confirmed that focusing the loss did not improve accuracy in the brain nor tumor, see figure 5 and table 1. The applied loss $\mathcal{L}_{SSIM}$ (figure 5e) resulted in relatively sharp images compared to the other loss functions but at the cost of enhancement in the training samples, as shown in figure 4e. For that reason, $\mathcal{L}_{1B}$ and $\mathcal{L}_{2B}$ (figures 5c and 5g) were considered the best-performing model that could generate enhancement on training samples. In the same way, we compared all these metrics on the evaluation dataset (table 1). The loss function $\mathcal{L}_{SSIM}$ generally showed the best performance, which corresponded with the visual assessment. Only $\mathcal{L}_2$ performed better regarding MSE, MSE_B, and MSE_T,

which we attribute to the close mathematical relation of $\mathcal{L}_2$ and MSE. Models trained with either $\mathcal{L}_{1T}$ or $\mathcal{L}_{2T}$ performed poor, as shown in figures 5d and 5h. These images demonstrate that excessive focus caused distortions due to insufficient information about brain structures. In conclusion, the models trained with different loss functions unfortunately could not generate enhancement in the tumors. However, slight differences between the loss functions were noticeable. In general, $\mathcal{L}_{1B}$, $\mathcal{L}_{2B}$, and $\mathcal{L}_{SSIM}$ were the best performing loss functions.

4.2 Model complexity

The results in section 4.1 showed no significant differences between loss functions. Therefore, we shifted our attention to variations in model architecture. Model architecture could limit model performance by being too shallow or non-specific. Specifically, shallow models may have insufficient parameters to precisely capture a problem, while non-specific models may not focus on the ROI. Both issues were addressed by implementing different attributes, as described in section 2.4. These different models were trained with $\mathcal{L}_{1B}$, $\mathcal{L}_{2B}$, and $\mathcal{L}_{SSIM}$ and evaluated with the evaluation metrics from section 2.7, with a focus on the tumor-specific metrics.

The scT1 images generated by the different models lacked enhancement. Only subtle differences were visible between models, as presented in table 2 and figure 6. Models were qualitatively assessed based on visual markers, such as overestimation of tumor area, under- and overestimation of enhancement in the great longitudinal fissure, high-intensity distortion, and presence of checkerboard artifacts. As can be seen in figure 6, the lower intensities around the tumor were overestimated compared to the original image in model 0. This Too high intensity decreased in the other models, becoming more similar to the ground truth, improving visual accuracy. With most models, there were cases of hallucinated enhancement in the great longitudinal fissure, for the same reason as described previously (4.1). In the case of the attention model, trained with $\mathcal{L}_{2B}$, distortion arose around high-intensity enhanced areas. The reason for this artifact could not be identified by us. A checkerboard artifact was visible in figure 4b and 5c for which additional convolutional layers were applied to prevent this. However, with models 2 and 3, this artifact was still visible in some generated images. Adding even more additional layers still did not solve this problem. At the same time, these artifacts were not visible in the results generated by models 1 and 4. To summarise, model complexity did not significantly improve overall model performance. Based on the shallow architecture and the lack of checkerboard artifacts, model 1 could be considered as the best-performing model.

4.3 Volume focus

Section 4.1 introduced brain and tumor-specific loss functions to cope with the lack of focus on enhancing voxels. Another approach was to implement patches to decrease the influence of healthy tissues. By selecting patches around the tumors, the model focus was targeted on the tumor and its close surrounding rather than the full image. We considered patch sizes of 64^3, 96^3, and 128^3 to investigate the influence of patching compared to exploiting the full images.

Unfortunately, visual inspection showed no improvement compared to the full image, and rather a decrease in performance. Figure 7c shows no enhancement in the tumor or great longitudinal fissure, checkerboard artifacts, and intensity differences between different results. Evaluation metrics confirm this decrease in performance, as seen in table 3. The decrease in the evaluation performance was inversely proportional to the patch size. By patching, we essentially selected the area surrounding the tumor containing the enhancement. However, these areas, due to the inability of generating enhancement, contained the largest error, resulting in a decrease in evaluation metric scores. In the end, using the full image size as input outperformed the other input sizes as reflected by the quantitative metrics.§

4.4 Mix loss function

In section 4.1, we concluded that models trained with a tumor-specific loss function did not improve model accuracy and introduced distortions in healthy tissues. Inspired by Chen et al., we implemented a mixed loss function that could focus on enhancement, preventing distortions as seen in figure 4d and 4h [19]. To reproduce the work of Chen et al. and compare it to our previous results, we trained model 0 with the same settings as in 4.1.

In figure 8, the model trained with the $\mathcal{L}_M$ was compared to the best-performing models from 4.1 ($\mathcal{L}_{1B}$, $\mathcal{L}_{2B}$, and $\mathcal{L}_{SSIM}$). It is visible that the model trained with $\mathcal{L}_M$ could generate a coarse representation of enhancement. In some cases, however, the model generates enhancement

on non-enhancing samples, creating false positives. While the figures clearly show differences, these are not noticeable in the evaluation metrics in table 4. While $\mathcal{L}_M$ outperforms in terms of SSIM and SSIM_B, SSIM_T did not improve. Given the enhancement in figure 8, we expected improvements in tumor-specific evaluation metrics. However, due to the coarse representation of the enhancement, a large area is essentially overestimated, decreasing the evaluation metric score.

4.5 Noise

Scanning through the samples, we noticed that some input images contained a significant amount of noise. Noise in images degrades the image's quality, in effect often limiting model performance. Deep learning models rely on large amounts of data to learn image patterns. However, when the images are degraded by noise, the model may learn incorrect patterns. Importantly, the level of noise also limits the approximation precision of the model. This limit can be determined from the variance in areas with "constant" intensity. These areas were selected based on visual inspection. The average variance over the evaluation samples was 0.029. This indicates that we are close to the limit with an MAE of ~ 0.05. The limit may be decreased by implementing denoising techniques. However, it is essential to note that reducing the noise can go at the expense of removing fine detail that should actually be predicted. Thus, a trade-off must be made between reducing the limit and preserving important features.

4.6 Registration

During the evaluation of synthetic cT1, it became clear that some errors were caused by the non-ideal registration of images. Therefore, it could be that structures between input and output did not align well. The model relies on the overall structure of the input to make an accurate prediction of cT1 images. However, the model cannot predict these random and unpredictable misalignments, limiting our model accuracy. An example of such misalignment can be found in appendix I, in which the image has a rotation difference and a right ventricle misalignment between T1 and cT1. To cope with this, the registration should be improved, or such images should be excluded during training and evaluation.

4.7 Evaluation metric limit

In most of our experiments, we would have preferred that the evaluation metrics were more conclusive, to enhance decision-making. Now we depend (too much) on visual assessment, making the results subjective. This was most evident in the results of 3.4. Apparent enhancement was visible in figure 8, but the evaluation metrics did not increase significantly in table 4.

A previous literature study showed that the most used evaluation metrics are MAE, MSE, and SSIM. No other evaluation metrics were found that were fitting with our data, and that amplified differences between models.

4.8 Loss complexity

In section 3.1, we showed that the models could initially not learn any enhancement. However, in section 3.4, the model was able to generate a coarse representation of the enhancement. One may observe that $\mathcal{L}_{1B}$, $\mathcal{L}_{2B}$, and $\mathcal{L}_{SSIM}$ are relatively simple loss functions for a complex problem like ours. It could be the case that these loss functions by themselves can not capture the detail needed to generate enhancement. Alternatively, potentially due to the increased complexity, the model trained with a combination of loss functions could generate enhancement, 3.4. However, the resolution was poorer than might be preferable.

5 Conclusion

We showed that it was possible to generate contrast-weighted T1 images using T1, T2, and FLAIR with model 0 and a mixed loss function. Compared to the work of Gong et al., our model performed slightly better regarding mean absolute error on training data:0.879 ± 0.024 compared to 0.85 ± 0.08, respectively. However, the proposed model by Gong et al. could also generate accurate enhancement in synthetic scans. Overall, the model previously proposed by Chen et al. had the best performance: 0.923 ± 0.041. An important aspect to note, however, is that our results were based on a weight of 1 for each loss function. Different weight combinations might be explored to optimize the accruacy.

Despite the efforts to optimize the models and loss functions, the results did not meet the desired performance. Further research and development of deep learning models and loss functions is needed to improve the accuracy and reliability of contrast-weighted T1 image generation. This might in par-

ticular require establishment of larger databases for training.

6 Further work

During the project, I implemented various approaches to achieve high-accuracy generated vcT1 scans. Some strategies could have further improved accuracy. Examples are EGD potential, Mix loss function, and model architecture.

The EGD dataset, in section 2.1, contains more than only MR images and masks. Sex, age, grade, gene mutation, and codeletion were known for each patient. Including these features during training could improve the model's performance and use the EGD dataset to its full potential.

During the project, the $\mathcal{L}_M$ outperforms all other loss functions without proper fine-tuning, as described in 4.4. For now, the implementation only considered an even weighing between loss functions. Performing an empirical study could lead to a more optimal weighing and increase model performance.

In section 4.2, we concluded that increasing model complexity did not improve accuracy. However, during the project, only U-Net was considered as a model. Another approach would have been a GAN-based architecture. Compared to the predefined loss function in U-Net, a GAN can learn its loss function during training. This approach could find more complex correlations in the data than a predefined loss function.

References

[1] "Integraal kankercentrum nederland." iknl.nl. (accessed: 04.01.2023).

[2] J. C. Buckner, "Factors influencing survival in high-grade gliomas," in *Seminars in oncology*, vol. 30, pp. 10–14, Elsevier, 2003.

[3] S. L. Hervey-Jumper and M. S. Berger, "Role of surgical resection in low-and high-grade gliomas," *Current treatment options in neurology*, vol. 16, no. 4, pp. 1–19, 2014.

[4] M. Verma, "Personalized medicine and cancer," *Journal of personalized medicine*, vol. 2, no. 1, pp. 1–14, 2012.

[5] P. Wesseling and D. Capper, "Who 2016 classification of gliomas," *Neuropathology and applied neurobiology*, vol. 44, no. 2, pp. 139–150, 2018.

[6] Y. Yang, M. Z. He, T. Li, and X. Yang, "Mri combined with pet-ct of different tracers to improve the accuracy of glioma diagnosis: a systematic review and meta-analysis," *Neurosurgical review*, vol. 42, no. 2, pp. 185–195, 2019.

[7] B. L. Dean, B. P. Drayer, C. R. Bird, R. A. Flom, J. A. Hodak, S. W. Coons, and R. G. Carey, "Gliomas: classification with mr imaging.," *Radiology*, vol. 174, no. 2, pp. 411–415, 1990.

[8] L. E. Ginsberg, G. N. Fuller, M. Hashmi, N. E. Leeds, and D. F. Schomer, "The significance of lack of mr contrast enhancement of supratentorial brain tumors in adults: histopathological evaluation of a series," *Surgical neurology*, vol. 49, no. 4, pp. 436–440, 1998.

[9] N. Nishida, H. Yano, T. Nishida, T. Kamura, and M. Kojiro, "Angiogenesis in cancer," *Vascular health and risk management*, vol. 2, no. 3, p. 213, 2006.

[10] W. A. Gibby, K. A. Gibby, and W. A. Gibby, "Comparison of gd dtpa-bma (omniscan) versus gd hp-do3a (prohance) retention in human bone tissue by inductively coupled plasma atomic emission spectroscopy," *Investigative radiology*, vol. 39, no. 3, pp. 138–142, 2004.

[11] S. Aime and P. Caravan, "Biodistribution of gadolinium-based contrast agents, including gadolinium deposition," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 30, no. 6, pp. 1259–1267, 2009.

[12] T. Grobner and F. Prischl, "Gadolinium and nephrogenic systemic fibrosis," *Kidney international*, vol. 72, no. 3, pp. 260–264, 2007.

[13] V. Gulani, F. Calamante, F. G. Shellock, E. Kanal, S. B. Reeder, *et al.*, "Gadolinium deposition in the brain: summary of evidence and recommendations," *The Lancet Neurology*, vol. 16, no. 7, pp. 564–570, 2017.

[14] K. Kümmerer and E. Helmers, "Hospital effluents as a source of gadolinium in the aquatic environment," *Environmental science*

& technology, vol. 34, no. 4, pp. 573–577, 2000.

[15] J. Rogowska, E. Olkowska, W. Ratajczyk, and L. Wolska, "Gadolinium as a new emerging contaminant of aquatic environments," *Environmental toxicology and chemistry*, vol. 37, no. 6, pp. 1523–1534, 2018.

[16] K. Bahrami, F. Shi, X. Zong, H. W. Shin, H. An, and D. Shen, "Reconstruction of 7t-like images from 3t mri," *IEEE transactions on medical imaging*, vol. 35, no. 9, pp. 2085–2097, 2016.

[17] G. M. Conte, A. D. Weston, D. C. Vogelsang, K. A. Philbrick, J. C. Cai, M. Barbera, F. Sanvito, D. H. Lachance, R. B. Jenkins, W. O. Tobin, *et al.*, "Generative adversarial networks to synthesize missing t1 and flair mri sequences for use in a multisequence brain tumor segmentation model," *Radiology*, vol. 299, no. 2, pp. 313–323, 2021.

[18] E. Gong, J. M. Pauly, M. Wintermark, and G. Zaharchuk, "Deep learning enables reduced gadolinium dose for contrast-enhanced brain mri," *Journal of magnetic resonance imaging*, vol. 48, no. 2, pp. 330–340, 2018.

[19] C. Chen, C. Raymond, W. Speier, X. Jin, T. F. Cloughesy, D. Enzmann, B. M. Ellingson, and C. W. Arnold, "Synthesizing mr image contrast enhancement using 3d high-resolution convnets," *IEEE Transactions on Biomedical Engineering*, 2022.

[20] S. R. van der Voort, F. Incekara, M. M. Wijnenga, G. Kapsas, R. Gahrmann, J. W. Schouten, H. J. Dubbink, A. J. Vincent, M. J. van den Bent, P. J. French, *et al.*, "The erasmus glioma database (egd): Structural mri scans, who 2016 subtypes, and segmentations of 774 patients with glioma," *Data in brief*, vol. 37, p. 107191, 2021.

[21] V. Fonov, A. Evans, R. McKinstry, C. Almli, and D. Collins, "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood," *NeuroImage*, vol. 47, p. S102, 2009. Organization for Human Brain Mapping 2009 Annual Meeting.

[22] J. P. Cohen, M. Luck, and S. Honari, "Distribution matching losses can hallucinate features in medical image translation," in *International conference on medical image computing and computer-assisted intervention*, pp. 529–536, Springer, 2018.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[25] "pix2pix: Image-to-image translation with a conditional gan." tensorflow.org/tutorials/generative/pix2pix. (accessed: 08.02.2023).

[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[28] G. van Tulder, "Shifting representations: Adventures in cross-modality domain adaptation for medical image analysis," 2022.

[29] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[33] M. P. A. Starmans, T. Phil, S. R. van der Voort, and S. Klein, "Workflow for optimal radiomics classification (worc)," 2018.

[34] M. P. A. Starmans, S. R. van der Voort, T. Phil, M. J. M. Timbergen, M. Vos, G. A. Padmos, W. Kessels, D. Hanff, D. J. Grunhagen, C. Verhoef, S. Sleijfer, M. J. van den Bent, M. Smits, R. S. Dwarkasing, C. J. Els, F. Fiduzi, G. J. L. H. van Leenders, A. Blazevic, J. Hofland, T. Brabander, R. A. H. van Gils, G. J. H. Franssen, R. A. Feelders, W. W. de Herder, F. E. Buisman, F. E. J. A. Willemssen, B. G. Koerkamp, L. Angus, A. A. M. van der Veldt, A. Rajicic, A. E. Odink, M. Deen, J. M. C. T., J. Veenland, I. Schoots, M. Renckens, M. Doukas, R. A. de Man, J. N. M. IJzermans, R. L. Miclea, P. B. Vermeulen, E. E. Bron, M. G. Thomeer, J. J. Visser, W. J. Niessen, and S. Klein, "Reproducible radiomics through automated machine learning validated on twelve clinical applications," 2021.

Appendix A Figure: Model 1

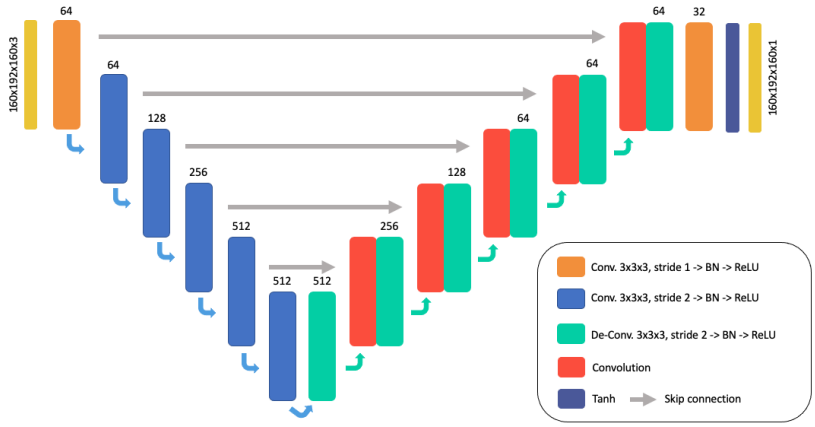


Figure 10: U-Net architecture based on the pix2pix interpretation of TensorFlow. The number above the blocks represent the number of filters in each layer. Compared to model 0, we parameter initialization, upsampling, kernel size, and high-resolution skip connection.

Appendix B Figure: Model 2

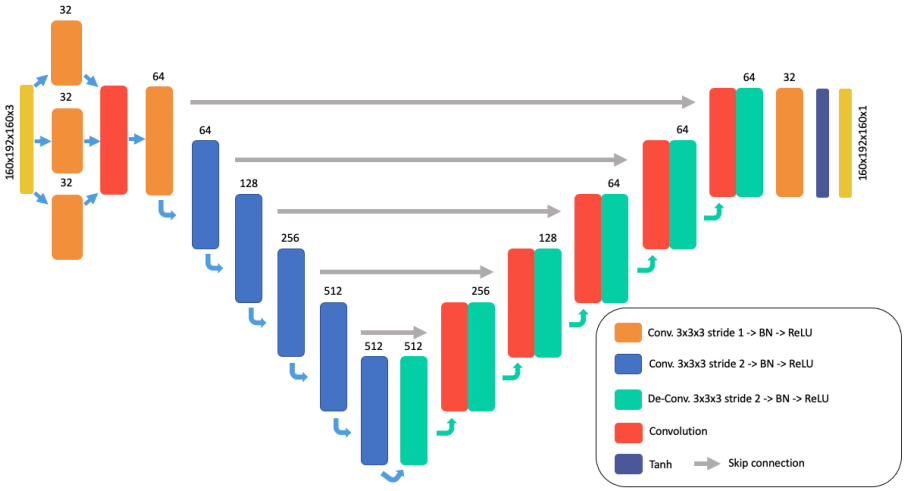


Figure 11: U-Net architecture based on the pix2pix interpretation of TensorFlow. The number above the blocks represent the number of filters in each layer. Compared to model 1, we added modality specific convolution layers to learn model-specific features.

Appendix C Figure: Model 3

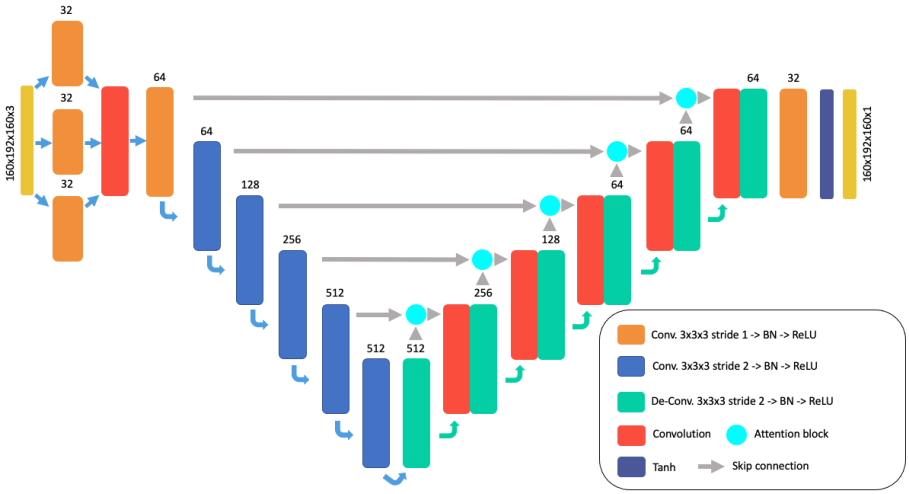


Figure 12: U-Net architecture based on the pix2pix interpretation of TensorFlow. The number above the blocks represent the number of filters in each layer. Compared to model 2, we added attention blocks to focus our attention on the tumor

Appendix D Figure: Model 4

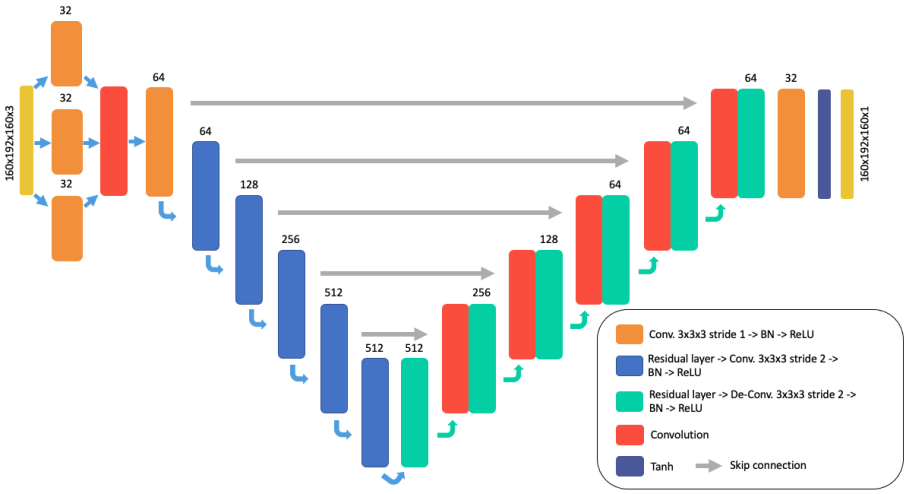


Figure 13: U-Net architecture based on the pix2pix interpretation of TensorFlow. The number above the blocks represent the number of filters in each layer. Compared to model 2, we residual layer to learn the residual between the input and output, rather than the complete transformation.

Appendix E Table: Loss functions

Table 5: Quantitative comparison, with the standard deviation, between the different loss functions with the same baseline model with the evaluation dataset. $\mathcal{L}_2$ and $\mathcal{L}_{SSIM}$ were, based on quantitative metrics, the best performing models.

	$MAE \downarrow$	$MSE \downarrow$	$SSIM \uparrow$	$MAE_B \downarrow$	$MSE_B \downarrow$	$SSIM_B \uparrow$	$MAE_T \downarrow$	$MSE_T \downarrow$	$SSIM_T \uparrow$
$\mathcal{L}_1$	0,047 (0,007)	0,014 (0,003)	0,874 (0,025)	0,121 (0,018)	0,035 (0,008)	0,727 (0,056)	0,165 (0,062)	0,065 (0,045)	0,632 (0,132)
$\mathcal{L}_2$	0,048 (0,007)	**0,014 (0,003)**	0,872 (0,024)	0,121 (0,017)	**0,034 (0,008)**	0,724 (0,056)	0,171 (0,059)	**0,064 (0,042)**	0,624 (0,131)
$\mathcal{L}_{1B}$	0,048 (0,007)	0,014 (0,003)	0,875 (0,024)	0,122 (0,017)	0,036 (0,008)	0,729 (0,056)	0,167 (0,061)	0,066 (0,046)	0,633 (0,131)
$\mathcal{L}_{2B}$	0,048 (0,006)	0,014 (0,003)	0,875 (0,024)	0,122 (0,016)	0,035 (0,008)	0,728 (0,056)	0,174 (0,057)	0,066 (0,041)	0,622 (0,125)
$\mathcal{L}_{1T}$	0,067 (0,005)	0,024 (0,004)	0,822 (0,027)	0,171 (0,013)	0,060 (0,010)	0,618 (0,062)	0,169 (0,063)	0,067 (0,046)	0,610 (0,129)
$\mathcal{L}_{2T}$	0,068 (0,006)	0,024 (0,004)	0,822 (0,025)	0,173 (0,016)	0,061 (0,010)	0,619 (0,058)	0,178 (0,060)	0,066 (0,041)	0,609 (0,124)
$\mathcal{L}_{SSIM}$	**0,047 (0,007)**	0,014 (0,003)	**0,878 (0,024)**	**0,120 (0,018)**	0,036 (0,009)	0,737 (0,055)	**0,162 (0,068)**	0,069 (0,052)	**0,642 (0,135)**

Appendix F Table: Model comparison

Table 6: Quantitative comparison, with the standard deviation, between the different loss functions and models. These metrics are based on 45 evaluation samples.

Model	$MAE \downarrow$	$MSE \downarrow$	$SSIM \uparrow$	$MAE_B \downarrow$	$MSE_B \downarrow$	$SSIM_B \uparrow$	$MAE_T \downarrow$	$MSE_T \downarrow$	$SSIM_T \uparrow$
				$\mathcal{L}_{1B}$					
0	0,048 (0,007)	0,014 (0,003)	**0,875 (0,024)**	0,122 (0,017)	0,036 (0,008)	**0,729 (0,056)**	0,167 (0,061)	0,066 (0,046)	0,633 (0,131)
1	**0,047 (0,007)**	0,014 (0,003)	0,874 (0,025)	**0,120 (0,017)**	0,035 (0,008)	0,728 (0,057)	0,164 (0,065)	0,064 (0,045)	0,631 (0,136)
2	0,048 (0,006)	0,014 (0,003)	0,872 (0,025)	0,121 (0,017)	0,036 (0,008)	0,722 (0,057)	**0,160 (0,070)**	0,068 (0,052)	0,636 (0,137)
3	0,048 (0,006)	0,014 (0,003)	0,869 (0,025)	0,122 (0,016)	0,036 (0,008)	0,717 (0,058)	0,168 (0,063)	0,064 (0,043)	0,622 (0,135)
4	0,047 (0,007)	**0,014 (0,003)**	0,873 (0,025)	0,121 (0,018)	**0,035 (0,008)**	0,724 (0,058)	0,164 (0,063)	**0,063 (0,044)**	**0,639 (0,137)**
				$\mathcal{L}_{2B}$					
0	0,048 (0,006)	0,014 (0,003)	**0,875 (0,024)**	0,122 (0,016)	0,035 (0,008)	**0,728 (0,056)**	0,174 (0,057)	0,066 (0,041)	**0,622 (0,125)**
1	0,048 (0,006)	0,014 (0,003)	0,869 (0,025)	0,124 (0,016)	0,035 (0,008)	0,716 (0,058)	0,173 (0,062)	0,064 (0,043)	0,616 (0,134)
2	0,048 (0,006)	0,013 (0,003)	0,870 (0,024)	0,123 (0,016)	0,034 (0,007)	0,718 (0,057)	0,173 (0,062)	0,064 (0,042)	0,619 (0,132)
3	0,048 (0,007)	0,014 (0,003)	0,871 (0,025)	0,123 (0,019)	0,035 (0,008)	0,720 (0,057)	0,175 (0,061)	0,064 (0,041)	0,617 (0,134)
4	**0,048 (0,007)**	**0,013 (0,003)**	0,872 (0,025)	**0,121 (0,017)**	**0,034 (0,008)**	0,720 (0,058)	**0,172 (0,062)**	**0,064 (0,042)**	0,620 (0,134)
				$\mathcal{L}_{SSIM}$					
0	0,047 (0,007)	0,014 (0,003)	0,878 (0,024)	0,120 (0,018)	0,036 (0,009)	0,737 (0,055)	0,162 (0,068)	0,069 (0,052)	0,642 (0,135)
1	0,048 (0,009)	0,015 (0,004)	0,876 (0,026)	0,122 (0,022)	0,039 (0,011)	0,731 (0,059)	0,165 (0,067)	0,071 (0,053)	0,644 (0,137)
2	**0,046 (0,008)**	**0,014 (0,004)**	**0,880 (0,025)**	**0,118 (0,021)**	**0,035 (0,009)**	**0,740 (0,058)**	**0,161 (0,063)**	**0,064 (0,045)**	0,648 (0,133)
3	0,047 (0,007)	0,014 (0,004)	0,878 (0,025)	0,120 (0,018)	0,036 (0,009)	0,736 (0,057)	0,163 (0,070)	0,070 (0,053)	0,643 (0,138)
4	0,048 (0,008)	0,014 (0,004)	0,879 (0,025)	0,122 (0,022)	0,037 (0,010)	0,739 (0,058)	0,169 (0,062)	0,068 (0,049)	**0,651 (0,132)**

Appendix G Table: Input size comparison

Table 7: Quantitative comparison, with the standard deviation, between the different loss functions and input sizes trained with model 2. These metrics are based on 45 evaluation samples.

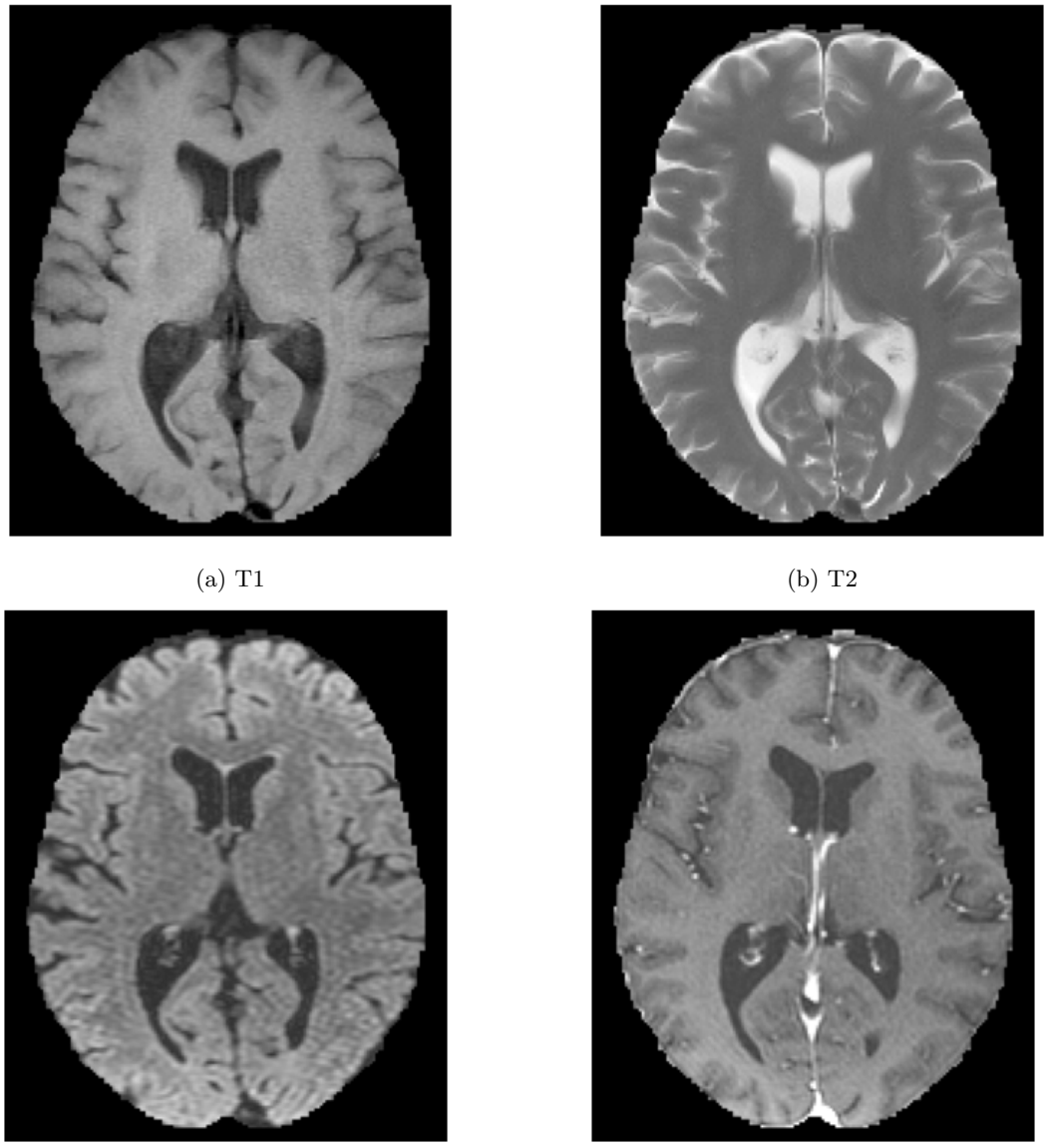
Size	$MAE \downarrow$	$MSE \downarrow$	$SSIM \uparrow$	$MAE_B \downarrow$	$MSE_B \downarrow$	$SSIM_B \uparrow$	$MAE_T \downarrow$	$MSE_T \downarrow$	$SSIM_T \uparrow$
					$\mathcal{L}_{1B}$				
full	**0,047 (0,007)**	**0,014 (0,003)**	**0,874 (0,025)**	**0,120 (0,017)**	**0,035 (0,008)**	**0,728 (0,057)**	0,164 (0,065)	**0,064 (0,045)**	0,631 (0,136)
128	0,077 (0,015)	0,023 (0,008)	0,798 (0,042)	0,123 (0,023)	0,037 (0,012)	0,710 (0,064)	**0,162 (0,068)**	0,067 (0,050)	**0,637 (0,140)**
96	0,097 (0,021)	0,031 (0,012)	0,747 (0,065)	0,133 (0,024)	0,043 (0,016)	0,683 (0,075)	0,163 (0,067)	0,066 (0,048)	0,631 (0,137)
64	0,136 (0,046)	0,050 (0,032)	0,663 (0,125)	0,153 (0,043)	0,056 (0,033)	0,647 (0,109)	0,172 (0,073)	0,071 (0,056)	0,62 (0,147)
					$\mathcal{L}_{2B}$				
full	**0,048 (0,006)**	**0,014 (0,003)**	**0,869 (0,025)**	**0,124 (0,016)**	**0,035 (0,008)**	**0,716 (0,058)**	0,173 (0,062)	0,064 (0,043)	**0,616 (0,134)**
128	0,081 (0,013)	0,023 (0,007)	0,787 (0,043)	0,128 (0,021)	0,037 (0,011)	0,694 (0,065)	0,176 (0,066)	0,064 (0,042)	0,614 (0,135)
96	0,101 (0,021)	0,031 (0,011)	0,741 (0,066)	0,139 (0,024)	0,042 (0,014)	0,674 (0,076)	**0,172 (0,064)**	**0,063 (0,041)**	0,613 (0,136)
64	0,142 (0,045)	0,049 (0,028)	0,656 (0,122)	0,159 (0,041)	0,054 (0,029)	0,639 (0,105)	0,180 (0,064)	0,067 (0,046)	0,611 (0,140)
					$\mathcal{L}_{SSIM}$				
full	**0,048 (0,009)**	**0,015 (0,004)**	**0,876 (0,026)**	**0,122 (0,022)**	**0,039 (0,011)**	**0,731 (0,059)**	**0,165 (0,067)**	**0,071 (0,053)**	**0,644 (0,137)**
128	0,078 (0,016)	0,026 (0,009)	0,803 (0,048)	0,124 (0,025)	0,041 (0,015)	0,716 (0,069)	0,171 (0,073)	0,077 (0,059)	0,644 (0,141)
96	0,096 (0,025)	0,034 (0,016)	0,759 (0,069)	0,132 (0,031)	0,047 (0,021)	0,694 (0,080)	0,169 (0,075)	0,076 (0,059)	0,641 (0,142)
64	0,140 (0,055)	0,061 (0,042)	0,659 (0,128)	0,158 (0,054)	0,068 (0,044)	0,646 (0,114)	0,180 (0,088)	0,086 (0,071)	0,621 (0,153)

Appendix H Table: Mix loss function

Table 8: Quantitative comparison, with the standard deviation, between model 1 trained with either $\mathcal{L}_{1B}$, $\mathcal{L}_{2B}$ $\mathcal{L}_{SSIM}$, and $\mathcal{L}_{M}$. These metrics are based on 45 evaluation samples.

	$MAE \downarrow$	$MSE \downarrow$	$SSIM \uparrow$	$MAE_B \downarrow$	$MSE_B \downarrow$	$SSIM_B \uparrow$	$MAE_T \downarrow$	$MSE_T \downarrow$	$SSIM_T \uparrow$
$\mathcal{L}_{1B}$	0,048 (0,007)	0,014 (0,003)	0,875 (0,024)	0,122 (0,017)	0,036 (0,008)	0,729 (0,056)	0,167 (0,061)	0,066 (0,046)	0,633 (0,131)
$\mathcal{L}_{2B}$	0,048 (0,006)	**0,014 (0,003)**	0,875 (0,024)	0,122 (0,016)	**0,035 (0,008)**	0,728 (0,056)	0,174 (0,057)	**0,066 (0,041)**	0,622 (0,125)
$\mathcal{L}_{SSIM}$	**0,047 (0,007)**	0,014 (0,003)	0,878 (0,024)	**0,120 (0,018)**	0,036 (0,009)	0,737 (0,055)	**0,162 (0,068)**	0,069 (0,052)	**0,642 (0,135)**
$\mathcal{L}_{M}$	0,048 (0,008)	0,014 (0,004)	**0,879 (0,024)**	0,122 (0,019)	0,036 (0,009)	**0,740 (0,054)**	0,172 (0,064)	0,070 (0,046)	0,630 (0,127)

Appendix I Figure: Misalignment figure



(a) T1

(b) T2

(c) T2-FLAIR

(d) cT1

Figure 14: Registration error between the different modalities. In d) is clearly a slight rotation visible compared to a), b), and c). Other differences are the sulci on the left side of the brain and the ventricle at the right posterior side. The model is unable to correct for these errors, decreasing evaluation metrics.