Brain Magnetic Resonance Imaging Generation using Generative Adversarial Networks

Emanuel Alogna
Dipartimento di Elettronica, Informazione e Bioinformatica
Politecnico di Milano
emanuel.alogna@mail.polimi.it

Edoardo Giacomello
Dipartimento di Elettronica, Informazione e Bioinformatica
Politecnico di Milano
edoardo.giacomello@polimi.it

Daniele Loiacono
Dipartimento di Elettronica, Informazione e Bioinformatica
Politecnico di Milano
daniele.loiacono@polimi.it

***Abstract*—Magnetic Resonance Imaging (MRI) is nowadays one of the most common medical imaging technology, due to its non-invasive nature and the many kind of supported sequences (modalities), that provide unique insights about a particular disease. However, it is not always possible to acquire all the sequences required, for several reasons such as prohibitive scan times or allergies to contrast agents. To overcome this problem and thanks to the recent improvements in Deep Learning, in the last few years researchers have been studying the application of Generative Adversarial Networks, a promising paradigm in deep learning, to generate the missing modalities. In this work we developed and trained two models of Generative Adversarial Networks, called MI-pix2pix and MI-GAN, to solve the problem of generating missing modalities for brain MRIs. In particular, our approaches are multi-input generative models, as they exploit as input several MRI modalities to generate the missing one. Our results are promising and show that the developed models are able to generate rather realistic and good quality images.**

I. Introduction

Medical imaging is crucial for clinical analysis and medical intervention since it gives important insights about some diseases whose structures might be hidden by the skin or by the bones. One of the most common imaging technology used nowadays is the Magnetic Resonance Imaging (MRI), ubiquitous in hospitals and medical centers, first because of its non-invasive nature, since, differently from other imaging technologies, doesn't make use of X-ray radiography and secondly because of the recent improvements in the software and hardware instrumentation used. In this type of imaging, various sequences (or modalities) can be acquired and each sequence can give useful and different insights about a particular problem of the patient. For example the T_1-weighted sequence can distinguish between gray and white matter tissues while T_2-weighted is more indicated to highlight fluid from cortical issue.

However, a relevant issue with MRI images is that often not all the modalities are available for each patient. This generally happens for several reasons, such as prohibitive scan times and costs, artifacts, data corruption, wrong machine settings, adverse reactions to contrast dye, etc. Accordingly, a computer assisted generation of the missing MRI modalities from the available ones, is a a problem of great interest. In fact, solving this problem would not only allow doctors to make a more effective diagnostic process, but it could be also very useful for the application of machine learning and deep learning models to MRI images.

978-1-7281-2547-3/20/$31.00 ©2020 IEEE

A promising solution to this problem comes from the field of Deep Learning, and more specifically, from the recent breakthrough of the generative models based on Generative Adversarial Networks (GANs).

In this work, we study and compare two different generative models based on GANs for the generation of missing modalities of brain MRIs. In particular, based on the work of Sharma et al. [1], we wanted to investigate the benefit of multi-input generative models, i.e., models that are able to generate a missing MRI modality from *more than one* available modality in input. To this purpose, we trained two models: MI-GAN, adapted from the approach introduced in [1], and MI-pix2pix that extends the well known approach introduced by Isola et al. [2]. Then, we compared the performance of these two models on the data provided by the 2015 Multimodal Brain Tumor Segmentation Challenge. Our results show that multi-input generative models are indeed a promising approach for the generation of missing modalities in brain MRI. However, differently from what found by Sharma et al. [1], our results suggest that the model based on the pix2pix architecture (MI-pix2pix) is able to achieve better performances than the ones achieve by MI-GAN.

II. Related Work

Generative Adversarial Networks (GANs) were proposed in 2014 by Ian GoodFellow [3] and represent a novel approach for estimating generative models in an adversarial setting. The system is composed by two neural networks: a discriminator D, typically a CNN, and a generator G that are trained simultaneously. In particular, G is trained to learn the probability distribution of the data given as input and, thus, to generate synthetic data that are similar to training data. At the same time, the discriminative model estimates the probability that a sample comes from the training data rather than G. The two networks are trained alternately using backpropagation. During the training process, they compete with each other in a minimax two-player game [4, p. 276]: the discriminator tries to distinguish true images, belonging to the input dataset, from fake images produced by the generator, whose objective is

instead to learn to generate the most realistic data possible to be able to fool D. Later, Radfrod et al. [5] improved GANs to work more effectively with larger dimensionality of data; instead, Mirza and Osindero introduced Conditional Generative Adversarial Networks (cGANs) [6], that extended GANs by conditioning the input of the model with additional data, that could be used to direct the data generation process.

Since the introduction of GANs and, in particular, of cGANs, many researchers started to apply this kind of architecture to the *Image-to-Image translation*, the task of translating one possible image representation into another one, using images as auxiliary information to control the data generation. A notable contribution is the work of Phillip Isola et al. [2] that developed a general purpose architecture, known as Pix2Pix, that employs a generator network based on *U-Net* [7] and a *PatchGAN* [8] classifier, as discriminator.

So far, several studies applied GANs to medical image synthesis. In particular, cross-modality image synthesis – i.e., the conversion of an input image from a source modality to a target modality – is one of the main application of this architecture to medical imaging. A survey published in 2018 [9] collects the major contributions in this field through the application of GANs. The authors describe MRI as the most common imaging modality explored in the literature related to this kind of generative approaches, probably due to the fact that cross-modality synthesis through GANs could reduce the amount of time requested from multiple-modality MRI acquisition.

Many different approaches and datasets have been used in the literature in order to overcome the problem of missing modalities. Since most datasets only contain T_1 or T_1/T_2/PD scans due to practical reasons, Orbes-Arteaga et al., in [10], implemented a GAN that generates Brain T_{2flair} using the T_1 modality. Camileri et al. [11] developed a variant of the original GAN, called Laplacian pyramid framework (LAP-GAN) that synthesizes images in a coarse-to-fine process by introducing progressive refinements. This method, as the name suggests, is based on Laplacian pyramid and allows to initially generate an image with low resolution and then, incrementally refining it by adding finer details. Another approach to generate missing modalities was proposed in [12] where the authors presented two possible scenarios, based on the given dataset: (i) when the multi-contrast images are spatially registered they use a model called pGAN, which incorporates a pixel-wise loss into the objective function, while they adopt (ii) a cycleGAN [13] in the more realistic scenario in which pixels are not aligned between modalities. Anmol Sharma and Ghassan Hamarneh [1] proposed a *many to many* generative model, capable of synthesizing multiple missing sequences given a combination of various input sequences. Furthermore, they also apply the concept of curriculum learning, based on the variation of the difficulty of the examples that are used during the network training.

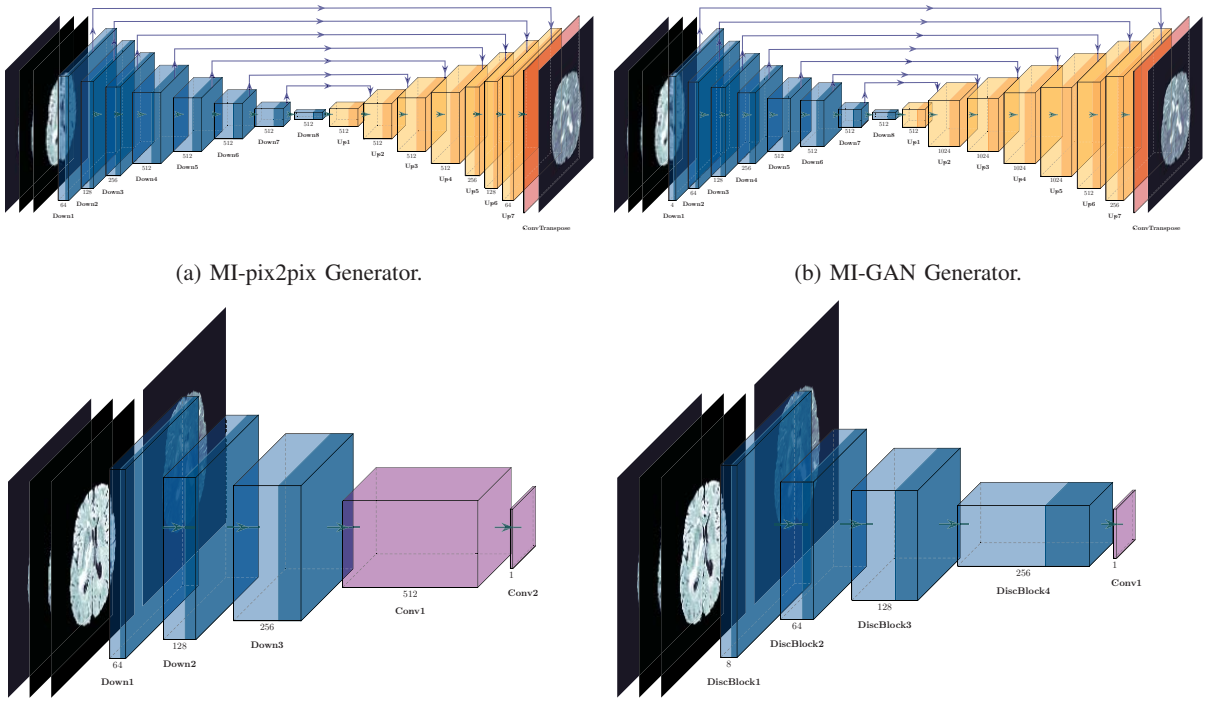
Finally, GANs have been successfully applied also to different images, such as PET, CT, and MRA images. Notable examples include the work of Olut et al. [14], that proved that GANs work efficiently even when the source imaging technique is different from the target one by synthesizing MRA brain images from T_1 and T_2 MRI modalities. Then, Ben-Cohen et al. [15] successfully generated PET images using CT scans through a fully connected neural network, whose output is improved and refined by cGANs.

III. Modality Generation with Adversarial Networks

In this work, we studied the problem of generating a target brain MRI modality among the four ones more commonly acquired: T_1-weighted (T_1), T_1-contrast-enhanced (T_{1c}), T_2-weighted (T_2), and T_2-fluid-attenuant inversion recovery (FLAIR). More specifically, we assumed that a missing modality has to be generated from the other three available ones. To this purpose, based on the findings in [1], we focused on multi-input generative models, i.e., models that receive as input multiple images (the available modalities) and generate as output a single image (the missing modality). In particular, we designed two generative models based on GANs: a multi-input version of pix2pix [2], dubbed *MI-pix2pix*, and a modified version of the MM-GAN introduced by Sharma et al. in [1], dubbed *MI-GAN*. In this section, we provide the details of these two generative models.

A. MI-pix2pix

In this GAN, the generator is based on U-net [7] that takes as input a tensor with size 32x256x256x3, where the first dimension indicates the batch size and the last one is the number of modalities in input. For example, when MI-pix2pix is used to generate the missing T_2 modality, the other three modalities, T_1, T_{1c}, T_{flair}, are provided as inputs respectively as the first, second and third channel. The symmetrical downsampling and upsampling blocks, typical of U-Net architecture, employ skip connections to concatenate the inputs of each downsampling block to the input of the corresponding upsampling block. The downsampling blocks are composed by three layers: $C_{n,k=4,s=2}, BatchNorm, LeakyReLU$, where $C_{n,k,s}$ is a convolutional layer with n filters, kernel size k and stride s. The sequence of downsampling blocks, with $n = \{64, 128, 256, 512, 512, 512, 512, 512\}$, reduces the spatial information while increases the feature dimension, until the last downsampling block that has output shape of 1x1x512. Instead, the upsampling blocks are composed by four layers: $D_{n,k=4,s=2}, BatchNorm, Dropout, ReLU$, where $D_{n,k,s}$ is a transposed convolution layer with n filters, kernel size k and stride s. The sequence of upsampling blocks, with $n = \{512, 512, 512, 512, 512, 256, 128, 64\}$, is followed by a last transposed convolutional layer and a *Tanh* activation. The output of the network is a batch of images with dimension 256x256x1. The discriminator is a 70x70 PatchGAN with two inputs: (i) the target image (fake or real) of shape 32x256x256x1, (ii) the concatenation of the three modalities (of shape 32x256x256x3) provided as input to the Generator. This network has only three downsampling blocks (with $n = \{64, 128, 256\}$), followed by these layers: $ZeroPadding$,



(a) MI-pix2pix Generator.

(b) MI-GAN Generator.

(c) MI-pix2pix Discriminator.

(d) MI-GAN Discriminator.

Fig. 1: Generator (top row) and Discriminator (bottom row) architectures of the two multi-input GANs studied in this work.

$C_{n=512,k=4,s=1}$, $BatchNorm$, $LeakyReLU$, $ZeroPadding$, $C_{n=1,k=4,s=1}$.

B. MI-GAN

This is based on the MM-GAN introduced in [1]: the only variation we applied was the replacement of every layer of Instance Normalization with a Batch Normalization, which is also used in MI-pix2pix, as we observed that normalizing the activations of each channel across the whole batch was more effective, in terms of quality in the generated samples, than computing the mean/standard deviation and normalizing across each channel for each training image.

The MI-GAN generator is a modified U-Net with concatenated skip connections and the typical U-shape architecture (Figure 1b), The building blocks for the encoding path are defined as: $C_{n,k=4,2}, BatchNorm, LeakyReLU, DropOut_{0.5}$. The *upsample block* has the same layers seen in the pix2pix architecture: $D_{n,k=4,s=2}, BatchNorm, ReLU, Dropout_{0.5}$ The last layer of the generator is a transposed convolution followed by a *Tanh* activation function. The output of the network is a batch of images with dimension 256x256x1. The discriminator takes two inputs X_t and X_i and produces one output D(X_t, X_i), each one of these with 4 channels - i.e. one per modality. Assuming the discrimination of a T_1' synthesized scan, the inputs were X_t: $\{T_1, T_2, T_{1c}, T_{2flair}\}$ and X_i: $\{T_1', T_2, T_{1c}, T_{2flair}\}$.

C. Loss Function

The generator loss and discriminator loss used with MI-GAN (III-B) are the ones proposed in [2] and defined as follows:

$$\begin{aligned} L_G &\leftarrow \lambda \mathcal{L}_1(G(x), y) + (1-\lambda)\mathcal{L}_2(D(x, G(x)), L_{ar}) \\ L_D &\leftarrow \mathcal{L}_2(D(x, y), L_{ar}) + \mathcal{L}_2(D(x, G(x)), L_r) \end{aligned} \tag{1}$$

where the input x is a concatenation of three sequences, while $G(x)$ represents the prediction generated by the GAN. L_{ar} is a tensor of unitary values that is used to encourage the generator to produce samples that the discriminator evaluates as real. $\mathcal{L}_2$ is the L2 norm - or *Mean Squared Error* -, while $\mathcal{L}_1$ denotes the L1 norm - or *Mean Absolute Error* that is useful to generate samples that are structurally similar to the target image. This term was chosen as reconstruction loss term because of its ability to prevent too much blurring, as compared to using a L2 loss [1]. L_r is a tensor with its entries equal to zero and it's use to encourage D to assign low values to the generated samples. In the same way, L_{ar} is used to induce the discriminator in assigning values close to 1 to the true samples.

Following the same notation, the loss of MI-pix2pix is instead defined as:

$$\begin{aligned} L_G &\leftarrow \lambda \mathcal{L}_1(G(x), y) + BCE(L_{ar}, D(x, G(x))) \\ L_D &\leftarrow BCE(D(x, y), L_{ar}) + BCE(D(x, G(x)), L_r) \end{aligned} \tag{2}$$

where $BCE(x, y)$ is the *Binary Cross-Entropy* as commonly implemented in most deep learning frameworks. The L_D of MI-pix2pix is computed as the sum between the binary crossentropy of $(D(x, y), L_{ar})$ and the one of $D((x, y'), L_r)$ where y' is the generated image. L_G, on the other hand, contains the reconstruction term between y and y', multiplied by an hyper parameter lambda and summed to the binary crossentropy of $D((x, y'), L_r)$.

IV. Experimental Design

A. Dataset

In this work, to test our approach we used data from the Multimodal Brain Tumor Segmentation Challenge 2015 (known as BraTS2015) [16], [17], that provides a rather large dataset of Brain MRIs of patients affected by Glioma. In particular, BraTS2015 includes 220 samples of high grade subjects (HG) and 54 samples of low grade subjects (LG). For each sample, 5 volumes are provided: 4 of these contain the different MRI sequences (T_1, T_{1c}, T_2, and T_{2flair}), while the last volume corresponds to the segmented area of the tumor. We split the dataset in three different sets: training (80%), validation (10%) and test (10%) resulting in 219 patients assigned to the first set, 27 to the validation one and 28 to the test. Since we believe it is important to maintain the balance between HG and LG subjects during training and evaluation phases, we split the dataset applying stratified sampling [18].

B. Preprocessing

All the images in the dataset were first *center cropped*: the outer parts of each volume were removed while the central region was retained along each dimension. We also discarded the external slices, that contained almost no useful information, reducing the number of slices from 155 to 128. As a result, the final shape of each volume is 180x180x128. Then, in order to provide the models data that has the same dynamic range of values, we applied to each volume a min-max normalization. Finally, as our GANs architecture only allows input images with dimensions that are power of 2, we padded our 180x180 images in order to obtain 256x256 images.

C. Evaluation Metrics

To evaluate the output of our generative models, we considered different metrics that could be used to assess three different objectives: (i) quality of the whole image, (ii) quality of the tumor area, and (iii) discriminative power of the generated image.

Image Metrics. The first metrics we considered aim at assessing the quality of the whole images generated. To compute these metrics we first had to crop and normalize the images as follows. We center-cropped the generated images to 155x194, which is the size of the largest bounding box to contain each brain in the BraTS2015 dataset [1]. Then, we applied *mean normalization* [19] to each image, either generated or real. Thus, we computed the following three metrics: (i) the mean squared error (MSE), (ii) the *Peak Signal-to-Noise Ratio* (PSNR) [20], and (iii) the *Structural Similarity* (SSIM) [21]:

$$MSE(\hat{I}, I) = \frac{1}{W \cdot H} \sum_{i=1}^{W} \sum_{j=1}^{H} (\hat{I}_{i,j} - I_{i,j})^2,$$

$$PSNR(\hat{I}, I) = 10 \log_{10} \frac{(\max(\hat{I}_{i,j}))^2}{MSE(\hat{I}, I)},$$

$$SSIM(\hat{I}, I) = \frac{(2\mu(I)) + \mu(\hat{I}) + c_1)(2\sigma(\hat{I}, I) + c_2)}{(\mu^2(I) + \mu^2(\hat{I}) + c_1)(\sigma^2(I) + \sigma^2(\hat{I}) + c_2)},$$

where W and H are the width and height of the images; I and $\hat{I}$ are respectively the real and generated image; $I_{i,j}$ and $\hat{I}_{i,j}$ are the pixel values of respectively the real and generated images; μ is the average of pixel values, σ^2 is the variance of pixel values, and $\sigma(\hat{I}, I)$ is the covariance of $\hat{I}$ and I pixel values. Both MSE and SSIM have values between 0 and 1. Lower values of MSE mean a better quality of the generated image. Instead, the greater the SNR and SSIM are, the better is the quality.

Tumor Metrics. In the dataset considered in this work, the tumor is the most relevant area of the images. Thus, we computed the MSE and PSNR metrics, described above, by restricting the computation only to the pixels that are inside the tumor area of the image.

Discriminative Metrics. In addition to the metrics described so far, to assess the amount of information contained in the generated images, we compared the performance achieved by a tumor segmentation model when generated images are used as input instead of real ones. More specifically, we computed the *Dice Similarity Coefficient* of the segmentations obtained from a model based on GANs, introduced in [22].

V. Results

In this work, we compared MI-GAN and MI-pix2pix on the generation of each one of the four MRIs modality available in the BRATS2015 dataset: T_1, T_{1c}, T_2, and T_{2flair}. We also compared the performance of MI-GAN and MI-pix2pix with some single-input pix2pix generative models and computed, as a performance baseline, the metrics of the images provided as input to the generative models – i.e., we evaluated how similar the input images are to the expected output of the generative models.

To train the pix2pix, MI-pix2pix and MI-GAN, we used the following parameters settings. The learning rate α was set to 0.0002, the exponential decay rate for the 1st moment estimates β_1 was set to 0.5, and the one for the 2nd moment estimates (β_2) was set to 0.999. The value of λ was set to 100 in the generator loss of pix2pix and MI-pix2pix, while it was set to 0.9 in the loss of the MI-GAN generator. The MI-GAN discriminator loss was multiplied by 0.5, to slow down the rate at which D learns compared to G. The convolutional layer weights have been initialized using a normal distribution with 0 mean and standard deviation equal to 0.05. Training

was performed for a number of epochs that range from 25 to 70, until the convergence of each model. We used [23] as reference for the implementation of the pix2pix model that, as MI-pix2pix and MI-GAN, was re-implemented in Tensorflow 2.1.0. We ran our experiments using Google Colaboratory, on a Nvidia Tesla P100-PCIE-16GB GPU with 26 GB of RAM available and an Intel(R) Xeon(R) CPU @ 2.30GHz.

In the results reported in this section, we used the following notation. The pix2pix models have been dubbed as P2P, and MI-pix2pix as MI-P2P. In addition for pix2pix and for the baseline performance, we reported between parenthesis the source modality used.

Generation of T_1 images. We trained four models to generate the missing modality T_1: two single-input pix2pix with different inputs (T_2 and T_{1c}, MI-pix2pix, and MI-GAN. We choose to train different pix2pix models to understand how the performances change when using an input, T_{1c}, that is similar to the target compared to using T_2. The results are reported in Table I.

Generation of T_2 images. Also for the generation of T_2 we trained four models: the first one is a pix2pix trained to receive T_1 as input, the second one takes as input the T_{2flair} modality, that captures more similar characteristics (of T_2) than T_1; the other models are MI-pix2pix and MI-GAN. The results are reported in Table II.

Generation of T_{1c} images. Table III shows instead the performances obtained by three models trained to generate T_{1c} slices: one pix2pix model and two multi-input models. We choose T_1 as input modality for the single-input GAN because it is the most similar sequence, among the ones available, to the target T_{1c}.

Generation of T_{2FLAIR} images. Tables IV and V summarize the performances reached by the models trained to generate T_{2flair}: in particular V shows the additional metric we implemented to evaluate the quality of the generated tumor area using a single-input segmentation model from [22]. The DSC is calculated between the ground truth and the segmentation of our synthesized images. Furthermore, a reference DSC score is computed between the ground truth and the segmentation of the original slice from BRATS2015. We trained only one pix2pix model, using as input T_2 that is the most similar sequence, among the ones available, to the target.

TABLE V: DSC performances, on the test set, obtained by comparing the ground truth and the segmentations, using segmentation model in [22], of T_{2flair} predictions.

Segmentation image	$\mathbf{DSC_{tumor}}$
Original T_{2flair}	0.8053 ± 0.1156
P2P(T_2)	0.6632 ± 0.1608
MI-P2P	$\mathbf{0.7427 \pm 0.1810}$
MI-GAN	0.6837 ± 0.2136

VI. Discussion

In this section, we discuss the result reported in the previous section. In the first part of the discussion, we discuss the quantitative results reported in Section V. In the last part, instead, we will provide a brief qualitative analysis of the generated images.

A. Quantitative Analysis

Pix2pix vs Baselines. Our results (see Table I, Table II, Table III, and Table IV) show that, as expected, pix2pix is indeed able to generate images that are more similar to the images of the target modality with respect to the ones received as input. In fact, all the metrics computed on the baselines are worse than the corresponding ones computed on the images generated with pix2pix generative models. As an example, when pix2pix is trained to generate T_{2flair} images from T_2 images (see P2P(T_2) in Table IV), it is able to achieve improvements over the baseline (i.e., T_2 images) of 35%, of 38%, and of 560% respectively in PSNR, SSIM, and MSE. These results suggest that the processing performed by the generative models is actually able to *translate* the input into images that are more similar to the target ones.

Single-Input vs Multi-Input Models. Our results show that both multi-input models generally outperform single-input ones. However, our results show at least two cases where a single-input model is able to achieve better performances than a multi-input one. In particular, the results in Table I show that, when trained to generate T_{1c} images from T_1, single-input pix2pix is able to outperform MI-pix2pix (but is outperformed by MI-GAN). Instead, Table III shows that the single-input pix2pix trained to generate T_{1c} from T_1 images outperforms both MI-pix2pix and MI-GAN on this task. These results are not very surprising as T_1 and T_{1c} are very similar modalities and the information content of T_2 and T_{2flair} modalities do not provide a clear advantage to generate the target images. Therefore, in this kind of tasks, a simpler single-input model can achieve better performance than a more complex one. These results suggest that the choice of single-input models over multi-input ones should be based on a careful analysis of the considered task.

MI-pix2pix vs MI-GAN. Our results show that MI-pix2pix performs slightly better than MI-GAN for the generation of all the modalities except for T_1, where MI-GAN achieves better performance. The results also suggest that the improved performances of MI-GAN with respect to pix2pix discussed in [1] are not due to the novel architecture and loss function of MI-GAN but rather to the benefit of using a multiple modalities as input. In fact, as our results show, when pix2pix is extended to use multiple modalities as input (MI-pix2pix), it generally outperforms MI-GAN.

Whole image vs tumor area. The comparison of the metrics computed on the whole image to the ones computed on the tumor area show that none of the trained models seems to perform much better (or much worse) than the others on the specific tumor area. On the other hand, a direct comparison of the values computed on the tumor area with the values computed on the whole image is not possible: while for the

TABLE I: Generation of T_1: performances on the test set. Best results are reported in bold.

Model	MSE	PSNR	SSIM	$\mathbf{MSE_{tumor}}$	$\mathbf{PSNR_{tumor}}$
Baseline(T_2)	0.0396 ± 0.0275	15.3286 ± 4.2134	0.5054 ± 0.2116	0.0594 ± 0.0523	13.6678 ± 3.6085
P2P(T_2)	0.0060 ± 0.0046	23.4967 ± 3.6754	0.8112 ± 0.1004	0.0199 ± 0.0187	18.5047 ± 3.7607
Baseline(T_{1c})	0.0058 ± 0.0050	23.8431 ± 4.0912	0.8096 ± 0.0984	0.0173 ± 0.0216	20.1544 ± 5.0543
P2P(T_{1c})	0.0044 ± 0.0041	25.0680 ± 3.8652	0.8403 ± 0.0856	0.0114 ± 0.0143	21.4485 ± 4.3380
MI−P2P	0.0044 ± 0.0040	24.9339 ± 3.6983	0.8413 ± 0.0838	0.0113 ± 0.0099	20.8938 ± 3.6111
MI−GAN	$\mathbf{0.0041 \pm 0.0038}$	$\mathbf{25.2569 \pm 3.6512}$	$\mathbf{0.8472 \pm 0.0830}$	$\mathbf{0.0102 \pm 0.0097}$	$\mathbf{21.5359 \pm 3.8620}$

TABLE II: Generation of T_2: performances on the test set. Best results are reported in bold.

Model	MSE	PSNR	SSIM	$\mathbf{MSE_{tumor}}$	$\mathbf{PSNR_{tumor}}$
Baseline(T_1)	0.0396 ± 0.0275	15.3286 ± 4.2134	0.5054 ± 0.2116	0.0594 ± 0.0523	13.6678 ± 3.6085
P2P(T_1)	0.0100 ± 0.0074	21.3182 ± 3.8023	0.7521 ± 0.1247	0.0476 ± 0.0397	14.3652 ± 3.3523
Baseline(T_{2flair})	0.0275 ± 0.0199	16.8268 ± 3.9727	0.6262 ± 0.1597	0.0464 ± 0.0500	15.1591 ± 4.0428
P2P(T_{2flair})	0.0087 ± 0.0076	21.9227 ± 3.7021	0.7567 ± 0.1287	0.0256 ± 0.0242	17.3035 ± 3.4584
MI−P2P	$\mathbf{0.0073 \pm 0.0063}$	$\mathbf{22.7645 \pm 3.8272}$	$\mathbf{0.8005 \pm 0.1112}$	0.0207 ± 0.0167	$\mathbf{18.1305 \pm 3.5930}$
MI−GAN	0.0077 ± 0.0061	22.3719 ± 3.5290	0.7835 ± 0.1141	$\mathbf{0.0205 \pm 0.0167}$	18.0725 ± 3.3763

TABLE III: Generation of T_{1c}: performances on the test set. Best results are reported in bold.

Model	MSE	PSNR	SSIM	$\mathbf{MSE_{tumor}}$	$\mathbf{PSNR_{tumor}}$
Baseline(T_1)	0.0058 ± 0.0050	23.8431 ± 4.0912	0.8096 ± 0.0984	0.0173 ± 0.0216	20.1544 ± 5.0543
P2P(T_1)	$\mathbf{0.0051 \pm 0.0048}$	$\mathbf{24.6165 \pm 4.0755}$	$\mathbf{0.8139 \pm 0.0996}$	$\mathbf{0.0155 \pm 0.0199}$	$\mathbf{20.6981 \pm 4.9762}$
MI−P2P	0.0052 ± 0.0040	24.1597 ± 3.8631	0.8110 ± 0.0963	0.0168 ± 0.0172	19.8441 ± 4.6258
MI−GAN	0.0054 ± 0.0040	23.9242 ± 3.6958	0.8027 ± 0.1003	0.0157 ± 0.0162	19.9779 ± 4.3568

TABLE IV: Generation of T_{2flair}: performances on the test set. Best results are reported in bold.

Model	MSE	PSNR	SSIM	$\mathbf{MSE_{tumor}}$	$\mathbf{PSNR_{tumor}}$
Baseline(T_2)	0.0275 ± 0.0199	16.8268 ± 3.9727	0.6262 ± 0.1597	0.0464 ± 0.0500	15.1591 ± 4.0428
P2P (T_2)	0.0090 ± 0.0065	21.5895 ± 3.4831	0.7518 ± 0.1211	0.0390 ± 0.0463	15.9946 ± 4.0459
MI−P2P	$\mathbf{0.0069 \pm 0.0049}$	$\mathbf{22.8165 \pm 3.7317}$	$\mathbf{0.7772 \pm 0.1094}$	$\mathbf{0.0221 \pm 0.0375}$	$\mathbf{19.0374 \pm 4.1582}$
MI−GAN	0.0072 ± 0.0050	22.5524 ± 3.5655	0.7610 ± 0.1175	0.0258 ± 0.0285	17.4694 ± 3.6137

tumor area the computation involves only relevant pixels, for the whole image it involves also pixel with no information (black pixels) as we don't dispose of a precise segmentation of the whole brain area. Therefore, further investigations will be necessary to get a better insight on this issue.

Discriminative Metrics. Finally, we also tried to assess the discriminative power of the generated images when used to take some decisions, more specifically, when used as input to a segmentation model. We performed this analysis only on T_{2flair} images, as they are the most effective ones in segmentation tasks [22]. The results in Table V show that the images generated with MI-pix2pix are the ones that allow to achieve the better segmentations, reaching an average performance (the DSC score) that is rather good with respect to the ones reached using the real images.

B. Qualitative Analysis

The qualitative analysis of the generated images allows to confirm the findings discussed above. Figure 2 shows two examples of T_2 images generated by pix2pix using either T_1 or T_{2flair} as input. A qualitative analysis shows how using as input a modality that is more similar to the target one, i.e., T_{2flair} instead of T_1, allows to generate much better quality images, especially in the tumor area.

On the other hand, a qualitative analysis of this kind does not always allow to appreciate the differences among the models. As an example, Figure 3 shows some T_2 images generated by different models.

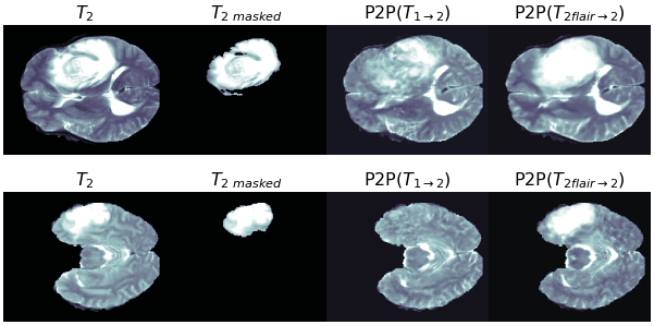


Fig. 2: Comparison of P2P($T_{1\to2}$) (i.e., pix2pix trained to generate T_2 from T_1) and P2P($T_{2flair\to2}$) (i.e., pix2pix trained to generate T_2 from T_1). $T_{2masked}$ is the tumor area.

Finally, Figure 4 shows some examples of segmentation obtained using either real and generated images. As expected, in general the segmentation obtained from generated images are not as good as the ones obtained from real images. However, sometimes the segmentation obtained from generated images are very similar (e.g., fourth row in Figure 4) or even better (e.g., second row in Figure 4) than the ones obtained from the real images. This might be also due to the fact that generated images include information from other modalities that might be not present in the real image.

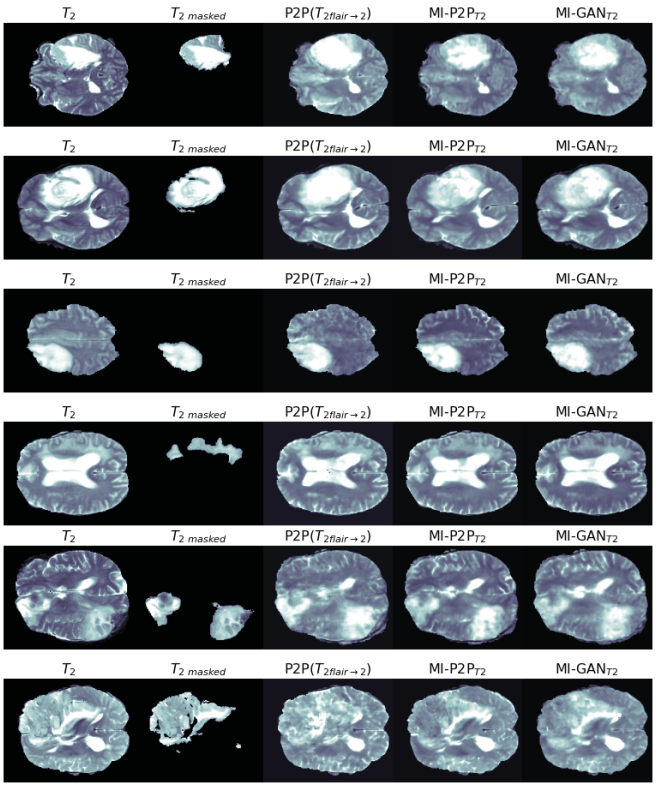


Fig. 3: Some examples of real and generated T_2 images. P2P($T_{2flair\rightarrow 2}$) is a single-input pix2pix model that uses T_{2flair} images as input, MI-P2P$_{T_2}$ and MI-GAN$_{T_2}$ are multi-input models trained to generate T_2 images. Each row corresponds to a different subject from the test set.

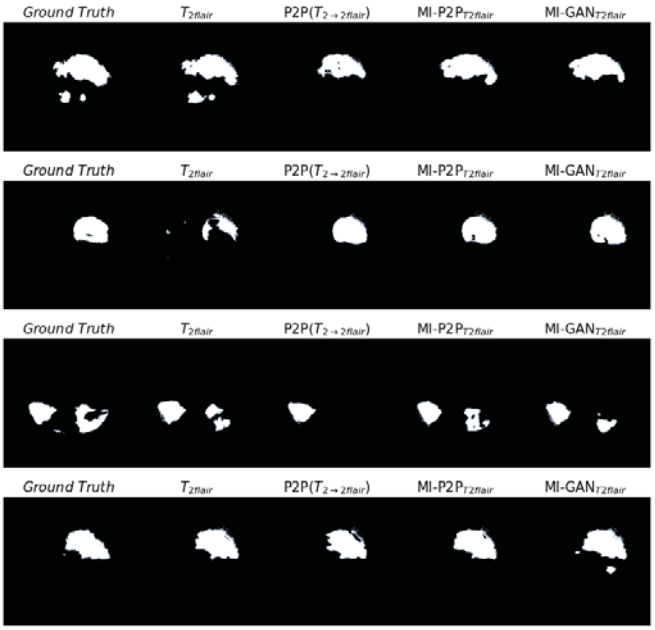


Fig. 4: From left to right: ground truth and the segmentations from T_{2flair}, pix2pix, MI-pix2pix and MI-GAN. Each row corresponds to a different subject from the test set.

VII. Conclusions

In this work we developed and compared different generative models based on GANs to synthesize missing MRI modalities. In particular we studied two multi-input generative models. The first, MI-pix2pix, is a multi-input extension of the well known pix2pix approach [2] for image-to-image translation problems. The second, MI-GAN, was adapted from the approach introduced in [1]. These two models were also compared to a single-input pix2pix model to better assess the benefits of using a multi-input approach. We trained these models to generate missing moadalities for brain MRIs, using the BraTS2015 dataset. We designed a set of quantitative metrics to assess the performance of the different approaches and performed also a qualitative analysis. Our results show that generated images are rather accurate and realistic, compared to real images available in the dataset, such that in some cases it might be difficult to distinguish between real and generate images. We also showed that generated images could be used to solve segmentation task reaching a rather good performance compared to using real images. Finally, our findings suggest that (i) multi-input models performs generally better than single-input ones with a few exceptions and (ii) MI-pix2pix model allows to achieve, in general, better results than MI-GAN. Future works will include a better qualitative assessment of the generated images, involving domain experts, and further investigations of the model performances on the tumor areas.

References

[1] A. Sharma and G. Hamarneh, "Missing mri pulse sequence synthesis using multi-modal generative adversarial network," 2019. [Online]. Available: https://arxiv.org/abs/1904.12200

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[4] C. Floudas and P. Pardalos, *Encyclopedia of Optimization*, ser. Encyclopedia of Optimization. Kluwer Academic, 2001, no. v. 1. [Online]. Available: https://books.google.com.mx/books?id=gtoTkL7heS0C

[5] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015. [Online]. Available: http://arxiv.org/abs/1511.06434

[6] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[8] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," 2016. [Online]. Available: https://arxiv.org/abs/1604.04382

[9] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, Dec 2019. [Online]. Available: http://dx.doi.org/10.1016/j.media.2019.101552

[10] M. Orbes-Arteaga, M. J. Cardoso, L. Srensen *et al.*, "Simultaneous synthesis of flair and segmentation of white matter hypointensities from t1 mris," 2018. [Online]. Available: https://arxiv.org/abs/1808.06519
[11] F. Calimeri, A. Marzullo, C. Stamile, and G. Terracina, "Biomedical data augmentation using generative adversarial neural networks," in *Artificial Neural Networks and Machine Learning – ICANN 2017*, 10 2017, pp. 626–634.
[12] S. U. H. Dar, M. Yurt, L. Karacan *et al.*, "Image synthesis in multi-contrast mri with conditional generative adversarial networks," 2018. [Online]. Available: https://arxiv.org/abs/1802.01221
[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2017. [Online]. Available: https://arxiv.org/abs/1703.10593
[14] S. Olut, Y. H. Sahin, U. Demir, and G. Unal, "Generative adversarial training for mra image synthesis using multi-contrast mri," 2018. [Online]. Available: https://arxiv.org/abs/1804.04366
[15] A. Ben-Cohen, E. Klang, S. P. Raskin *et al.*, "Cross-modality synthesis from ct to pet using fcn and gan networks for improved automated lesion detection," 2018. [Online]. Available: https://arxiv.org/abs/1802.07846
[16] B. Menze, A. Jakab, S. Bauer *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, p. 33, 2014. [Online]. Available: https://hal.inria.fr/hal-00935640
[17] M. Kistler, S. Bonaretti, M. Pfahrer *et al.*, "The virtual skeleton database: An open access repository for biomedical research and collaboration," *J Med Internet Res*, vol. 15, no. 11, p. e245, Nov 2013. [Online]. Available: http://www.jmir.org/2013/11/e245/
[18] Y. Xue, T. Xu, H. Zhang *et al.*, "Segan: Adversarial network with multi-scale l1 loss for medical image segmentation," *Neuroinformatics*, vol. 16, no. 3-4, p. 383392, May 2018. [Online]. Available: http://dx.doi.org/10.1007/s12021-018-9377-x
[19] A. Ng, "Gradient descent in practice I - feature scaling," Coursera. [Online]. Available: https://www.coursera.org/learn/machine-learning/lecture/xx3Da/gradient-descent-in-practice-i-feature-scaling
[20] "Psnr," Mathworks. [Online]. Available: https://it.mathworks.com/help/vision/ref/psnr.html
[21] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, pp. 600 – 612, 05 2004. [Online]. Available: https://ieeexplore.ieee.org/document/1284395
[22] E. Giacomello, D. Loiacono, and L. Mainardi, "Brain mri tumor segmentation with adversarial networks," 2019. [Online]. Available: https://arxiv.org/abs/1910.02717
[23] "Pix2pix." [Online]. Available: https://www.tensorflow.org/tutorials/generative/pix2pix