Synthetic Cross-sequence Generation of MRI Images using Deep Learning Networks

Jacobo Romero Díaz, jromerdi7@alumnes.ub.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisors: Aida Niñerola Baizán, ninerola@clinic.cat, Arnau Farré Melero, afarrem@clinic.cat

Abstract: Magnetic Resonance Imaging (MRI) is a non-invasive diagnostic imaging modality that employs multiple acquisition sequences to generate complementary tissue contrasts. However, in clinical practice, not all sequences are available due to time, cost, or patient-related issues. In this work, we investigate deep learning–based supervised image-to-image translation for synthetic cross-sequence MRI generation using a dataset of 148 patients. T2, FLAIR, and contrast-enhanced T1 (T1GD) images are synthesized from T1 inputs using U-Net–based encoder–decoder models and conditional GANs (pix2pix), under full-resolution and patch-based training strategies. Evaluation using Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) shows that U-Net models with tailored loss functions achieve results comparable to state-of-the-art approaches, outperforming pix2pix. Finally, qualitative evaluation revealed that conventional metrics fail to capture sequence-specific localized regions, highlighting the need for task-aware evaluation criteria.
Keywords: Image-to-Image Translation, Magnetic Resonance Imaging, Deep Learning, Medical Physics
SDGs: Good health and wellbeing, Industry, innovation and Infrastructure

I. INTRODUCTION

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging technique that provides high-resolution images of soft tissues without ionizing radiation [1]. MRI contrast is based on the magnetic properties of hydrogen nuclei, whose nuclear spins align with a strong external magnetic field and are perturbed by radiofrequency pulses. Their return to equilibrium depends on tissue-specific nuclear magnetic relaxation processes, characterized by longitudinal (T1) and transverse (T2) relaxation times. Variations in molecular composition and environments across tissues lead to differences in T1 and T2 values, and therefore differences in contrast across MRI acquisition sequences. It is one of the most versatile imaging modalities in medicine, providing a wide spectrum of tissue contrasts in different acquisition sequences. MRI can target tissue relaxation properties by adjusting the radiofrequency excitation and acquisition, producing sequences such as T1-weighted images that highlight general structures such as gray matter and white matter, T2-weighted images that are used for locating tumors, and Fluid-Attenuated Inversion Recovery (FLAIR) images that eliminate fluid signals to improve the location of lesions [1]. Therefore, acquiring different sequences is needed for accurate disease diagnosis, treatment and monitoring.

Nevertheless, acquiring all MRI sequences is often affected by limited scanning time, cost, or patient motion, leading to incomplete or distorted studies [2]. In addition, some sequences require the administration of contrast agents such as gadolinium (T1GD), used to highlight blood–brain barrier disruption, inflammation or tumors [3]. Besides increasing scan time and cost, contrast agent administration may be contraindicated in some patients due to allergic reactions or nephrotoxicity [3]. Contrast-enhanced sequences are then often omitted, which leads to incomplete MRI acquisitions.

In this context, synthesizing missing or noisy MRI sequences from available ones, commonly referred to as cross-sequence or cross-modality [2], has gained increasing attention with the introduction of deep learning. Broadly, the task of learning a mapping between aligned source and target images is known as supervised image-to-image (I2I) translation [4]. Encoder–decoder architectures initially dominated this paradigm, with U-Net [**?**] emerging as the gold standard in medical imaging for capturing image context and fine details.

More recently, Generative Adversarial Networks (GANs) have been introduced to improve visual realism. The adversarial loss favors generated images to resemble real ones, reducing blurring effects sometimes observed with pixel-wise objectives [5]. Conditional GANs (cGANs) extend this framework by conditioning the generation on an input image, making them useful for I2I translation tasks [5].

Despite recent advances, MRI I2I translation remains challenging due to the need for high anatomical and structural accuracy, as synthesis errors may directly impact clinical interpretation. In this work, we focus on deep learning–based architectures for cross-sequence MRI synthesis that aim to reliably translate contrast information while preserving anatomical realism.

II. METHODOLOGY

A. Data

The private *Hospital-Clinic* dataset used in this work consists of MRI scans from 148 patients, each acquired with a varying subset of imaging sequences. The modalities include T1 (n=139), T1GD (n=143), T2 (n=132),

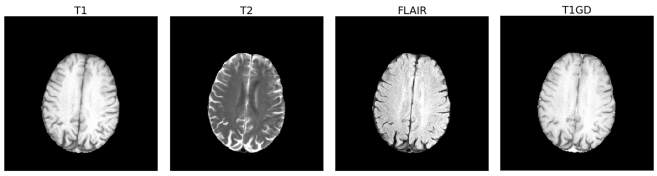

FIG. 1: Representative sample of each image modality included in the dataset.

FLAIR (n=135). A representative example of each modality is shown in Fig. 1. The volumes have an average slice-direction coverage of 150 ± 20 mm, with a mean voxel spacing of 5 ± 2 mm along the slice direction (z), indicating substantial variability across acquisitions. To increase the effective number of training samples, each 3D volume is decomposed into 2D slices along the three principal anatomical planes (axial, sagittal, and coronal), resulting in approximately 1200 2D images per patient. All image volumes were spatially aligned prior to slice extraction using affine registration with SPM.

Sequences All proposed models use T1 images as input and synthesize T2, FLAIR, and T1GD sequences as outputs. T2 and FLAIR were selected to represent complementary tissue contrast characteristics, while T1GD synthesis explores whether this sequence can be approximated without gadolinium administration.

Partition For all sequences, the data is divided into training (70%), validation (20%), and test (10%) sets. The split is performed at the patient level to avoid bias.

B. Model architectures

We introduce two of the most commonly used deep learning models in I2I tasks.

Encoder–decoder These models work in two main symmetrical stages [6]. The encoder analyzes the input image and reduces it to a dense representation that contains the most important information, such as shapes, structures, and global context. It is composed of a sequence of convolutional layers that progressively reduce the spatial resolution of the image, by using strided convolutions, while simultaneously increasing the number of feature channels, each one representing a different learned visual characteristic of the input image. Shallow layers capture low-level information such as edges, gradients, and simple textures, while deeper ones encode higher-level objects, shapes, and context relationships across the image [6]. The decoder performs the inverse operation of the encoder. It starts from the compact representation outputted by the decoder and it restores the spatial resolution of the image while reducing the number of feature channels. At each stage, the decoder combines the learned feature information using transposed convolutions to reconstruct spatial details, producing an output image with the same resolution as the input. U-Net is adopted as the backbone architecture in this work. As shown in Fig. 2, U-Net follows an encoder–decoder architecture with skip connections between corresponding encoder and decoder layers, which preserve spatial information that may be lost during the encoding process. For a detailed architectural description, see [**?**].

Conditional Generative Adversarial Networks cGANs consist of two neural networks that are trained simultaneously and complement each other [6]. The generator learns to map an input image to its output image while the discriminator assesses if this output is indistinguishable from a real image.

Unlike common GANs, which generate images from random noise without control over the output [4], cGANs condition both networks on an input image. This forces the output image to be not only realistic but also consistent with the structure of the input image. In this work, we use the pix2pix framework, a well-established cGAN approach for paired image translation [4]. As illustrated in Fig. 2, the generator follows a U-Net architecture, while the discriminator follows a PatchGAN architecture. The PatchGAN discriminator operates on local image patches instead of assessing the realism of the image at full-resolution, producing a set of real/fake predictions over the image. The final decision is obtained by aggregating these local predictions, which pushes the generator to produce realistic details and textures, while preserving the structure imposed by the input image. PatchGAN follows a fully convolutional architecture composed of successive convolutional layers. For a detailed architectural description, see [4].

C. Loss functions

The parameters of the network are optimized by minimizing a loss function using backpropagation and gradient-based optimization. The loss function defines which image properties are targeted during training and therefore plays a critical role in I2I translation. During training, it quantifies the discrepancy between the predicted image and the target and the model that achieves the lowest validation loss is selected. Common objectives in I2IT include reconstruction losses, which directly compare the input and target image at the pixel level. We employ ℓ_1,

$$\mathcal{L}_{\ell_1}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{1}$$

which penalizes deviations linearly, making it is less sensitive to large errors and preserving intensity transitions. However, it does not account for the preservation of overall structural information, which is particularly relevant in MRI. Enforcing structural similarity can be incorporated using the Structural Similarity Index Measure (SSIM),

$$\mathrm{SSIM}(y, \hat{y}) = \frac{(2\mu_y\mu_{\hat{y}} + C_1)(2\sigma_{y\hat{y}} + C_2)}{(\mu_y^2 + \mu_{\hat{y}}^2 + C_1)(\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2)}, \tag{2}$$

where μ_y and $\mu_{\hat{y}}$ denote the mean intensities of the images, σ_y^2 and $\sigma_{\hat{y}}^2$ their variances, and $\sigma_{y\hat{y}}$ the covariance between them. The constants C_i are stabilization terms.

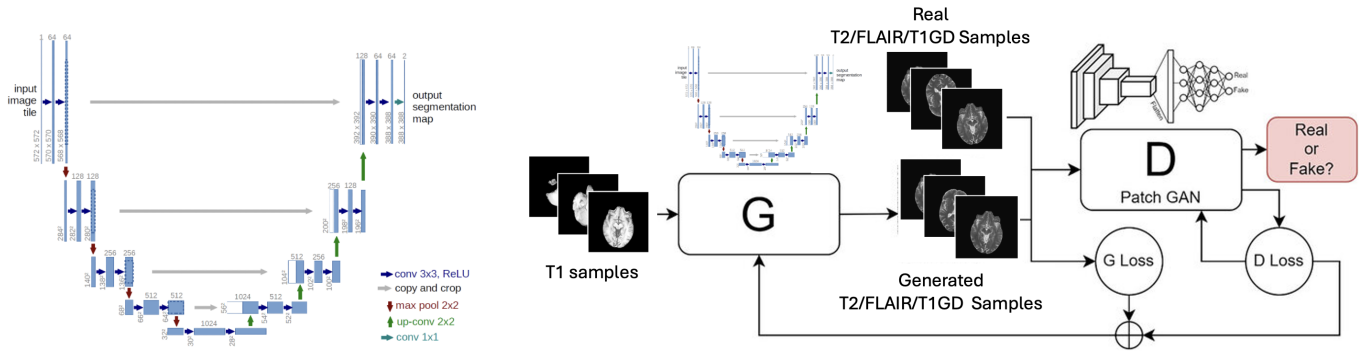

FIG. 2: Architectural overview of the models. U-Net architecture (left) [7] and cGAN (pix2pix) with U-Net generator and PatchGAN discriminator for adversarial learning (right) [8].

SSIM provides a perceptual measure of image similarity [2] by comparing luminance, contrast, and structural correlation. Therefore, the corresponding loss is $\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}$.

Other aspects of the image can also be prioritized such as preserving edges and sharp transitions, for which a gradient-based loss can be introduced:

$$\mathcal{L}_{\nabla}(y, \hat{y}) = \frac{1}{N} \sum_{i} \left(|\nabla_x y_i - \nabla_x \hat{y}_i| + |\nabla_y y_i - \nabla_y \hat{y}_i| \right), \quad (3)$$

where ∇_x and ∇_y represent the horizontal and vertical image gradients, which capture local intensity changes, promoting the preservation of tissue boundaries.

We propose a composite loss that integrates pixel-wise accuracy, global similarity, and edge preservation:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\ell_1} + \beta \mathcal{L}_{\text{SSIM}} + \gamma \mathcal{L}_{\nabla}, \quad (4)$$

where these weights are manually set to $\alpha = 1.0$, $\beta = 0.8$, and $\gamma = 0.4$. Loss functions alone do not enforce visual realism, which is a limitation of encoder–decoder architectures. PIX2PIX incorporates an adversarial loss via the discriminator, which provides another learning signal [4]. The adversarial loss is defined as

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x}[\log(1 - D(x, G(x)))], \quad (5)$$

where $\mathbb{E}$ is the expectation or average over the training samples, G denotes the generator and D the discriminator. The discriminator receives patches of images and learns to distinguish real input–target pairs (x, y) from pairs where the target image is generated by the model $(x, G(x))$. The first term encourages the discriminator to assign high scores to real pairs, while the second term penalizes it when generated pairs are incorrectly classified as real. The generator is trained to deceive the discriminator by producing images indistinguishable from real ones. The final loss of PIX2PIX combines this adversarial loss with an ℓ_1 reconstruction term:

$$\mathcal{L}_{\text{pix2pix}} = \mathcal{L}_{\text{GAN}} + \lambda \mathcal{L}_{\ell_1}. \quad (6)$$

helping to preserve the spatial alignment and intensities, while the adversarial loss focuses on visual realism, particularly details and textures. The weighting parameter λ controls the balance between these two. In its original article, this parameter is set to $\lambda = 100$ [4].

D. Evaluation metrics

We evaluate the quality of the synthesized images using two widely used metrics in I2I translation, Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR), which quantifies pixel-wise intensity differences between the image and target, derived from the mean squared error (MSE) [2]:

$$\text{PSNR}(y, \hat{y}) = 10 \log_{10} \left(\frac{I_{\text{max}}^2}{\text{MSE}(y, \hat{y})} \right) \quad (7)$$

$$\text{MSE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \quad (8)$$

where I_{max} denotes the maximum possible image intensity. Higher PSNR values indicate more accurate intensity reconstruction at the pixel level.

III. EXPERIMENTAL SETUP

A. Training configurations

Although convolutional architectures such as U-Net are often defined for a fixed input size, our implementation allows flexible spatial dimensions, enabling different training and inference strategies.

U-NET-BASE is trained and evaluated on full 512×512 slices, using data augmentation (resize to 582×582, random crops to 512×512, 90^{o} rotations, and flips) to promote sensitivity to fine details.

U-NET-PATCH uses the same architecture but is trained on 256×256 patches extracted from 512×512 slices, while inference is performed on full-resolution inputs. This increases the effective spatial detail seen during training while preserving 512×512 outputs, using the same augmentation pipeline as UNET-BASE.

In contrast, **PIX2PIX** enforces a fixed input size due to its coupled generator–discriminator design. We therefore

TABLE I: Results for cross-sequence MRI synthesis using SSIM (↑) and PSNR (↑) evaluated on our private *Hospital Clínic* dataset. Best results are in **bold**, second best are underlined.

	SSIM (↑)				PSNR (↑)			
	T1→T2	T1→FLAIR	T1→T1GD	AVG.	T1→T2	T1→FLAIR	T1→T1GD	AVG.
UNet-Base	0.87(0.05)	0.88(0.06)	**0.92(0.05)**	0.89(0.05)	**24(4)**	**25(4)**	**27(5)**	**25(4)**
UNet-Patch	**0.89(0.05)**	**0.89(0.06)**	**0.92(0.05)**	**0.90(0.05)**	23(5)	24(4)	26(5)	24(5)
pix2pix	0.42(0.07)	0.70(0.09)	0.80(0.09)	0.64(0.08)	12(1)	14(4)	19(5)	15(3)

train and test on 256×256 patches only, which reduces resolution and complicates comparisons. Patch-wise inference at 256×256 followed by reassembly to 512×512 was explored but introduced visible stitching artifacts and misalignments. Moreover, a high-resolution pix2pix configuration was also investigated but failed to converge due to training instabilities, consistent with reported behavior in high-resolution settings [9].

B. Training details

Hyperparameters All models are trained using the AdamW optimizer. UNet-Base and pix2pix are trained for 5 epochs, while UNet-Patch is trained for 2 epochs, as the patching strategy provides four samples per original image. The LR schedule consists of a cosine schedule with a warmup for 10%, initial value of 1×10^{-4} that decays to 1×10^{-5}. The batch size is 16 for all models.

Transfer Learning To reduce training time and computational cost, we adopt a transfer learning strategy. We train a base model on the T1→T2 task for 5 epochs until convergence. These weights are used to initialize the remaining models (T1→FLAIR and T1→T1GD), which are fine-tuned for an additional 2 epochs.

Implementation We use keras and tensorflow python packages for training. We run all trainings on NVIDIA H100 GPUs. The full codebase used in this work is publicly available to ensure reproducibility: `https://github.com/jacoboromerodiaz/mri-crossmodality`.

IV. RESULTS & DISCUSSION

UNet models outperform pix2pix. As shown in Table I, pix2pix does not achieve competitive performance across any modalities, due to the following reasons. First the fact that this model operates at a lower effective resolution limits its ability to preserve high-definition details compared to UNet. Also, the adversarial loss prioritizes realism over pixel-wise similarity, which could not be optimal for this I2I task. Moreover, as the adversarial loss in pix2pix focuses on visual realism rather than medical fidelity, it can penalize intensity variations that are clinically meaningful but uncommon in the data. Therefore, diagnostically relevant elements may be eliminated, making this loss less suitable for MRI I2I translation. We also believe that imperfect alignment in some of the image-target pairs degraded pix2pix adversarial training. Overall, UNet-based models achieve performances comparable with state-of-the-art results for supervised MRI cross-sequence synthesis [10, 11], even though direct benchmark comparison was not performed.

Patch-based training improves high-resolution synthesis. As shown in Table I, the patch-based strategy achieves higher SSIM scores than the full-resolution baseline, while exhibiting slightly lower PSNR values. Most clinical MRI important information is contained on local image patterns, so despite some image context being reduced, patch-level training is effective. The skip connections in the U-Net architecture preserve spatial information in the decoder, which could also compensate for the reduced context and facilitate image reconstructions. In addition, this approach reduces memory requirements and increases the number of training samples. One limitation is that patch-based training can lead to contrast inconsistencies when local artifacts are present, as observed in the T1→FLAIR example in Fig. 3. In the sample outputted by the UNet-Patch model the global intensity coherence is lost. Samples like this one could explain the slightly lower PSNR results reported in Table I. This effect can be mitigated by aggregating all patient slices and applying normalization or histogram equalization.

Limitations of quantitative evaluation metrics Although the proposed models achieve high quantitative performance, some modality-specific structures are not preserved. In the T1→T1GD task, the two modalities share nearly identical overall appearance and differ only in localized regions affected by contrast injection. As illustrated in Fig. 3, these contrast-enhanced regions are not obtained in any of the outputs, despite high SSIM reported in Table I. This highlights the need for ROI-aware losses for explicitly targeting contrast-enhanced regions or incorporating evaluations by an radiologist to better judge realism and clinical fidelity. In contrast, for the T1→FLAIR task the meningioma present in the ground-truth image is consistently reconstructed across all models, with varying levels of definition. This suggests that brain regions that differ between sequences in textural changes can be effectively translated using the proposed approaches, although further improvements are needed for their accurate reconstruction.

V. CONCLUSIONS

We developed a lightweight U-Net model with a tailored loss function that showed stable training and achieved performance comparable to state-of-the-art methods on a private multi-sequence MRI dataset.

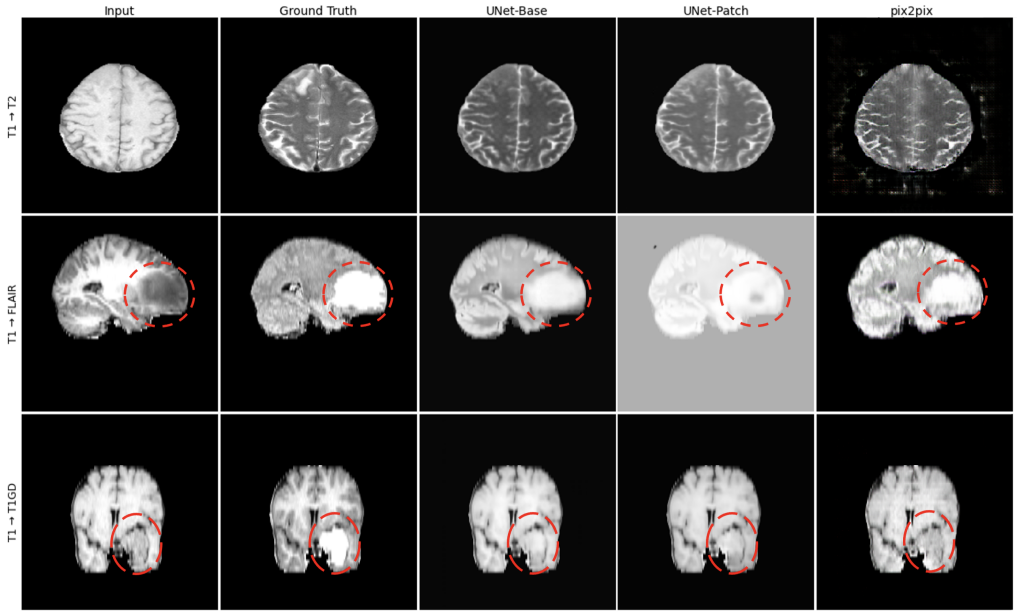

FIG. 3: Qualitative comparison of cross-sequence MRI synthesis results. Columns show, from left to right: input image, ground truth, UNet-Base, UNet-Patch, and pix2pix. Each row represents a different translation task: axial T1→T2 (top), sagittal T1→FLAIR (middle), and coronal T1→T1GD (bottom). Dashed red line highlights a meningioma in the FLAIR example and a contrast-enhanced region in the T1GD example.

Benchmarking on public datasets, such as ADNI, remains for future study, although notable differences are not expected. Despite being the gold-standard in I2IT, pix2pix does not achieve a competitive performance in this medical imaging setting, in addition to generating low-resolution images. Addressing high-resolution medical image synthesis with cGANs remains an open direction for future work. Despite reporting strong scores, subtle but clinically relevant features are poorly captured by traditional metrics, particularly in modalities where localized contrast variations dominate. This suggests the need for region-aware approaches such as evaluating segmentation quality of contrast-enhanced regions or meningioma; or incorporating radiologist evaluations.

VI. ACKNOWLEDGEMENTS

I would like to thank Aida and Arnau for their support throughout this project. I also thank Dr. Josep Puig for collecting and segmenting the dataset used in this study.

[1] Brown RW, Cheng YC, *et al.* Magnetic resonance imaging: physical principles and sequence design. John Wiley & Sons; 2014 Jun 23.
[2] Chen J, Ye Z, *et al.* Medical image translation with deep learning: Advances, datasets and perspectives. Medical Image Analysis. 2025 Apr 27:103605.
[3] Islam MT, Tsnobiladze V. The application, safety, and recent developments of commonly used gadolinium-based contrast agents in MRI: a scoping review. EMJ. 2024 Sep;9(3):63–73.
[4] Isola P, Zhu JY, *et al.* Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 1125–1134.
[5] Wang TC, Liu MY, *et al.* High-resolution image synthesis and semantic manipulation with conditional GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 8798–8807.
[6] Atienza R. Advanced deep learning with TensorFlow 2 and Keras. Birmingham, UK: Packt Publishing; 2020.
[7] Ronneberger O, Fischer P, *et al.* U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention; 2015 Oct 5. p. 234–241.
[8] Image adapted from: Adıyaman H, Emre Varul Y, *et al.* Stripe Error Correction for Landsat-7 Using Deep Learning. PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science. 2025 Mar;93(1):51-63.
[9] Chen Q, Koltun V. Photographic image synthesis with cascaded refinement networks. In Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 1511–1520.
[10] Huang Y *et al.*. Simultaneous super-resolution and cross-modality synthesis in magnetic resonance imaging. In Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics 2019 Sep 20 (pp. 437-457). Cham: Springer International Publishing.
[11] Yan Y, Wang H, *et al.* Cross-modal vertical federated learning for mri reconstruction. IEEE Journal of Biomedical and Health Informatics. 2024 Jan 31;28(11):6384-94.

Generació sintètica creuada de seqüències d'imatges de ressonància magnètica mitjançant xarxes de deep learning

Jacobo Romero Díaz, jromerdi7@alumnes.ub.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisors: Aida Niñerola Baizán, ninerola@clinic.cat, Arnau Farré Melero, afarrem@clinic.cat

Resum: La imatge per ressonància magnètica (MRI) és una modalitat diagnòstica no invasiva que utilitza diferents seqüències d'adquisició per ressaltar diferents contrasts entre teixits. En l'àmbit clínic, no totes les seqüències estan disponibles per raons de temps, cost o factors del pacient. En aquest treball, s'investiga la traducció supervisada d'imatge a imatge per generar seqüències de MRI amb un dataset de 148 pacients. Les seqüències T2, FLAIR i T1 amb contrast (T1GD) es sintetitzen a partir d'imatges T1 mitjançant models encoder-decoder (U-Net) i xarxes adversàries condicionals (pix2pix), amb entrenaments a resolució completa i en fragments. L'avaluació mitjançant l'índex de similitud estructural (SSIM) i el ratio senyal-soroll pic (PSNR) demostra que els models U-Net assoleixen resultats comparables a l'estat de l'art i superen al pix2pix. Finalment, l'avaluació qualitativa indica que les mètriques convencionals no capten regions localitzades específiques de cada seqüència, evidenciant la necessitat de criteris d'avaluació més conscients de la tasca.
Paraules clau: Traducció d'Imatge a Imatge, Imatge per Ressonància Magnètica, Aprenentatge Profund, Física Mèdica
ODSs: Aquest TFG està relacionat amb els Objectius de Desenvolupament Sostenible (SDGs)

Objectius de Desenvolupament Sostenible (ODSs o SDGs)

1. Fi de les desigualtats		10. Reducció de les desigualtats	
2. Fam zero		11. Ciutats i comunitats sostenibles	
3. Salut i benestar	X	12. Consum i producció responsables	
4. Educació de qualitat		13. Acció climàtica	
5. Igualtat de gènere		14. Vida submarina	
6. Aigua neta i sanejament		15. Vida terrestre	
7. Energia neta i sostenible		16. Pau, justícia i institucions sòlides	
8. Treball digne i creixement econòmic		17. Aliança pels objectius	
9. Indústria, innovació, infraestructures	X		

Es relaciona amb l'ODS 9, concretament amb la fita 9.5, ja que contribueix a la recerca científica i a la innovació tecnològica mitjançant el desenvolupament de mètodes d'intel·ligència artificial aplicats a la imatge mèdica. També es relaciona amb l'ODS 3, especialment amb 3.8, perquè afavoreix l'accés equitatiu a serveis de diagnòstic d'alta qualitat, i amb la fita 3.4, en donar suport al diagnòstic i seguiment de malalties no transmissibles, especialment neurològiques i oncològiques.