Contrast-enhanced to non-contrast-enhanced image translation to exploit a clinical data warehouse of T1-weighted brain MRI

Simona Bottani, Elina Thibeau-Sutre, Aurélien Maire, Sebastian Ströer, Didier Dormont, Olivier Colliot, Ninon Burgos

Homogenization of brain MRI from a clinical data warehouse using contrast-enhanced to non-contrast-enhanced image translation

Simona Bottani[a,b,c,d,e,f], **Elina Thibeau-Sutre**[b,c,d,e,f,a], **Aurélien Maire**[g], **Sebastian Ströer**[h], **Didier Dormont**[h,b,c,d,e,f,a], **Olivier Colliot**[b,c,d,e,f,a], **Ninon Burgos**[b,c,d,e,f,a,*], **the APPRIMAGE Study Group**[†]

[a]Inria, Aramis project-team, 47 boulevard de l'hôpital, 75013 Paris, France
[b]Sorbonne Université, 47 boulevard de l'hôpital, 75013 Paris, France
[c]Institut du Cerveau - Paris Brain Institute, 47 boulevard de l'hôpital, 75013 Paris, France
[d]Inserm, U 1127, 47 boulevard de l'hôpital, 75013 Paris, France
[e]CNRS, UMR 7225, 47 boulevard de l'hôpital, 75013 Paris, France
[f]AP-HP, Hôpital Pitié-Salpêtrière, 47 boulevard de l'hôpital, 75013 Paris, France
[g]AP-HP, WIND department, 33 boulevard Picpus, 75012 Paris, France
[h]AP-HP, Hôpital Pitié Salpêtrière, Department of Neuroradiology, 47 boulevard de l'hôpital, 75013 Paris, France
[†]Members of the study group are listed in section 5

Abstract.

Purpose Clinical data warehouses (CDW) provide access to massive amounts of medical images, but these images are often heterogeneous. They can for instance include images acquired both with or without the injection of a gadolinium-based contrast agent. Harmonizing such data sets is thus fundamental to guarantee unbiased results, for example when performing differential diagnosis. Furthermore, classical neuroimaging software tools for feature extraction are typically applied only to images without gadolinium. The objective of this work is to homogenize images from a CWD and enable the extraction of consistent features from brain MR images, no matter the initial presence or absence of gadolinium.

Approach We propose and compare different 3D U-Net and conditional GAN models to convert contrast-enhanced T1-weighted (T1w-ce) into non-contrast-enhanced (T1w-nce) brain MRI. These models were trained using 230 image pairs and tested on 77 image pairs from the CDW of the Greater Paris area.

Results Validation using standard image similarity measures demonstrated that the similarity between real and synthetic T1w-nce images was higher than between real T1w-nce and T1w-ce images for all the models compared. The best performing models were further validated on a segmentation task. We showed that tissue volumes extracted from synthetic T1w-nce images were closer to those of real T1w-nce images than volumes extracted from T1-ce images.

Conclusions We showed that deep learning models could synthesize T1w-nce from T1w-ce images and that reliable features could be extracted from the synthetic images, thus demonstrating the ability of such methods to homogenize a data set coming from a CDW.

Keywords: Image translation, Clinical data warehouse, Gadolinium injection, Anatomical MRI, Brain, Deep learning.

*Corresponding author: Ninon Burgos (ninon.burgos@cnrs.fr)

1 Introduction

Clinical data warehouses, gathering hundreds of thousands of medical images from numerous hospitals, offer unprecedented opportunities for research. They can for example be used to develop and validate machine learning and deep learning algorithms for the computer-aided diagnosis of neurological diseases. However, they also pose important challenges, a major challenge being their heterogeneity. Neurological diseases can result in a variety of brain lesions that are each studied with specific magnetic resonance imaging (MRI) sequences. For example, T1-weighted (T1w) brain MR images enhanced with a gadolinium-based contrast agent are used to study lesions such as tumors, and T1w images without gadolinium are used to study neurodegenerative diseases.

To perform differential diagnosis using classification algorithms, homogeneous features must be extracted from the images, no matter the disease, otherwise a link could be established between MRI sequence and pathology, which would create bias. This is critical as differential diagnosis in a clinical setting can be more challenging than in a research setting as different diseases may co-exist. Software tools such as SPM,[1] ANTs[2] or FSL[3] have been widely used for feature extraction but they were largely validated using structural T1w MRI without gadolinium, to the best of our knowledge, and their good performance on images with gadolinium is thus not guaranteed. A solution could then be to convert contrast-enhanced T1w (T1w-ce) into non-contrast-enhanced T1w (T1w-nce) brain MRI before using such tools.

Deep learning has been widely used in the image translation domain. The U-Net and conditional generative adversarial networks (GANs) appear as the two most popular options. The U-Net was originally proposed for image segmentation:[4] an encoder with convolutional and downsampling blocks is followed by a decoder with upsampling and convolutional layers. The skip con-

nections linking the encoder and decoder blocks at the same level enable the reconstruction of fine-grained details, explaining the popularity of this architecture for image translation.[5–12] Conditional GANs consist of a generator, which may adopt the U-Net architecture, followed by a discriminator in charge of distinguishing synthetic from real images and challenging the generator so that it improves the quality of the generated images. The good results obtained with conditional GANs explain their wide use for image translation.[13–22]

Both U-Net like models and conditional GANs have been proposed for diverse applications. Some aim to enhance the quality of the input images, for example by reducing noise in MRI[23–25] or positron emission tomography[26] images or by performing super-resolution.[13,15,27–29] Other works aim to translate an image of a particular modality into another modality, such as an MRI into an X-ray computed tomography (CT)[7,8,12,17,18,30] or a particular MRI sequence into another sequence.[19–22] The U-Net architecture has also been used for the data harmonization: Dewey et al.[31] built Deep-Harmony that aims to homogenize the contrast between images coming from different sites.

Closer to our application, various deep learning models have been developed for the synthesis of images with gadolinium from images without gadolinium: they include reinforcement learning for liver MRI,[32] or Gaussian mixture modeling for CT images.[33] As for the other image translation tasks, 3D U-Net like models have also been used to convert T1w-nce into T1w-ce images.[34–36] In two studies,[34,35] multimodal MRI sequences were used as input of the 3D U-Net that was trained and tested on patients with brain cancers. More specifically, the 3D U-Net proposed in the first study[35] predicts patches of T1w-ce, while the one in the second study[34] directly predicts the full 3D T1w-ce image. The residual attention U-Net described in the last work[36] outputs synthetic T1w-ce that are used for the evaluation of cerebral blood volume in mice, instead of the real T1w-ce.

Our objective in this work was to obtain a homogeneous data set of T1w-nce images from very heterogeneous images coming from a clinical data warehouse. This homogenization step should enable a consistent extraction of features that would later be used for computer-aided diagnosis in a clinical setting. We thus developed and compared different deep learning models that rely on typical architectures used in the medical image translation domain to convert T1w-ce into T1w-nce images. In particular, we implemented 3D U-Net like models with the addition of residual connections, attention modules or transformer layers. We also used these 3D U-Net like models in a conditional GAN setting. We trained and tested our models using 307 pairs of T1w-nce and T1w-ce images coming from a very large clinical data warehouse (39 different hospitals of the Greater Paris area). We first assessed synthesis accuracy by comparing real and synthetic T1w-nce images using standard metrics. We tested our models both on images of good or medium quality and on images of bad quality to ensure that deep learning models could generate accurate T1w-nce images no matter the quality of the input T1w-ce images. We then compared the volumes of gray matter, white matter and cerebrospinal fluid obtained by segmenting the real T1w-nce, real T1w-ce and synthetic T1w-nce images using SPM[37] in order to verify that features extracted from synthetic T1w-nce were reliable. Preliminary work is accepted for publication in the proceedings of the SPIE Medical Imaging 2022 conference.[38] Contributions specific to this paper include the development of additional models (a 3D U-Net like model with the addition of transformer layers, and three conditional GAN models using 3D U-Net like models as generators and a patch-based discriminator) and an extended validation of the segmentation task with a deeper analysis the tissue volume differences.

2 MATERIALS AND METHODS

2.1 Data set description

This work relies on a large clinical data set containing all the T1w brain MR images of adult patients scanned in one of the 39 hospitals of the Greater Paris area (Assistance Publique-Hôpitaux de Paris [AP-HP]). The data were made available by the AP-HP data warehouse and the study was approved by the Ethical and Scientific Board of the AP-HP. According to French regulation, consent was waived as these images were acquired as part of the routine clinical care of the patients.

Images were acquired as part of the routine clinical care in the different hospital sites and gathered in a central hospital PACS. Images relevant to the research project were copied to the research PACS and pseudonymized. They always remain within the hospital network that we accessed remotely. Images from this clinical data warehouse are very heterogeneous:[39] they include images of patients with a wide range of ages (from 18 to more than 90 years old) and diseases, acquired with different scanners (more than 30 different models) from 1980 up to now.

In a previous work,[39] we developed a quality control framework to identify images that are not proper T1w brain MRIs, to identify acquisitions for which gadolinium was injected, and to rate the overall image quality defined based on three characteristics: motion, contrast and noise. We did so by manually annotating 5500 images (out of a batch of 9941 images that were available) to train and test convolutional neural network (CNN) classifiers. The graphical interface used to manually annotate the images is publicly available (https://github.com/SimonaBottani/Quality_Control_Interface).

The data set used in this work is composed of 307 pairs of T1w-ce and T1w-nce images that were extracted from the batch of 9941 images made available by the AP-HP data warehouse. We

first selected all the images of low, medium and good quality, excluding images that were not proper T1w brain MRI,[39] resulting in 7397 images. This selection was based on manual quality control for 5500 images and on automatic quality control for the remaining 4441 images.[39] In the same way, the presence or absence of gadolinium-based contrast agent was manually noted for 5500 images, while it was obtained through the application of a CNN classifier for the remaining 4441 images. We then considered only patients having both a T1w-ce and a T1w-nce image at the same session, with a T1w-nce image of medium or good quality. Finally, to limit heterogeneity in the training data set, we visually checked all the images and excluded 52 image pairs that were potential outliers because of extremely large lesions. Among the selected images, 256 image pairs were of medium and good quality, and 51 image pairs had a T1w-ce of low quality and a T1w-nce of good or medium quality. In total the data set comprises 614 images: 534 images were acquired at 3 T and 80 at 1.5 T, 556 images were acquired with a Siemens machine (with seven different models) and 58 with a GE Healthcare machine (with five different models).

2.2 Image preprocessing

All the images were organised using the Brain Imaging Data Structure (BIDS).[40] We applied the following pre-processing using the 't1-linear' pipeline of Clinica,[41] which is a wrapper of the ANTs software.[2] Bias field correction was applied using the N4ITK method.[42] An affine registration to MNI space was performed using the SyN algorithm.[43] The registered images were further rescaled based on the min and max intensity values, and cropped to remove background resulting in images of size $169\times208\times179$, with 1 mm isotropic voxels.[44] Finally all the images were resampled to have a size of $128\times128\times128$ using trilinear interpolation in Pytorch.

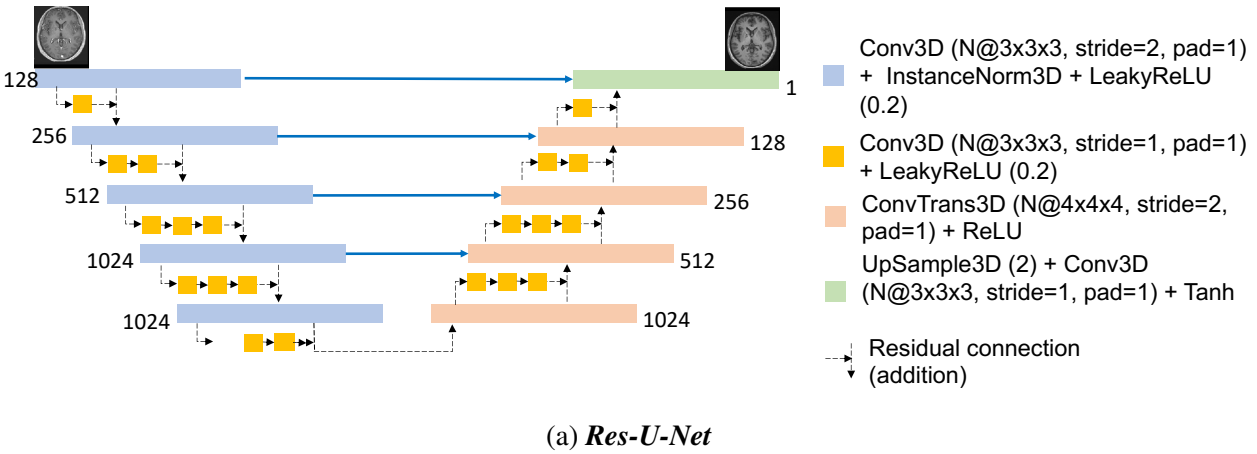
2.3 Network architecture

To generate T1w-nce from T1w-ce images, both 3D U-Net like models and conditional GANs were developed and compared. The code used to implement all the architectures and perform the experiments is openly available (https://github.com/SimonaBottani/image_synthesis).

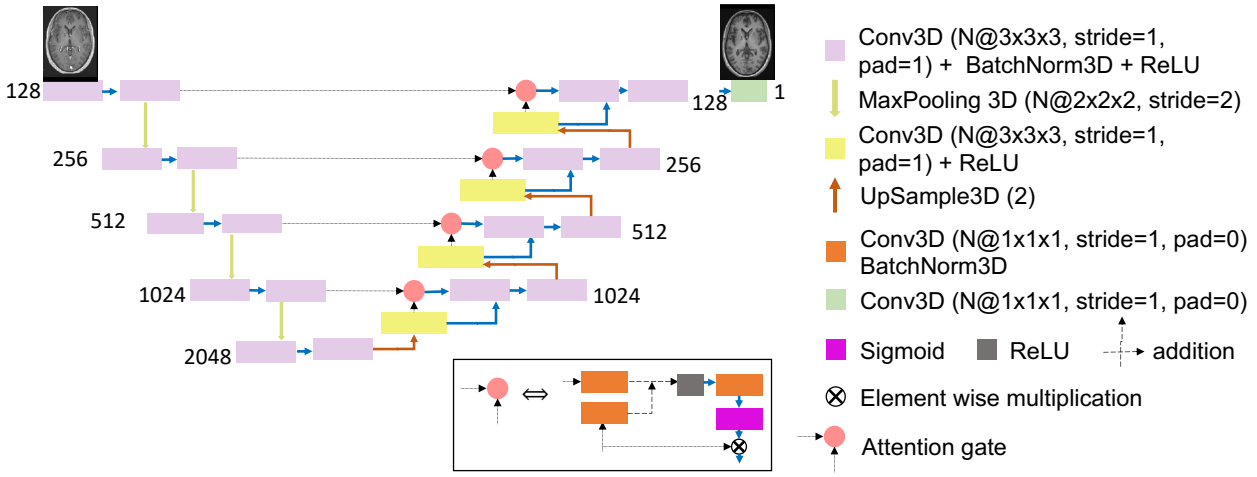
2.3.1 3D U-Net like structures

We implemented three models derived from the 3D U-Net:[4] a 3D U-Net with the addition of residual connections (called *Res-U-Net*), a 3D U-Net with the addition of attention mechanisms (called *Att-U-Net*), a 3D U-Net with both transformer and convolutional layers (called *Trans-U-net*). The U-Net structure allows preserving the details present in the original images thanks to the skip connections[4] and has shown good performance for image-to-image translation.[5–12] Here we detail the three architectures, which are also shown in Figure 1.

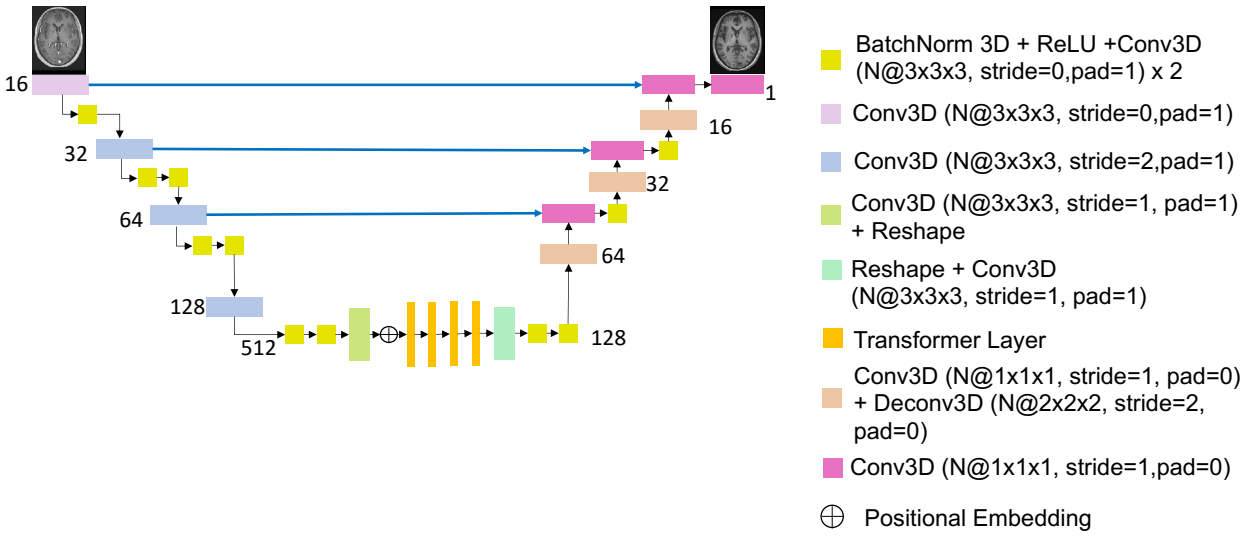
Res-U-Net The *Res-U-Net* we implemented is based on the architecture first proposed by Milletari et al.[45] and later used in Bône et al.[34] The five descending blocks are composed of 3D convolutional layers followed by an instance normalization block and a LeakyReLU (negative slope coefficient $\alpha = 0.2$). The four ascending blocks are composed of transposed convolutional layers followed by a ReLU. The final layer is composed of an upsample module (factor of 2), a 3D convolutional block and a hyperbolic tangent module. Each descending or ascending block is followed by a residual module, which can vary from one to three blocks composed of a 3D convolutional layer and a LeakyReLU ($\alpha = 0.2$). Residual blocks were introduced to avoid the problem of the vanishing gradients in the training of deep neural network:[46] they ease the training since they improve the flow of the information within the network.



(a) ***Res-U-Net***



(b) ***Att-U-Net***



(c) ***Trans-U-Net***

Fig 1: Architectures of the proposed 3D U-Net like models. The models take as input a real T1w-nce image of size 128×128×128 and generate a synthetic T1w-nce of size 128×128×128. *Res-U-Net*: images pass through five descending blocks, each one followed by a residual module, and then through four ascending blocks and one final layer. *Att-U-Net*: images pass through five descending blocks and then through four ascending blocks and one final layer. One of the input of each ascending block is the result of the attention gate. *Trans-U-Net*: images pass through four descending blocks, four transformer layers and four ascending layers. All the parameters such as kernel size, stride, padding, size of each feature map (N) are reported.

Att-U-Net We implemented the *Att-U-Net* relying on the work of Oktay et al.[47] In this architecture, the five descending blocks are composed of two blocks with a 3D convolutional layer followed by a batch normalization layer and a ReLU. They are followed by four ascending blocks. Each ascending block is composed of an upsample module (factor of 2), a 3D convolutional layer followed by a ReLU, an attention gate and two 3D convolutional layers followed by a ReLU. The attention gate is composed of two 3D convolutional layers, a ReLU, a convolutional layer and a sigmoid layer. Its objective is to identify only salient image regions: the input of the attention gate is multiplied (element-wise multiplication) by a factor (in the range 0–1) resulting from the training of all the blocks of the networks. In this way it discards parts of the images that are not relevant to the task at hand.

Trans-U-Net The *Trans-U-Net* was implemented by Wang et al.[48] (who called the model *Trans-BTS*). They proposed a 3D U-Net like structure composed of both a CNN and a transformer. The CNN is used to produce an embedding of the input images in order not to loose local information across depth and space. The features extracted by the CNN are the input of the transformer whose aim is to model the global features. The descending blocks are composed of four different blocks, each being composed of a 3D convolutional layer and one, two or three blocks composed of a batch normalization layer, a ReLU and another 3D convolutional layer. The model is then composed of four transformer layers, after a linear projection of the features. Each transformer layer is itself composed of a multi-head attention block and a feed forward network. The four ascending blocks are composed of a 3D convolutional layer and one or two blocks with a batch normalization layer, a ReLU, a 3D convolutional layer followed by a 3D deconvolutional layer. The final layer is composed of a 3D convolutional

layer and a soft-max layer.

For the three 3D U-Net like models we used the same training parameters. We used the Adam optimizer, the L1 loss, a batch size of 2 and trained during 300 epochs. The model with the best loss, determined using the training set, was saved as final model. We relied on Pytorch for the implementation.

2.3.2 Conditional GANs

Generative adversarial networks (GANs) were firstly introduced by Goodfellow et al.[49] They are generative deep learning models composed of two elements: a generator for synthesizing new examples and a discriminator for classifying whether examples are real, i.e. the original ones, or fake, i.e. synthesized by the generator. Conditional GANs (cGANs)[50] are a variant of GANs where the generator and the discriminator are conditioned by the true samples. They can only be used with paired data sets.

We propose three different cGAN models that differ in the architecture of the generators, which correspond to the three architectures presented above. The discriminator is the same for all the cGANs: it is a 3D patch CNN, first proposed by Isola et al.[51] and used in the medical image translation domain.[52,53] Its aim is to classify if each pair of patches contains two real images or a real and a fake image. The advantages of working with patches is that the discriminator focuses on the details of the images and the generator must improve them to fool the discriminator.

Our discriminator is composed of four blocks: the first three blocks are composed of a 3D convolutional layer followed by a LeakyReLU (negative slope coefficient $\alpha = 0.2$), and the last block is composed of a 3D convolutional layer and a 3D average pooling layer. From images of size $128 \times 128 \times 128$, we created eight patches of size $64 \times 64 \times 64$ with a stride of 50.

For the training of the discriminator we used the least-square-loss as proposed in 54 in order to increase the stability, thus avoiding the problem of vanishing gradients that occurs with the usual cross-entropy loss. Stability of the training was also improved using soft labels: random numbers between 0 and 0.3 represented real images and random numbers between 0.7 and 1 represented fake images.

The total loss of the cGANs combines

- the loss of the generator composed of the sum of the L1 loss (i.e. pixel-wise absolute error) computed between the generated and true images, and the least-square loss computed between the predicted probabilities of the generated images and positive labels.

- the loss of the discriminator composed of the mean of the least-square loss computed between the predicted probabilities of the true images and positive labels and the least-square loss computed between the predicted probabilities of the generated images and negative labels.

At first, both the generators and discriminators were pretrained separately. Regarding each generator, we reused the best model obtained previously. The discriminators were pretrained for the recognition of real and fake patches (fake images were obtained from each pretrained generator). The generators and discriminators were then trained together. The generator models with the best loss, determined using the training set, were saved as final models. Note that the batch size was set to 1 due to limited computing resources.

2.4 *Experiments and validation measures*

The experiments relied on 307 pairs of T1w-ce and T1w-nce images. We randomly selected 10% of the 256 image pairs of medium and good quality for testing (data set called $\text{Test}_{\text{good}}$), the other 230 image pairs being used for training. Only images of good and medium quality were used for training to ensure that the model focuses on the differences related to the presence or absence of gadolinium, and not to other factors. The remaining 51 image pairs with a T1w-ce of low quality and a T1w-nce of good or medium quality were used only for testing (data set called Test_{low}).

2.4.1 *Synthesis accuracy*

Image similarity was evaluated using the mean absolute error (MAE), peak signal-to-noise ratio (PNSR) and structural similarity (SSIM).[55] The MAE is the mean of each absolute value of the difference between the true pixel and the generated pixel and PSNR is a function of the mean squared error: these two metrics allows a direct comparison between the synthetic image and the real one. The SSIM aims to measure quality by capturing the similarity of images, it is a weighted combination of the luminance, contrast and structure. For the MAE, the minimum value is 0 (the lower, the better), for PSNR the maximum value is infinite (the higher, the better) and for SSIM the maximum value is 1 (the higher, the better). We calculated these metrics both between the real and synthetic T1w-nce images and between the real T1w-nce and T1w-ce images (as reference). These metrics were calculated within the brain region. A brain mask was obtained for each subject by skull-stripping the T1w-nce and T1w-ce images using HD-BET[56] and computing the union of the two resulting brain masks.

2.4.2 *Segmentation fidelity*

Our goal is to obtain gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) segmentations from T1w-ce images using widely-used software tools that are consistent with segmentations obtained from T1w-nce images. We thus assessed segmentation consistency by analyzing the tissue volumes resulting from the segmentations, which are important features when studying atrophy in the context of neurodegenerative diseases.

The volumes of the different tissues were obtained as follows. At first, synthetic T1w-nce images were resampled back to a size of 169×208×179 using trilinear interpolation in Pytorch so that real and synthetic images have the same grid size. We processed the images using the 't1-volume-tissue-segmentation' pipeline of Clinica.[41,57] This wrapper of the Unified Segmentation procedure implemented in SPM[37] simultaneously performs tissue segmentation, bias correction and spatial normalization. Once the probability maps were obtained for each tissue, we computed the maximum probability to generate binary masks and we multiplied the number of voxels by the voxel dimension to obtain the volume of each tissue. We calculated both the relative absolute difference (rAD) and the relative difference (rD) for each tissue between the real T1w-ce or synthetic T1w-nce and the real T1w-nce as follows:

$$\text{rAD} = \frac{|V_t^I - V_t^J|}{TIV^I} \times TIV \ , \tag{1a}$$

$$\text{rD} = \frac{V_t^I - V_t^J}{TIV^I} \times TIV \ , \tag{1b}$$

where V_t^I is the volume of tissue t extracted from the real T1w-nce image I, V_t^J is the volume of tissue t extracted from image J, J being the synthetic T1w-nce or real T1w-ce image. TIV^I corresponds to the total intracranial volume obtained from the real T1w-nce image I and TIV

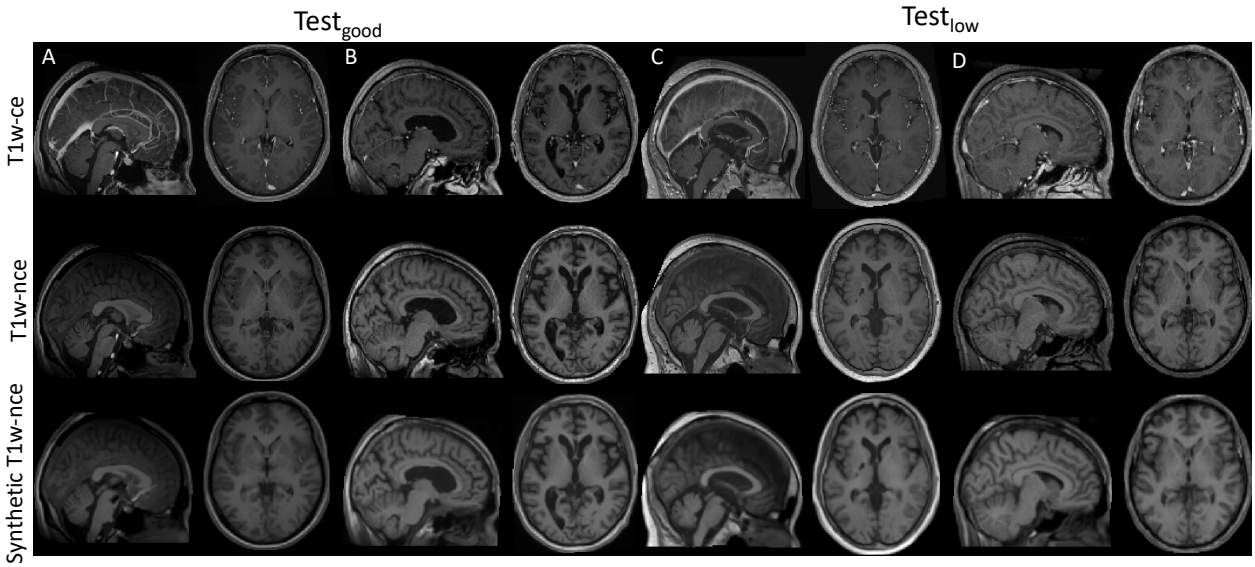


Fig 2: Examples of real T1w-ce (top), real T1w-nce (middle) and synthetic T1w-nce obtained with the *cGAN Att-U-Net* model (bottom) images in the sagittal and axial planes. Images of patients A and B belong to Test$_{good}$ (left) while images of patients C and D belong to Test$_{low}$ (right).

corresponds to the average total intracranial volume computed across the two test sets. The multiplication by the average total intracranial volume (TIV) aims at obtaining volumes (in cm^3) rather than fractions of the TIV of each subject, which is easier to interpret. Since this is a multiplication by a constant, it has not impact on the results. To assess whether the tissue volumes presented a statistically significant difference in terms of rAD depending on the images they were obtained from, we performed paired t-tests using Bonferroni correction for multiple comparisons.

In addition, we compared the binary tissue maps extracted from the real T1w-ce or synthetic T1w-nce image to those extracted from the real T1w-nce using the Dice score.

3 RESULTS

We report results for the proposed 3D U-Net like models and cGANs trained on 230 image pairs of good and medium quality, and tested on Test$_{good}$ and Test$_{low}$ obtained from a clinical data set.

Examples of synthetic T1w-nce images obtained with the *cGAN Att-U-Net* model together with the real T1w-ce and T1w-nce images are displayed in Figure 2. Images of patients A and

B belong to $\text{Test}_{\text{good}}$ while images of patients C and D belong to Test_{low}. We note the absence of contrast agent in the synthetic T1w-nce, while it is clearly visible in the sagittal slice of the T1w-ce (particularly visible for patients A and C) and that the anatomical structures are preserved between the synthetic and real T1w-nce, even in the case of a disease (as for patient B). We also note that contrast between gray and white matter is preserved in the synthetic T1w-nce (particularly visible for patients B and D). For Test_{low}, the contrast seems improved in the synthetic compared with the real T1w-ce image (especially for patient D).

3.1 Synthesis accuracy

Table 1 reports the image similarity metrics obtained for the two test sets within the brain region. We computed these metrics to assess the similarity between real and synthetic T1w-nce images, but also between T1w-nce and T1w-ce images to set a baseline. We observe that, for all models, the similarity is higher between real and synthetic T1w-nce images than between T1w-nce and T1w-ce images according to all three metrics on both test sets. The differences observed in terms of MAE, PSNR and SSIM between the baseline and each image translation approach are statistically significant (corrected p-value <0.05 according to a paired t-test corrected for multiple comparisons using the Bonferroni correction).

Among the generators composed of 3D U-Net like models, the *Att-U-Net* performed slightly better than the others, both for $\text{Test}_{\text{good}}$ (mean MAE: 2.73%, PSNR: 29.07 dB, SSIM: 0.96) and Test_{low} (mean MAE: 2.89%, PSNR: 27.18 dB, SSIM: 0.95). The performance of the cGANs were comparable to their counterparts composed only of the generator. *cGAN Att-U-Net* had a lower MAE for both test sets (mean MAE: 2.69% for $\text{Test}_{\text{good}}$ and mean MAE: 2.86% for Test_{low}). There was no statistically significant difference observed, no matter the synthesis accuracy mea-

Table 1: MAE, PSNR and SSIM obtained on the two independent test sets with various image quality. For each metric, we report the average and standard deviation across the corresponding test set. We compute the metrics for both T1w-ce and synthetic T1w-nce in relation to the real T1w-nce, and so within the brain region.

Test set	Compared images	Model	MAE (%)	PSNR (dB)	SSIM
$\text{Test}_{\text{good}}$	T1w-nce / T1w-ce	-	4.14 ± 1.59	23.03 ± 2.83	0.90 ± 0.05
	T1w-nce / Synthetic T1w-nce	*Res-U-Net*	3.06 ± 1.50	26.89 ± 4.30	0.95 ± 0.04
		Att-U-Net	2.73 ± 1.69	29.07 ± 4.53	0.96 ± 0.05
		Trans-U-Net	2.80 ± 1.42	28.00 ± 4.13	0.96 ±0.04
		cGAN Res-U-Net	3.47 ± 1.59	23.89 ± 4.30	0.95 ± 0.04
		cGAN Att-U-Net	2.69 ± 1.68	28.89 ± 4.44	0.97 ± 0.05
		cGAN Trans-U-Net	2.86±1.59	28.00 ±4.32	0.96 ± 0.04
Test_{low}	T1w-nce / T1w-ce	-	3.71 ± 1.99	24.20 ± 3.85	0.91 ± 0.06
	T1w-nce / Synthetic T1w-nce	*Res-U-Net*	2.93 ± 1.77	26.71 ± 4.32	0.95 ± 0.05
		Att-U-Net	2.89 ± 1.85	27.15 ± 4.57	0.95 ± 0.05
		Trans-U-Net	2.98 ± 1.89	26.71 ± 4.38	0.94 ± 0.05
		cGAN Res-U-Net	3.20 ± 1.96	26.20 ± 4.42	0.93 ± 0.05
		cGAN Att-U-Net	2.86 ± 1.83	27.12 ± 4.50	0.95 ± 0.05
		cGAN Trans-U-Net	2.97 ± 1.83	26.68 ± 4.40	0.94 ± 0.05

sure, between *cGAN Att-U-Net*, the best performing model according to the MAE, and the other approaches for both test sets (corrected p-value > 0.05). For further validation we kept only *Att-U-Net* and *cGAN Att-U-Net*.

3.2 Segmentation fidelity

Absolute volume differences (rAD) obtained between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the *Att-U-Net* model and the *cGAN Att-U-Net*) for GM, WM and CSF are reported in Table 2. For both test sets and all tissues, the absolute volume differences are smaller between T1w-nce and synthetic T1w-nce images than between T1w-nce and T1w-ce images for the two models. Using the *Att-U-Net* on $\text{Test}_{\text{good}}$, absolute volume differences of GM and CSF between T1w-nce/T1w-ce and T1-nce/Synthetic T1w-nce are statistically significantly different (corrected p-value <0.01 according to a paired t-test corrected for multiple comparisons using the Bonferroni correction), while on Test_{low} absolute volume dif-

Table 2: Absolute volume difference (mean $\pm$ standard deviation in cm^3) between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the *Att-U-Net* and *cGAN Att-U-Net* models) for the gray matter, white matter and cerebrospinal fluid (CSF). * indicates that the absolute volume difference between T1w-nce and synthetic T1w-nce images is statistically significantly different from that of the baseline (corrected p-value <0.01) according to a paired t-test corrected for multiple comparisons using the Bonferroni correction.

	Compared images	Model	$Test_{good}$ [cm^3]	$Test_{low}$ [cm^3]
Gray matter	T1w-nce / T1w-ce	-	26.68 ± 15.92	49.63 ± 49.38
	T1w-nce / Synthetic T1w-ce	*Att-U-Net*	10.36 ± 6.98 *	19.61 ± 29.54 *
		cGAN Att-U-Net	9.24 ± 6.10 *	19.67 ± 28.32 *
White matter	T1w-nce / T1w-ce	-	10.81 ± 3.71	25.36 ± 27.73
	T1w-nce / Synthetic T1w-ce	*Att-U-Net*	7.79 ± 5.87	13.95 ± 24.74 *
		cGAN Att-U-Net	6.40 ± 4.43 *	14.49 ± 21.06 *
CSF	T1w-nce / T1w-ce	-	61.62 ± 34.61	69.55 ± 37.77
	T1w-nce / Synthetic T1w-ce	*Att-U-Net*	13.37 ± 10.18 *	12.25 ± 7.72 *
		cGAN Att-U-Net	18.27 ± 17.20 *	17.10 ± 18.45 *

ferences of all the tissues are statistically significantly different (corrected p-value <0.01). Using the *cGAN Att-U-Net* model, absolute volume differences of all the tissues are statistically significantly different (corrected p-value <0.01) for both test sets. This means that there is an advantage in using synthetic T1w-nce images rather than T1w-ce images, no matter the model used for the synthesis: segmentation of GM, CSF and WM is more reliable since closer to the segmentation of the tissues in the real T1w-nce.

Volume differences (rD) computed between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the *Att-U-Net* and *cGAN Att-U-Net*) for GM, WM and CSF are reported in Figure 3. We observe that volumes extracted from T1w-ce images tend to be over-estimated (GM) or under-estimated (CSF) and that most of these biases disappear when tissues are extracted from synthetic T1w-nce images (mean rD closer to 0).

The Dice scores obtained when comparing the GM, WM and CSF segmentations between

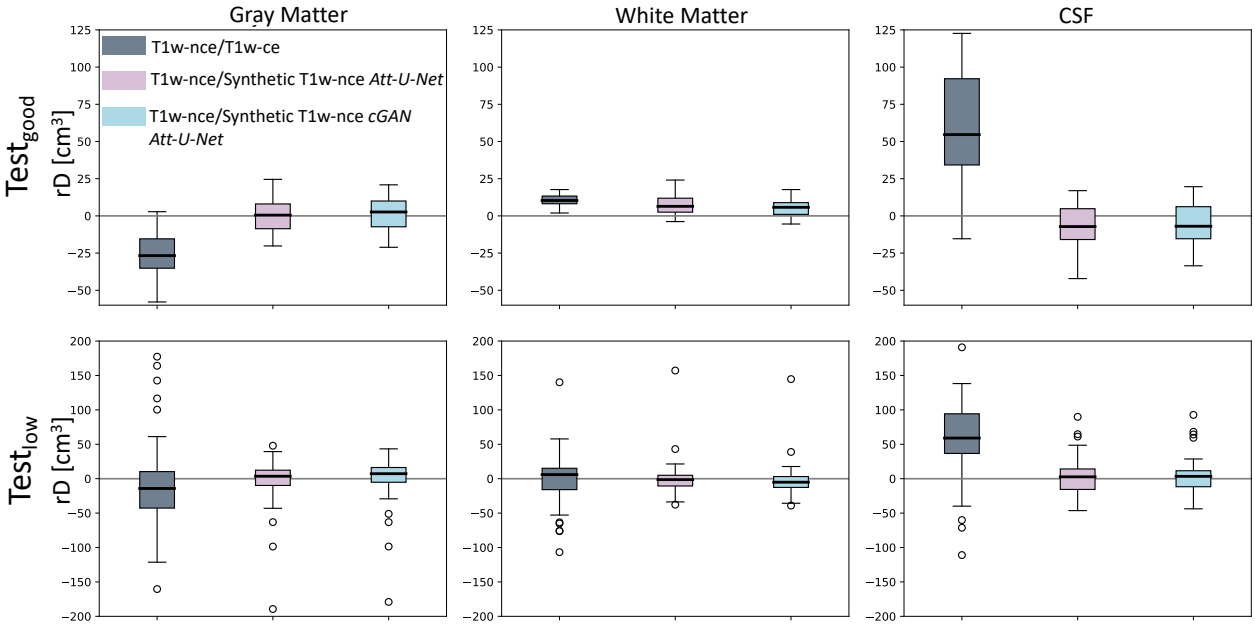


Fig 3: Volume differences (rD) in cm^3 between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the *Att-U-Net* and the *cGAN Att-U-Net* models) for gray matter (left), white matter (middle) and cerebrospinal fluid (CSF, right) for both $\text{Test}_{\text{good}}$ (top) and Test_{low} (bottom).

Table 3: Dice scores obtained when comparing the gray matter, white matter and cerebrospinal fluid (CSF) segmentations between T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the *Att-U-Net* and the *cGAN Att-U-Net*)

	Compared images	Model	$\text{Test}_{\text{good}}$	Test_{low}
Gray matter	T1w-nce / T1w-ce	-	0.88 ± 0.02	0.77 ± 0.12
	T1w-nce / Synthetic T1w-ce	*Att-U-Net*	0.87 ± 0.02	0.81 ± 0.07
		cGAN Att-U-Net	0.87 ± 0.02	0.81 ± 0.07
White matter	T1w-nce / T1w-ce	-	0.93 ± 0.01	0.85 ± 0.10
	T1w-nce / Synthetic T1w-ce	*Att-U-Net*	0.90 ± 0.02	0.86 ± 0.04
		cGAN Att-U-Net	0.91 ± 0.02	0.86 ± 0.03
CSF	T1w-nce / T1w-ce	-	0.63 ± 0.10	0.62 ± 0.10
	T1w-nce / Synthetic T1w-ce	*Att-U-Net*	0.80 ± 0.05	0.78 ± 0.07
		cGAN Att-U-Net	0.80 ± 0.05	0.78 ± 0.07

T1w-nce and T1w-ce images and between T1w-nce and synthetic T1w-nce images (obtained with the *Att-U-Net* and the *cGAN Att-U-Net*) are displayed in Table 3. We observe that for both gray

and white matter, the Dice scores are similar between T1w-nce and T1w-ce or synthetic T1w-nce images, while for CSF higher Dice scores are obtained using synthetic T1w-nce images.

4 DISCUSSION

The use of clinical images for the validation of computer-aided diagnosis (CAD) systems is still largely unexplored. One of the obstacles lies in the heterogeneity of the data acquired in the context of routine clinical practice. Post-acquisition homogenization is crucial because, contrary to research data, no strict acquisition protocols, that would ensure a certain homogeneity among the images, exist for clinical data. Heterogeneity originates from the fact that images are acquired with different scanners at different field strengths during a large period of time and because patients may suffer from a large variety of diseases. Homogenization of clinical data sets of 3D T1w brain MRI, and consequently of the features extracted from them, is an important step for the development of reliable CAD systems. Indeed, when training a CAD system, the algorithms must not be affected by the data set variations even though clinical images may greatly vary.

A source of heterogeneity among clinical data sets is the fact that they contain a mix of images acquired with and without gadolinium-based contrast agent. In our case, among the 7397 proper T1w brain images made available by the AP-HP data warehouse out of a batch of 9941 images, 59% of the images were contrast-enhanced.[39] To homogenize this data set, we thus proposed a framework to convert T1w-ce images into T1w-nce images using deep learning models. The choice to synthesize T1w-nce images from T1w-nce images was constrained by the fact that software tools for feature extraction in the neuroimaging community were developed for T1w-nce MRI. To the best of our knowledge, none of these tools has largely been applied to the extraction of features from T1w-ce MRI data and their performance in this scenario is thus mostly unknown.

The contribution of our work consists in the development and validation of deep learning models (U-Net models and conditional GANs) for the translation of T1w-ce to T1w-nce images coming from a clinical data warehouse. We compared three 3D U-net models differentiated by the addition of residual modules, of attention modules or of transformer layers, used as simple generators and also within a conditional GAN setting with the addition of a patch-based discriminator. These models have widely been used for the image translation of medical images,[58,59] but their application to clinical data has not been proven yet. The proposed models were trained using 230 image pairs and tested on two different test sets: 26 image pairs had both a T1w-nce and T1w-ce of good or medium quality and 51 image pairs had a T1w-nce of good or medium quality and a T1w-ce of bad quality. Having two test sets of different qualities is a key point since we are dealing with a real clinical heterogeneous data set where images of low quality, corresponding in majority to T1w-ce images with a low contrast, may represent 30% of the data.[39]

We first assessed the similarity between real and synthetic T1w-nce images and between real T1w-nce and T1w-ce images using three similarity metrics, MAE, PSNR and SSIM. We showed that the similarity between real and synthetic T1w-nce images was higher than the similarity between real T1w-nce and T1w-ce images according to all the metrics, no matter the models used nor the quality of the input image. The synthesis accuracy obtained with the models evaluated was of the same order as the one reached in recent works on non-contrast-enhanced to contrast-enhanced image translation.[34,35] The performance of all the models was equivalent (no statistically significant difference observed), meaning that all were able to synthesize T1w-nce images. Slightly better performance was reached with the addition of attention modules (*Att-U-Net* and *cGAN Att-U-Net* models), these models were thus further evaluated.

In the second step of the validation, we assessed the similarity of features extracted from the

different images available using a widely adopted segmentation framework, SPM.[1] We showed that the absolute volume differences of GM, WM and CSF were larger between real T1w-nce and Tw-ce images than between real and synthetic T1w-nce images (statistically significant difference most of the times). This confirms the hypothesis that gadolinium-based contrast agent may alter the contrast between the different brain tissues, making features extracted from such images with standard segmentation tools, here SPM,[1] unreliable. At the same time, we validated the suitability of the synthetic images since their segmentation was consistent with those obtained from real T1w-nce images as the volume differences were small. In particular we see that for both test sets, volume differences are statistically significantly different (corrected p-value$<$0.01 according to a paired t-test corrected for multiple comparisons using the Bonferroni correction) for GM which is the main feature when studying atrophy in neurodegenerative diseases. The fact that the relative differences between the volumes extracted from the real and synthetic T1w-nce images are relatively close to zero show that the tissue volumes are not systematically under- or over-estimated when extracted from the synthetic images.

Even though the synthetic T1w-nce images enable the extraction of reliable features, their quality could still be improved. Many constraints exist when working with data from a clinical data warehouse. One is the fact that these data are accessible only through a closed environment provided by the IT department of the AP-HP as described in.[60] Limitations in computational resources and storage space make training deep learning models difficult and thus limits the experiments that can be performed to find the optimal model. The proposed models could be improved by better optimizing the hyperparameters (such as the learning rate or the size of the kernels), adding a perceptual loss when training the conditional GANs[61] or adding more layers in the patch-based discriminator. Other architectures could also be explored. We have restricted our work to condi-

tional GANs, which need paired data to be trained, but we could exploit more data working with cycle GANs[62] as they can deal with unpaired data.

Several steps remain to be performed before using synthetic T1w-nce images for the differential diagnosis of neurological diseases. First, the performance of CAD systems trained with a mix of real T1w-nce and T1w-ce images should be compared with the performance of CAD systems trained with a mix of real and synthetic T1w-nce images. To prevent introducing a correlation between image properties (e.g. smoothness) and pathology, which would bias the classification performance, it may be necessary to also feed the real T1w-nce images to the neural network and use the resulting images as inputs of the CAD system, as suggested in 31.

5 CONCLUSION

Clinical data warehouses offer fantastic opportunities for computer-aided diagnosis of neurological diseases but their heterogeneity must be reduced to avoid biases. In this work we proposed to homogenize such a large clinical data set by converting images acquired after the injection of gadolinium into non-contrast-enhanced images using 3D U-Net models and conditional GANs. Validation using standard image similarity measures demonstrated that the similarity between real and synthetic T1w-nce images was higher than between real T1w-nce and T1w-ce images for all the models compared. We also showed that features extracted from the synthetic images (GM, WM, CSF volumes) were closer to those obtained from the T1w-nce brain MR images (considered as reference) than the original T1w-ce images. These results demonstrate the ability of deep learning methods to homogenize a data set coming from a clinical data warehouse.

DISCLOSURES

Competing financial interests related to the present article: none to disclose for all authors.

Competing financial interests unrelated to the present article: OC reports having received consulting fees from AskBio (2020), having received fees for writing a lay audience short paper from Expression Santé (2019). Members from his laboratory have co-supervised a PhD thesis with myBrainTechnologies (2016-2019) and with Qynapse (2017-present). OC's spouse is an employee and holds stock-options of myBrainTechnologies (2015-present). O.C. holds a patent registered at the International Bureau of the World Intellectual Property Organization (PCT/IB2016/0526993, Schiratti J-B, Allassonniere S, Colliot O, Durrleman S, A method for determining the temporal progression of a biological phenomenon and associated methods and devices) (2017).

ACKNOWLEDGMENTS

The research was done using the Clinical Data Warehouse of the Greater Paris University Hospitals. The authors are grateful to the members of the AP-HP WIND and URC teams, and in particular Stéphane Bréant, Florence Tubach, Jacques Ropers, Antoine Rozès, Camille Nevoret, Christel Daniel, Martin Hilka, Yannick Jacob, Cyrina Saussol, Julien Dubiel, Faerber Philippe, Mmaka Ibrahim and Gozlan Rafel. They would also like to thank the "Collégiale de Radiologie of AP-HP" as well as, more generally, all the radiology departments from AP-HP hospitals.

The research leading to these results has received funding from the Abeona Foundation (project Brain@Scale), from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6).

APPRIMAGE Study Group

Olivier Colliot, Ninon Burgos, Simona Bottani [1]

Didier Dormont [1,2], Samia Si Smail Belkacem, Sebastian Ströer [2]

Nathalie Boddaert [3]

Farida Benoudiba, Ghaida Nasser, Claire Ancelet, Laurent Spelle [4]

Hubert Ducou-Le-Pointe[5]

Catherine Adamsbaum[6]

Marianne Alison[7]

Emmanuel Houdart[8]

Robert Carlier [9,17]

Myriam Edjlali[9]

Betty Marro[10,11]

Lionel Arrive[10]

Alain Luciani[12]

Antoine Khalil[13]

Elisabeth Dion[14]

Laurence Rocher[15]

Pierre-Yves Brillet[16]

Paul Legmann, Jean-Luc Drape [18]

Aurélien Maire, Stéphane Bréant, Christel Daniel, Martin Hilka, Yannick Jacob, Julien Dubiel, Cyrina Saussol [19]

Florence Tubach, Jacques Ropers, Antoine Rozès, Camille Nevoret [20]

[1] Sorbonne Université, Institut du Cerveau - Paris Brain Institute, Inserm, CNRS, AP-HP, Hôpital de la Pitié Salpêtrière, Inria, Aramis project-team, F-75013, Paris, France

[2] AP-HP, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France

[3] AP-HP, Hôpital Necker, Department of Radiology, F-75015, Paris, France

[4] AP-HP, Hôpital Bicêtre, Department of Radiology, F-94270, Le Kremlin-Bicêtre, France

[5] AP-HP, Hôpital Armand-Trousseau, Department of Radiology, F-75012, Paris, France

[6] AP-HP, Hôpital Bicêtre, Department of Pediatric Radiology, F-94270, Le Kremlin-Bicêtre, France

[7] AP-HP, Hôpital Robert-Debré, Department of Radiology, F-75019, Paris, France

[8] AP-HP, Hôpital Lariboisière , Department of Neuroradiology, F-75010, Paris, France

[9] AP-HP, Hôpital Raymond-Poincaré, Department of Radiology, F-92380, Garches, France

[10] AP-HP, Hôpital Saint-Antoine, Department of Radiology, F-75012, Paris, France

[11] AP-HP, Hôpital Tenon, Department of Radiology, F-75020, Paris, France

[12] AP-HP, Hôpital Henri-Mondor, Department of Radiology, F-94000, Créteil, France

[13] AP-HP, Hôpital Bichat, Department of Radiology, F-75018, Paris, France

[14] AP-HP, Hôpital Hôtel-Dieu, Department of Radiology, F-75004, Paris, France

[15] AP-HP, Hôpital Antoine-Béclère, Department of Radiology, F-92140, Clamart, France

[16] AP-HP, Hôpital Avicenne, Department of Radiology, F-93000, Bobigny, France

[17] AP-HP, Hôpital Ambroise Paré, Department of Radiology, F-92100 104, Boulogne-Billancourt, France

[18] AP-HP, Hôpital Cochin, Department of Radiology, F-75014, Paris, France

[19] AP-HP, WIND department, F-75012, Paris, France

[20] AP-HP, Unité de Recherche Clinique, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France

476

477 *References*

478 1 W. D. Penny, K. J. Friston, J. T. Ashburner, *et al.*, *Statistical parametric mapping: the analy-*
479 *sis of functional brain images*, Elsevier (2011).

480 2 B. B. Avants, N. J. Tustison, M. Stauffer, *et al.*, "The Insight ToolKit image registration
481 framework," *Frontiers in Neuroinformatics* **8**, 44 (2014).

482 3 J. Mark, F. B. Christian, E. B. Timothy, *et al.*, "FSL," *NeuroImage* **62**(2), 782–790 (2012).

483 4 O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical im-
484 age segmentation," in *International Conference on Medical image computing and computer-*
485 *assisted intervention*, 234–241, Springer (2015).

486 5 X. Han, "MR-based synthetic CT generation using a deep convolutional neural network
487 method," *Medical Physics* **44**(4), 1408–1419 (2017).

488 6 I. Shiri, P. Ghafarian, P. Geramifar, *et al.*, "Direct attenuation correction of brain PET images
489 using only emission data via a deep convolutional encoder-decoder (Deep-DAC)," *European*
490 *radiology* **29**(12), 6867–6879 (2019).

491 7 K. Gong, J. Yang, K. Kim, *et al.*, "Attenuation correction for brain PET imaging using deep
492 neural network based on Dixon and ZTE MR images," *Physics in Medicine & Biology* **63**(12),
493 125011 (2018).

494 8 C. N. Ladefoged, L. Marner, A. Hindsholm, *et al.*, "Deep learning based attenuation correc-
495 tion of PET/MRI in pediatric brain tumor patients: evaluation in a clinical setting," *Frontiers*
496 *in neuroscience* **12**, 1005 (2019).

9 K. D. Spuhler, J. Gardus, Y. Gao, *et al.*, "Synthesis of patient-specific transmission data for PET attenuation correction for PET/MRI neuroimaging using a convolutional neural network," *Journal of nuclear medicine* **60**(4), 555–560 (2019).

10 J. Yang, D. Park, G. T. Gullberg, *et al.*, "Joint correction of attenuation and scatter in image space using deep convolutional neural networks for dedicated brain 18F-FDG PET," *Physics in medicine & biology* **64**(7), 075019 (2019).

11 S. Neppl, G. Landry, C. Kurz, *et al.*, "Evaluation of proton and photon dose distributions recalculated on 2D and 3D Unet-generated pseudoCTs from T1-weighted MR head scans," *Acta Oncologica* **58**(10), 1429–1434 (2019).

12 J. M. Wolterink, A. M. Dinkla, M. H. Savenije, *et al.*, "Deep MR to CT synthesis using unpaired data," in *International workshop on simulation and synthesis in medical imaging*, 14–23, Springer (2017).

13 Y. Chen, F. Shi, A. G. Christodoulou, *et al.*, "Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 91–99, Springer (2018).

14 J. Gu, Z. Li, Y. Wang, *et al.*, "Deep generative adversarial networks for thin-section infant MR image reconstruction," *IEEE Access* **7**, 68290–68304 (2019).

15 K. H. Kim, W.-J. Do, and S.-H. Park, "Improving resolution of MR images with an adversarial network incorporating images with different contrast," *Medical physics* **45**(7), 3120–3131 (2018).

16 A. M. Dinkla, J. M. Wolterink, M. Maspero, *et al.*, "MR-only brain radiation therapy: dosi-

metric evaluation of synthetic CTs generated by a dilated convolutional neural network," *International Journal of Radiation Oncology* Biology* Physics* **102**(4), 801–812 (2018).

17 H. Emami, M. Dong, S. P. Nejad-Davarani, *et al.*, "Generating synthetic CTs from magnetic resonance images using generative adversarial networks," *Medical physics* **45**(8), 3627–3636 (2018).

18 D. Nie, R. Trullo, J. Lian, *et al.*, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering* **65**(12), 2720–2730 (2018).

19 S. U. Dar, M. Yurt, L. Karacan, *et al.*, "Image synthesis in multi-contrast MRI with conditional generative adversarial networks," *IEEE Transactions on Medical Imaging* **38**(10), 2375–2388 (2019).

20 B. Yu, L. Zhou, L. Wang, *et al.*, "Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis," *IEEE transactions on medical imaging* **38**(7), 1750–1762 (2019).

21 H. Li, J. C. Paetzold, A. Sekuboyina, *et al.*, "DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 795–803, Springer (2019).

22 A. Sharma and G. Hamarneh, "Missing MRI pulse sequence synthesis using multi-modal generative adversarial network," *IEEE transactions on medical imaging* **39**(4), 1170–1183 (2019).

23 A. Benou, R. Veksler, A. Friedman, *et al.*, "Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced MRI sequences," *Medical Image Analysis* **42**, 145–159 (2017).

24 D. Jiang, W. Dou, L. Vosters, *et al.*, "Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network," *Japanese journal of radiology* **36**(9), 566–574 (2018).

25 M. Ran, J. Hu, Y. Chen, *et al.*, "Denoising of 3D magnetic resonance images using a residual encoder–decoder Wasserstein generative adversarial network," *Medical image analysis* **55**, 165–180 (2019).

26 F. Hashimoto, H. Ohba, K. Ote, *et al.*, "Dynamic PET image denoising using deep convolutional neural networks without prior training datasets," *IEEE Access* **7**, 96594–96603 (2019).

27 J. Du, L. Wang, Y. Liu, *et al.*, "Brain mri super-resolution using 3d dilated convolutional encoder–decoder network," *IEEE Access* **8**, 18938–18950 (2020).

28 C.-H. Pham, A. Ducournau, R. Fablet, *et al.*, "Brain MRI super-resolution using deep 3D convolutional networks," in *2017 IEEE ISBI*, 197–200 (2017).

29 K. Zeng, H. Zheng, C. Cai, *et al.*, "Simultaneous single-and multi-contrast super-resolution for brain MRI images based on a convolutional neural network," *Computers in Biology and Medicine* **99**, 133–141 (2018).

30 X. Han, "MR-based synthetic CT generation using a deep convolutional neural network method," *Medical physics* **44**(4), 1408–1419 (2017).

31 B. E. Dewey, C. Zhao, J. C. Reinhold, *et al.*, "DeepHarmony: a deep learning approach to contrast harmonization across scanner changes," *Magnetic Resonance Imaging* **64**, 160–170 (2019).

32 C. Xu, D. Zhang, J. Chong, *et al.*, "Synthesis of gadolinium-enhanced liver tumors on nonen-

562 hanced liver mr images using pixel-level graph reinforcement learning," *Medical Image Anal-*
563 *ysis* **69**, 101976 (2021).

564 33 M. Seo, D. Kim, K. Lee, *et al.*, "Neural contrast enhancement of ct image," in *Proceedings of*
565 *the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3973–3982 (2021).

566 34 A. Bône, S. Ammari, J.-P. Lamarque, *et al.*, "Contrast-enhanced brain MRI synthesis with
567 deep learning: key input modalities and asymptotic performance," in *2021 IEEE ISBI*, (2021).

568 35 J. Kleesiek, J. N. Morshuis, F. Isensee, *et al.*, "Can virtual contrast enhancement in brain MRI
569 replace gadolinium?: a feasibility study," *Investigative Radiology* **54**(10), 653–660 (2019).

570 36 H. Sun, X. Liu, X. Feng, *et al.*, "Substituting Gadolinium in Brain MRI Using DeepContrast,"
571 in *2020 IEEE ISBI*, 908–912 (2020).

572 37 J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage* **26**(3), 839–851 (2005).

573 38 S. Bottani, E. Thibeau-Sutre, A. Maire, *et al.*, "Homogenization of brain MRI from a clinical
574 data warehouse using contrast-enhanced to non-contrast-enhanced image translation with U-
575 Net derived models," in *SPIE Medical Imaging 2022*, (2022).

576 39 S. Bottani, N. Burgos, A. Maire, *et al.*, "Automatic quality control of brain T1-weighted mag-
577 netic resonance images for a clinical data warehouse," *Medical Image Analysis* **75**, 102219
578 (2022).

579 40 K. J. Gorgolewski, T. Auer, V. D. Calhoun, *et al.*, "The brain imaging data structure, a format
580 for organizing and describing outputs of neuroimaging experiments," *Scientific data* **3**(1), 1–9
581 (2016).

582 41 A. Routier, N. Burgos, M. Díaz, *et al.*, "Clinica: An Open-Source Software Platform for
583 Reproducible Clinical Neuroscience Studies," *Frontiers in Neuroinformatics* **15**, 39 (2021).

42 N. J. Tustison, B. B. Avants, P. A. Cook, *et al.*, "N4ITK: improved N3 bias correction," *IEEE Transactions on Medical Imaging* **29**(6), 1310–1320 (2010).

43 B. B. Avants, C. L. Epstein, M. Grossman, *et al.*, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis* **12**(1), 26–41 (2008).

44 J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, *et al.*, "Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation," *Medical Image Analysis* **63**, 101694 (2020).

45 F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, 565–571, IEEE (2016).

46 K. He, X. Zhang, S. Ren, *et al.*, "Identity mappings in deep residual networks," in *European conference on computer vision*, 630–645, Springer (2016).

47 O. Oktay, J. Schlemper, L. L. Folgoc, *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," in *Medical Imaging with Deep Learning - MIDL 2018*, (2018).

48 W. Wang, C. Chen, M. Ding, *et al.*, "Transbts: Multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 109–119, Springer (2021).

49 I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems* **27** (2014).

50 M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784* (2014).

606 51 P. Isola, J.-Y. Zhu, T. Zhou, *et al.*, "Image-to-image translation with conditional adversarial
607 networks," in *Proceedings of the IEEE conference on computer vision and pattern recogni-*
608 *tion*, 1125–1134 (2017).

609 52 W. Wei, E. Poirion, B. Bodini, *et al.*, "Predicting PET-derived demyelination from multi-
610 modal MRI using sketcher-refiner adversarial training for multiple sclerosis," *Medical image*
611 *analysis* **58**, 101546 (2019).

612 53 H. Choi and D. S. Lee, "Generation of structural MR images from amyloid PET: application
613 to MR-less quantification," *Journal of Nuclear Medicine* **59**(7), 1111–1117 (2018).

614 54 X. Mao, Q. Li, H. Xie, *et al.*, "Least squares generative adversarial networks," in *Proceedings*
615 *of the IEEE international conference on computer vision*, 2794–2802 (2017).

616 55 Z. Wang, A. C. Bovik, H. R. Sheikh, *et al.*, "Image quality assessment: from error visibility
617 to structural similarity," *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004).

618 56 F. Isensee, M. Schell, I. Pflueger, *et al.*, "Automated brain extraction of multisequence MRI
619 using artificial neural networks," *Human Brain Mapping* **40**(17), 4952–4964 (2019).

620 57 J. Samper-González, N. Burgos, S. Bottani, *et al.*, "Reproducible evaluation of classifica-
621 tion methods in Alzheimer's disease: Framework and application to MRI and PET data,"
622 *NeuroImage* **183**, 504–521 (2018).

623 58 X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A re-
624 view," *Medical image analysis* **58**, 101552 (2019).

625 59 N. Burgos, S. Bottani, J. Faouzi, *et al.*, "Deep learning for brain disorders: from data pro-
626 cessing to disease treatment," *Briefings in Bioinformatics* **22**(2), 1560–1576 (2021).

627 60 C. Daniel and E. Salamanca, "Hospital Databases," in *Healthcare and Artificial Intelligence*,
628 B. Nordlinger, C. Villani, and D. Rus, Eds., 57–67, Springer (2020).
629 61 H. Zhao, O. Gallo, I. Frosio, *et al.*, "Loss functions for image restoration with neural net-
630 works," *IEEE Transactions on computational imaging* **3**(1), 47–57 (2016).
631 62 J.-Y. Zhu, T. Park, P. Isola, *et al.*, "Unpaired image-to-image translation using cycle-
632 consistent adversarial networks," in *Proceedings of the IEEE international conference on*
633 *computer vision*, 2223–2232 (2017).

Simona Bottani is a PhD student in the ARAMIS Lab, a joint laboratory between CNRS, Inria, Inserm and Sorbonne University within the Paris Brain Institute. She received a master degree in Biomedical Engineering at Politecnico di Torino in 2016. Her research focuses on deep learning, clinical data warehouse and neurodegenerative diseases.

Elina Thibeau-Sutre was a PhD student in the ARAMIS Lab, a joint laboratory between CNRS, Inria, Inserm and Sorbonne University within the Paris Brain Institute. Her research interest includes deep learning application to neuroimaging data, its interpretability and reproducibility. She received master degrees from Ecole Nationale des Mines de Paris and Ecole supérieure de physique et de chimie industrielles (Paris, France).

Aurélien Maire is the head of the imaging group of the Assistance Publique-Hôpitaux de Paris (AP-HP) Megadata Platform, which exploits the AP-HP data warehouse.

Sebastian Ströer is a neuroradiologist working in the neuroradiology department of the Pitié Salpêtrière hospital (Assistance Publique-Hôpitaux de Paris).

Didier Dormont is Professor of Neuroradiology at the Pitié Salpêtrière hospital (Assistance Publique-Hôpitaux de Paris) and affiliated to Sorbonne Université. He is the Medical Director of the DIAMENT Department (Radiology, Nuclear Medicine and Pathologic Anatomy), AP-HP/Sorbonne Université.

Olivier Colliot is a Research Director at CNRS and the co-head of the ARAMIS Lab, a joint laboratory between CNRS, Inria, Inserm and Sorbonne University within the Paris Brain Institute (ICM), France. He received his PhD in Signal Processing from Telecom ParisTech in 2003 and the Habilitation degree from University Paris-Sud in 2011. His research interests include machine

learning, medical image analysis and their applications to neurological disorders.

Ninon Burgos is a CNRS researcher in the ARAMIS Lab, a joint laboratory between CNRS, Inria, Inserm and Sorbonne University within the Paris Brain Institute, France. She completed her PhD at University College London, UK, in 2016. Her research focuses on the development of computational imaging tools to improve the understanding and diagnosis of dementia.

List of Figures

List of Tables