Project title: Accelerating the lab to market transition of AI tools for cancer management

Grant Agreement: 952172

Call identifier: H2020-SC1-FA-DTS-2019-1

Topic: DT-TDS-05-2020 AI for Health Imaging

D6.7 COMPARATIVE ASSESSMENT OF THE PROPOSED HARMONISATION APPROACHES

Leader partner:	Guang Yang
Author(s):	Yang Nan, Eduardo Ibor Crespo
Work Package:	WP6
Due date:	Month 36
Actual delivery date:	31/08/2023
Type:	R (Report)
Dissemination level:	PU (Public)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952172.

Tables of contents

Abbreviations

AI	Artificial Intelligence
GAN	Generative adversarial network
LDM	Latent Diffusion Models
GLCM	grey level co-occurrence matrix
CT	computerised tomography
MRI	magnetic resonance imaging

Disclaimer

The opinions stated in this report reflects the opinions of the authors and not the opinion of the European Commission.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952172.

1. Introduction

Eliminating the inconsistency and variation in multicentre data has been a challenge for large scale clinical research. This involves synthesizing clinical attributes from data gathered using various equipment and methodologies, aiming to enhance reliability and resilience. Such variance mainly arises due to the diverse machinery used in capturing data, evident in signals, CT scans, MRIs, and pathological imagery. Several factors contribute to this inconsistency, including differences in vendor detection systems, coil sensitivities, changes in position and physiology during data capture, and magnetic field shifts in MRI [1-4]. Research indicates that machine learning techniques, particularly deep neural networks, are profoundly impacted by the nature of their training data. Consequently, strategies that can reduce inconsistency and variances across devices and sites are urgently needed.

In digital healthcare, data harmonisation aims to eliminate biases (non-biological discrepancies) resulting from varied data collection methods. This involves the use of computational techniques, like machine learning and image/signal processing, to merge multicentre data and decrease its non-biological discrepancies. Implementing such computational solutions encompasses stages like dataset collection, pre-processing, modelling, and analysis. Data harmonisation can be executed by processing images/signals/genetic data matrices (on a sample basis) or by aligning derived data features (on a feature basis). In the case of the synthesis of artificial images, especially in the field of image harmonization, the state of the art is mainly composed by GAN-type architectures. GAN-type architectures are artificial intelligence models designed to work in an adversarial way, one way to visualize it would be to imagine two architectures in which one of them generates a result and the other evaluates or validates this result, in such a way that the first one will generate each time more realistic results and the second will be increasingly demanding with the results of the first. In the case of image harmonization, this architecture has been used on many occasions to conveniently separate content from style, that is, being able to maintain the image content of a sample, but based on contrast, textures, and subtle image features of the second sample.

This deliverable (D6.7) aims to compare different approaches for data harmonisation to improve the performance of computational modules and provide harmonised data for clinical practice. These algorithms will be applied as a supervised-learning technique to generate a converted image where image quality has been learned from a pre-defined hand-crafted ground truth based on best state-of-the-art existing filtering, registration and normalization algorithms. The hand-crafted images can be acquired by adding some artefacts (e.g., some noise or linear/nonlinear transformation) to the original raw images.

2. Methodology

Computational approaches for data harmonisation mainly include plain CNN-based approaches (U-Net, Auto-Encoder, etc.) and generative-based approaches (WGAN, CycleGAN, LDM). In this study, we first evaluated these two schemes separately, followed by comparing the best solutions for these two schemes.

2.1 CNN-based method for harmonisation

In this section, we illustrate the details of plain CNN-based models in the experiments. We first build a baseline model, “Plain CNN”, upon the simplest backbone to avoid any possible benefits brought by the architecture itself. All Auto-Encoder (AE) models are then constructed by introducing resizing layers into Plain CNN to ensure the feature resizing operation is the only control variable in the comparison experiments. Moreover, we adapt the U-Net, primarily to investigate the effect of skip connections on the performance of AE models.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 952172.

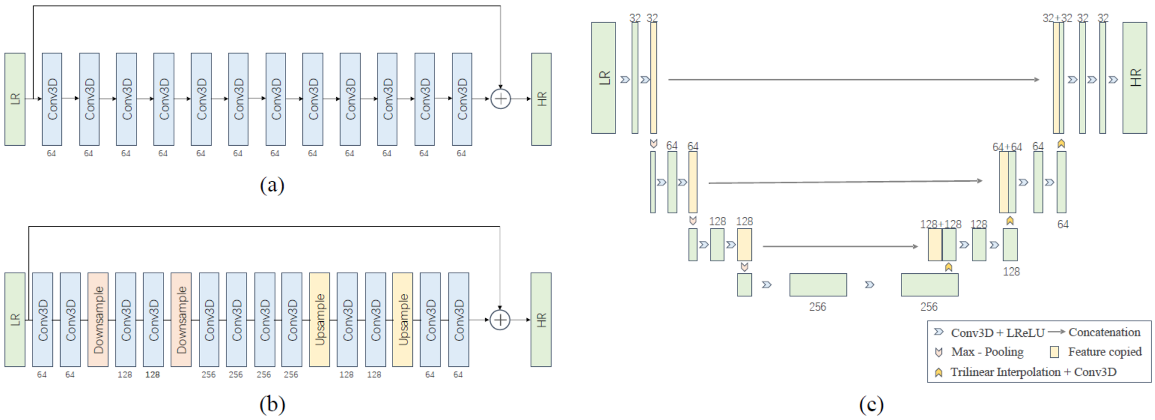


Fig. 1: Architectures of different models used in this study. (a) Plain CNN; (b) AE models; (c) U-Net.

Plain CNN. Fig. 1 (a) illustrates the structure of Plain CNN, which consists of 12 basic building blocks connected in series. Each block consists of a standard 3D convolutional layer with 64 filters of size $3 \times 3 \times 3$ and a Leaky Rectified Linear Unit (Leaky ReLU) with slope of 0.1 as the nonlinear activation function. To avoid resizing effects, we keep the same dimension for all feature maps by setting the stride to 1 for every convolutional layer and adding zero padding before convolution. We abandon the use of Batch Normalisation (BN) layer within the model. According to [5], BN not only occupies too much memory but also discards valuable feature range flexibility in SR. Finally, we apply the global residual learning as suggested by [6] to ease training and prevent the gradient vanishing problem.

Auto Encoder. Given the baseline model, we insert a downsampling layer after every 2 building blocks and its corresponding upsampling layer symmetrically, turning the model into the AE architecture shown in Fig. 1 (b). Each of these layers resizes the feature map in all dimensions by a factor of 2 and the channel number is adjusted to compensate for this effect. We consider 2 options for the downsampling layer: 1) max-pooling with a filter of size $2 \times 2 \times 2$, stride of 2 and dilation of 1; and 2) 3D convolutional layer with a filter of size $3 \times 3 \times 3$, setting stride to 2 and padding to 1. To prevent any checkerboard artifacts generated by the transpose convolution [7] during the upsampling operation, we use trilinear interpolation to resize feature maps followed by the standard convolution.

UNet. Our implementation of the U-Net model is based on [8], which is shown in Fig. 1 (c). We replace all 2D convolutional layers and pooling layers with their corresponding 3D versions without changing their configurations. For the reasons mentioned above, BN layers are not used and all transpose convolution operations are substituted with “trilinear interpolation + convolution” in the decoder pathway. To ensure a relatively fair comparison with AE models, we simplify the U-Net to have a similar model size and depth by 1) setting the channel number in the convolutional layer at the first level to 32 instead of 64; 2) reducing the level of hierarchical feature maps from 5 to 4.

Evaluation Metrics. The performance of each model is evaluated by using: Peak Signal-To-Noise-Ratio (PSNR); Structural Similarity Index (SSIM); and the Root Mean Square Error (RMSE). Both PSNR and RMSE focus on the pixel-level error between the reconstructed volume and the ground truth label while SSIM is more suitable for reflecting the structural correspondences.

2.2 GAN-based method for harmonisation

GAN-based methods have been widely applied in image super-resolution [9], data synthesis [10] and reconstruction [11]. However, the vanilla GAN architecture may suffer from unstable training and collapse mode. Moreover, the vanilla GAN is also suffered from prolonged training and complicated hyper-parameters tuning. In this section we tested the performance of Wasserstein-based Cycle-GAN, Cycle-GAN, Pix2Pix, and Latent diffusion model (LDM) for data harmonisation, using the unpaired brain MR data from TCGA (The Cancer Genome Atlas) dataset. The preliminary results showed that the Wasserstein-based Cycle-GAN approach can handle successfully the data harmonisation task that can both improve the image quality and the performance of computational modules.

GAN [12] architecture is an approach to train a model for image synthesis that is comprised of two submodules: a generator and a discriminator. The generator takes a point from a latent space as input and generates new plausible images from the domain, and the discriminator takes an image as input and predicts whether it is real (from a dataset) or fake (generated). Both models are trained in a game, such that the generator is updated to better fool the discriminator and the discriminator is updated to better detect generated images.

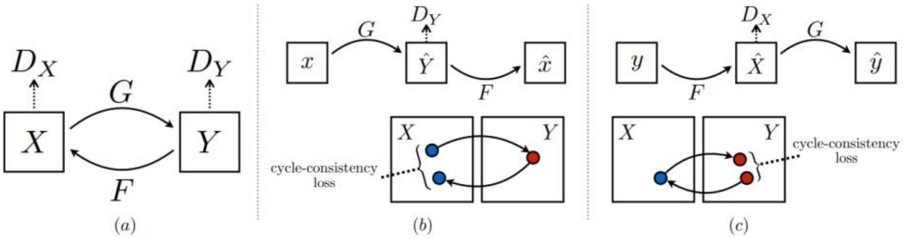


Fig. 2 Illustration of Cycle-GAN [13].

CycleGAN [13] is an extension of the GAN architecture that involves the simultaneous training of two generator models *G* and *F* and two discriminator models D_X and D_Y. It is a technique that involves the automatic training of image-to-image translation models without paired examples. The models are trained in a self-supervised manner using a collection of images from the source (*X*) and target (*Y*) domain that do not need to be related in any way. One generator *G* takes images from the first domain as input and outputs images for the second domain $\hat{Y} = G(x)$, and the other generator *F* takes images from the second domain as input and generates images for the first domain $\hat{X} = F(y)$. Discriminator models are then used to determine how plausible the generated images are and update the generator models accordingly. The CycleGAN uses an additional extension to the architecture called cycle consistency. This is the idea that an image output by the first generator could be used as input to the second generator and the output of the second generator should match the original image.

$$x \approx \hat{x} = F\big(G(x)\big), \tag{1}$$

$$y \approx \hat{y} = G\big(F(y)\big). \tag{2}$$

The reverse is also true: that an output from the second generator can be fed as input to the first generator and the result should match the input to the second generator.

Pix2Pix. Pix2Pix [14] is a conditional Generative Adversarial Network (cGAN) designed for image-to-image translation tasks. Originating from the realm of computer vision research, the model effectively transforms one type of image into another. For instance, it can be trained to convert a sketch into a colourful picture, or a satellite image into a map. The model comprises two primary components: Generator: It takes an image as input and produces a transformed image as its output. Discriminator: It distinguishes between real (ground truth) images and synthetic images produced by the generator.

During the training phase, the generator aims to create images that the discriminator cannot distinguish from real images, while the discriminator tries to get better at telling real from synthesized images. This adversarial process continues until a balance is achieved, often resulting in the generator producing highly realistic images. The "condition" in the cGAN is the input image, ensuring that the output image is not just a random generation but specifically a

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No. 952172.

transformation of the input. This conditioning makes Pix2Pix highly effective for tasks where the input and output images have a clear relationship or correspondence. One of the significant advantages of Pix2Pix is its ability to work on paired data, where every input image in the training dataset has a corresponding desired output. This pair-wise data training approach ensures that the model has a clear guideline on how to perform the transformation for various input scenarios.

Wasserstein-based Cycle-GAN. In this project, we aim to further improve the quality and stability of synthesised data by introducing the Wassertein GAN (WGAN) to CycleGAN scheme. The instability of training procedure of GAN has been reported in [15], including the gradient vanishing issue and mode collapse issue. The Wassertein distance, also known as the Earth-Mover distance, is introduced to the CycleGAN to alleviate unstable training process (gradient vanishing and mode collapse issue) for better synthesis quality. The definition of EM distance is

$$W\left(P_r, P_g\right) = \underset{\gamma \sim \Pi\left(P_r, P_g\right)}{inf} E_{(x,y)\sim\gamma}\left[||x-y||\right], \tag{3}$$

which can be relaxed the Kantorovich-Rubinstein duality as

$$W\left(P_r, P_g\right) = \frac{1}{K} \underset{||f||_{L\leq K}}{sup} E_{x\sim P_\gamma}[f(x)] - E_{x\sim P_g}[f(x)]. \tag{4}$$

and the objective function of discriminator D can be defined as

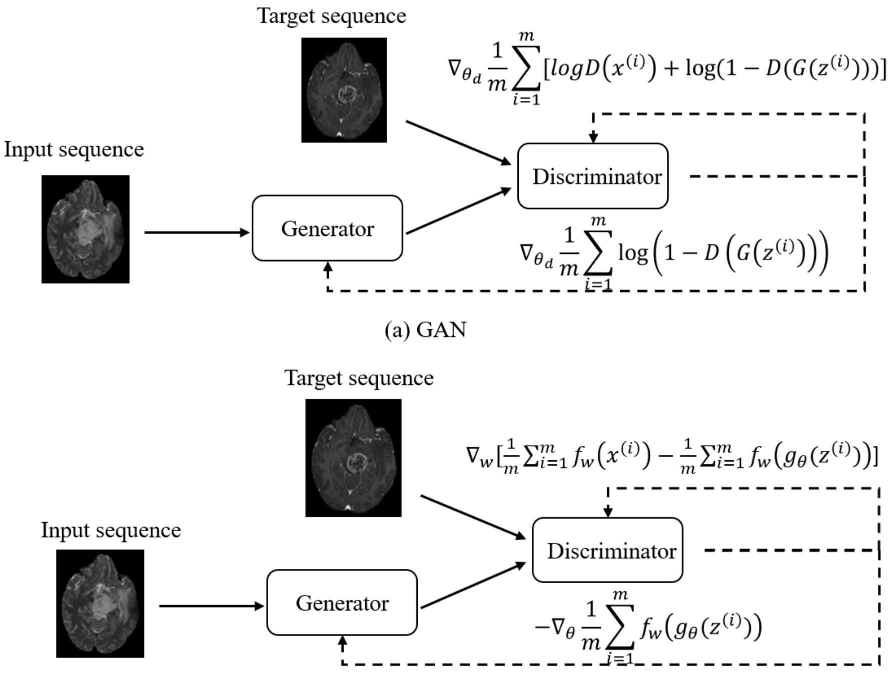
$$L = E_{x\sim P_{data(x)}}\left[D_\beta(x)\right] - E_{x\sim P_{G(x)}}\left[D_\beta(x)\right].$$

Overall, the adversarial loss function of WGAN is given by

$$\min_{D} \max_{D\in L}\{V(D,G)\} = \min_{D} \max_{D\in L}\left\{E_{x\sim P_{data(x)}}\left[D_\beta(x)\right] - E_{Z\sim P_{Z(z)}}\left[D_\beta\left(G_\theta(z)\right)\right]\right\}. \tag{5}$$

where θ represents the parameters in generator G.

The difference between GAN and WGAN is shown as


Target sequence
$\nabla_{\theta_d} \frac{1}{m}\sum_{i=1}^{m}[logD(x^{(i)}) + \log(1 - D(G(z^{(i)})))]$
Input sequence
Discriminator
Generator
$\nabla_{\theta_d} \frac{1}{m}\sum_{i=1}^{m} \log\left(1 - D\left(G(z^{(i)})\right)\right)$

(a) GAN

Target sequence
$\nabla_w[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))]$
Input sequence
Discriminator
Generator
$-\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$

(b) WGAN

Fig. 3. Differences between the GAN and WGAN

The backbone of two Generators in the proposed method is a U-Net based architecture, and that of the two discriminators is based on pixelGAN, which classifies whether a pixel in the synthesised image is real or not. The loss function of applied CycleGAN includes three parts: adversarial loss, cycle consistency loss and identity loss. The Wassertein based adversarial loss is presented as

$$L_{adv} = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) \tag{6}$$

where

$$L_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y\sim p_{data}(y)}[D_Y(y)] - \mathbb{E}_{x\sim p_{data}(x)}\left[D_Y\left(G(x)\right)\right] \tag{7}$$

$$L_{GAN}(F, D_X, Y, X) = \mathbb{E}_{x\sim p_{data}(x)}[D_X(x)] + \mathbb{E}_{y\sim p_{data}(y)}\left[D_X\left(F(Y)\right)\right]. \tag{8}$$

The cycle consistency loss is calculated through

$$L_{cyc}(G, F) = \mathbb{E}_{x\sim p_{data}(x)}\left[\left\|F\left(G(x)\right) - x\right\|_1\right] + \mathbb{E}_{y\sim p_{data}(y)}\left[\left\|G\left(F(x)\right) - y\right\|_1\right] \tag{9}$$

and the identity loss is computed by

$$L_{idt}(G, F) = \mathbb{E}_{x\sim p_{data}(x)}[\|G(x) - x\|_1] + \mathbb{E}_{y\sim p_{data}(y)}[\|F(y) - y\|_1]. \tag{10}$$

2.3 Data harmonisation in downstream tasks

- To evaluate the effectiveness of data harmonisation, we trained a 3D airway extraction algorithm on lung CT data to segment the airway trees. The harmonisation modules were trained to harmonise low-resolution CT scans to high-resolution CT scans. The effectiveness of data harmonisation was assessed by comparing the airway segmentation performance of the segmentation model on harmonised/unharmonised test datasets, respectively. The experiment was trained and tested on a public-available dataset ATM22. The performance was assessed by calculating the branch scores and IoU scores of the prediction.
- In addition, we also evaluate the segmentation performance on harmonised MR images, a model has been trained for each synthetic and original set. The model has been trained several times with different shuffle on train-test split, and the shown results are the mean of the best scores for each training. The set consist of 200 prostate T2 volumes split in 80%-20% train-test shuffled randomly.

3. Experimental results and discussion

3.1 CNN-based approaches for CT data harmonisation.

Imaging Data Source and Pre-processing. We use the AAPM-Mayo Clinic Low-Dose CT Grand Challenge Dataset[1] provided by Mayo Clinic for model training and testing. From the CT scans collected from 140 patients, we use 48 chest CT scans and further split these into 37 training, 5 validation and 6 test volumes. The data for each patient consists of normal-dose CT (NDCT) scans and the corresponding synthetic low-dose CT (LDCT) scans with additional Poisson noise. Only NDCT are used as ground truth HR data in the experiments. All volumes within the dataset contain an uneven number of slices of size 512×512 with 1.5mm thickness. To enable a feasible training time and solve the memory limitation, we pre-downsampled each slice from 512×512 to 256×256 using bilinear interpolation. Our LR data are degraded from ground truth HR data by the following steps: 1) Truncate the leading and trailing slices evenly so that the dimension of each volume can perfectly fit the non-overlapping patch extraction algorithm; 2) Downsample the volume in the axial direction by removing slices at a constant interval; 3) Clip all HU values into the range [-1024, 1476] and normalize them to [0,1]; and 4) Upsample the volume to its original dimension either by trilinear interpolation or by inserting the same slice at the previous position.

Statistical analysis. We use Shapiro-Wilk test to check the normality of the difference in the image quality between each AE model and Plain CNN and paired Student's t-test to determine whether there is statistically significant evidence to support this difference. We use Wilcoxon

[1] https://www.aapm.org/grandchallenge/lowdosect/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952172.

signed-rank test instead when the normality of the sample cannot be satisfied. We set the significance level to 0.05 in all statistical tests.

In Table 1, we compare the quantitative performance of all AE models with Plain CNN under $\times 2$, $\times 4$, and $\times 8$ scaling factors. Next, we present the total number of parameters and the average inference time used by one volume for every model in Table 2. Finally, we show visual comparisons of different models in Fig. 1.

Table 1: Quantitative comparisons of different models. Best results are shown in Bold. * indicates statistically significant evidence to support the difference with Plain CNN. Top: Mean (STD) using trilinear interpolation in LR generation; Bottom: Mean (STD) using same insertion in LR generation.

Scale	Methods	PSNR	SSIM	RMSE
×2	AE-Maxpool	35.14 (2.50)*	0.9649 (0.0062)*	4.64 (1.20)*
	AE-Conv	35.38 (2.53)*	0.9641 (0.0072)*	4.52 (1.21)*
	U-Net	39.40 (1.23)*	0.9795 (0.0040)*	2.76 (0.40)*
	Plain CNN	**43.52 (0.95)**	**0.9839 (0.0049)**	**1.71 (0.19)**
×4	AE-Maxpool	26.89 (2.51)*	0.8711 (0.0213)*	11.99 (3.12)*
	AE-Conv	25.49 (1.29)*	0.8741 (0.0207)*	13.71 (2.1)*
	U-Net	29.31 (1.73)*	0.9131 (0.0112)*	8.92 (1.94)*
	Plain CNN	**34.51 (0.65)**	**0.9345 (0.0130)**	**4.81 (0.35)**
×8	AE-Maxpool	24.20 (1.86)*	0.7890 (0.0320)*	16.08 (3.22)*
	AE-Conv	23.34 (1.59)*	0.7838 (0.0306)*	17.63 (2.88)*
	U-Net	30.24 (1.20)*	**0.8663 (0.0220)**	7.92 (1.19)*
	Plain CNN	**31.03 (1.25)**	0.8644 (0.0226)	**7.23 (0.96)**
×2	AE-Maxpool	33.66 (2.16)*	0.9539 (0.0094)*	5.44 (1.14)*
	AE-Conv	32.25 (3.23)*	0.9468 (0.0108)*	6.63 (2.13)*
	U-Net	37.52 (1.43)*	0.9770 (0.0035)*	3.44 (0.63)*
	Plain CNN	**43.16 (1.02)**	**0.9837 (0.0048)**	**1.79 (0.21)**
×4	AE-Maxpool	23.52 (1.06)*	0.8434 (0.0260)*	17.12 (1.96)*
	AE-Conv	25.70 (2.54)*	0.8420 (0.0290)*	13.77 (3.57)*
	U-Net	33.45 (0.97)*	0.9287 (0.0152)*	5.45 (0.61)*
	Plain CNN	**36.54 (0.69)**	**0.9422 (0.0134)**	**3.81 (0.30)**
×8	AE-Maxpool	22.36 (1.48)*	0.7562 (0.0302)*	19.70 (2.98)*
	AE-Conv	20.86 (1.42)*	0.7400 (0.0368)*	23.38 (3.55)*
	U-net	27.21 (2.14)*	0.8538 (0.0198)*	11.50 (3.36)*
	Plain CNN	**31.35 (1.34)**	**0.8757 (0.0197)**	**6.98 (0.99)**

Table 2: Computational comparisons of different models

Methods	#Parameter (M)	Inference Time (s)
AE-Maxpool	6.88	8.42
AE-Conv	7.44	5.25
U-Net	5.30	8.51
Plain CNN	1.11	27.72

The results in Table 1 show that AE is unsuitable for 3D CT SISR. It can be seen that there is an obvious performance drop for all AE models compared with Plain CNN in almost all cases. At the same time, there also exists statistically significant evidence to support this performance drop in almost every comparison experiment. We contend that the main reason for the performance drop is the resizing operation within AE since this is the only architectural difference between AE models and the baseline model. Surprisingly, those results also reveals the fact that skip connections, which are designed to benefit U-Net by increasing the high-resolution feature re-usability, cannot fully compensate for the aforementioned performance gap. From Table 2, AE models have a lower computational cost than Plain CNN reflected by the reduced average inference time, but at the cost of a larger model size.

We show comparisons between AE models and Plain CNN visually in Fig. 4. These results again argue that AE, including U-Net, is not suitable basis for 3D SISR tasks. It is clearly shown that AE models generate noticeable artifacts around edges and significant deviations in regions with abundant textures. This shows the loss of diagnostically important information due to resizing. In contrast, Plain CNN can recover comparatively more natural textures and smoothing edges and produce results that are almost indistinguishable from the ground truth HR data.

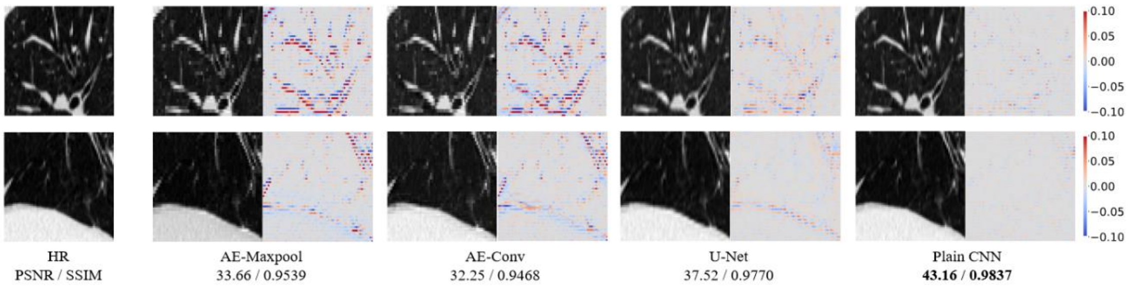


Fig. 4: Visual comparisons of different models using same insertion in LR generation under ×2 scaling factor.

The results showed that AE models are unsuitable due to the information loss in feature resizing operations. Therefore, the best practice would be plain convolutional neural network (CNN) without any down sampling operations. The experiments were carefully designed with adjusted model architectures for a fair comparison. The models were evaluated on a publicly available CT lung dataset, and the findings concluded that although AE models can achieve faster inference, they do so at the cost of inferior performance compared to the baseline CNN.

3.2 CNN-based approaches for MR data harmonisation.

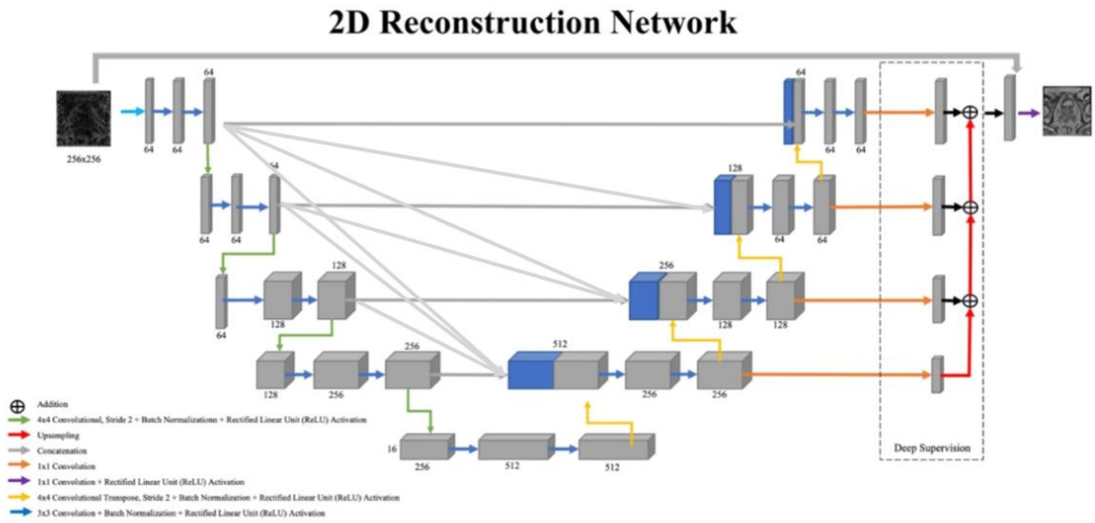

2D Reconstruction Network
256x256
64
128
256
512
16
Deep Supervision
⊕ Addition
4x4 Convolutional, Stride 2 + Batch Normalizationn + Rectified Linear Unit (ReLU) Activation
Upsampling
Concatenation
1x1 Convolution
1x1 Convolution + Rectified Linear Unit (ReLU) Activation
4x4 Convolutional Transpose, Stride 2 + Batch Normalization + Rectified Linear Unit (ReLU) Activation
3x3 Convolution + Batch Normalization + Rectified Linear Unit (ReLU) Activation

Fig. 5 Network architecture for prostate MR harmonisation

For this case, a variation of UNET known as UNET3+ has been used (Fig. 5), which has the particularity of a total connection between all the encoder blocks with all the decoder blocks, which is commonly known as skip connections, but in a general way, to which a deep supervision block has been added and another extra connection between the input, prior to the first convolutional layer, and the last concatenation layer after deep supervision.

As of visual results, a subset of highly different and variated image sources has been selected to harmonize, to see how each of the specific contrasts are brought to the same tonality.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952172.

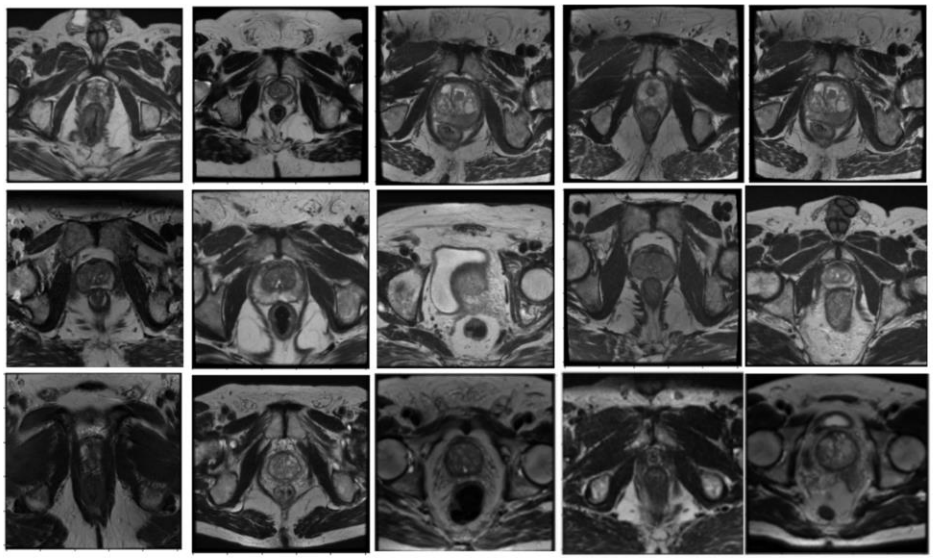


Fig 6. Subset of a wide variated input sources for prostate T2 samples

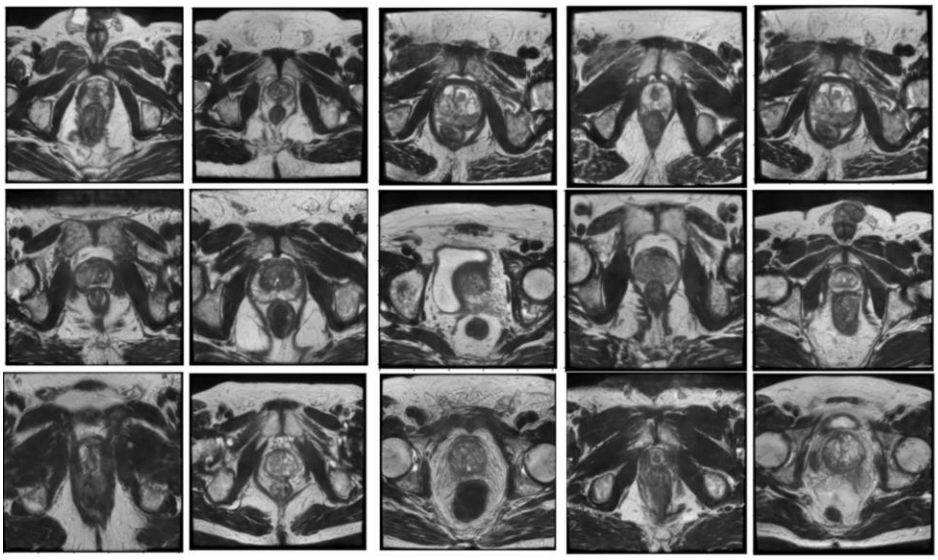


Fig 7. Same subset after harmonization algorithm execution

To evaluate the harmonization of the radiomic characteristics, the following process has been followed.

- Extraction of radiomic features and qualitative evaluation in box plots.
- Group radiomic features into two corresponding subgroups with the original split that has been taken to generate the training set for the harmonization model (repetition time > 6000, echo time < 133).
- Perform Levene test for each feature column iterating over the split of the previous step between the same groups. It tests the null hypothesis that the population variances are equal (called homogeneity of variance or homoscedasticity). If the resulting p-value of Levene's test is less than some significance level (typically 0.05), the obtained differences in sample variances are unlikely to have occurred based on random sampling from a population with equal variances. Thus, the null hypothesis of equal variances is rejected and it is concluded that there is a difference between the variances in the population.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952172.

- Based on results in Levene tests, T-Test evaluation takes place for each feature comparing the two variable groups based on split of step 1 and P-values from step 2, if Levene's test shows equal variance between groups T-Test with equal variance assumption is performed, otherwise it is assumed difference variance.
- Cohens_D factor is used as size effect, is calculated and thresholded accordingly each time to have a better weight depending on population differences between the two features compared, it is used a medium size effect
- Samples that fall before 0.05 threshold in T-test's P-value and are above 0.5 in cohens_D threshold are labeled as feature with significative differences over both subgroups, otherwise are counted as non-significative differences for this feature.

Table 3. Harmonisation results evaluated by radiomic features.

	Number of features without significative differences	Number of features with significative differences
Before harmonization set#1	81	**11**
After harmonization set#1	91	**1**
Before harmonization set#2	55	**39**
After harmonization set#2	87	**5**

The results visually show a systematic reduction of variance and outliers in most of the features extracted per image (Table 3 and Fig. 8).

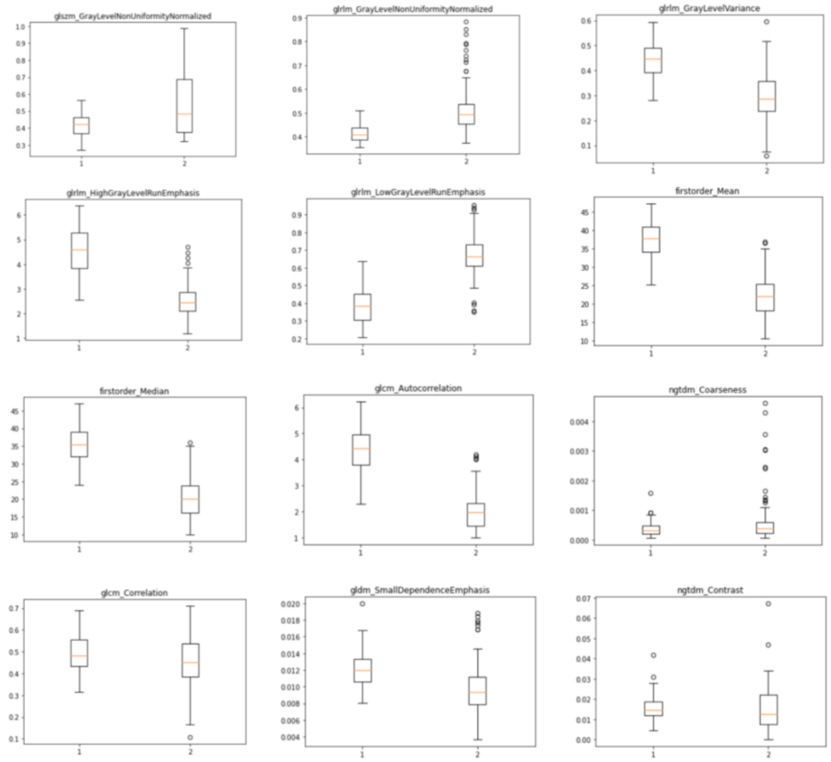


Fig 8. Plot of a subset of radiomic features after and before harmonization

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952172.

3.3 Generative-based approaches for data harmonisation

Imaging Data Source and Pre-processing. The experiment was designed based on UCSF PDGM. The UCSF-PDGM dataset includes 501 subjects with histopathologically-proven diffuse gliomas who were imaged with a standardized 3 Tesla preoperative brain tumor MRI protocol featuring predominantly 3D imaging, as well as advanced diffusion and perfusion imaging techniques. We randomly split the dataset into train, valid and test, following the ratio of 7:2:1.

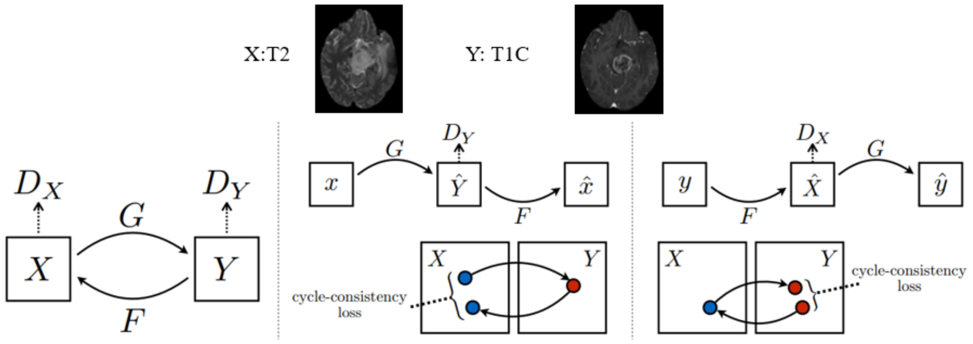
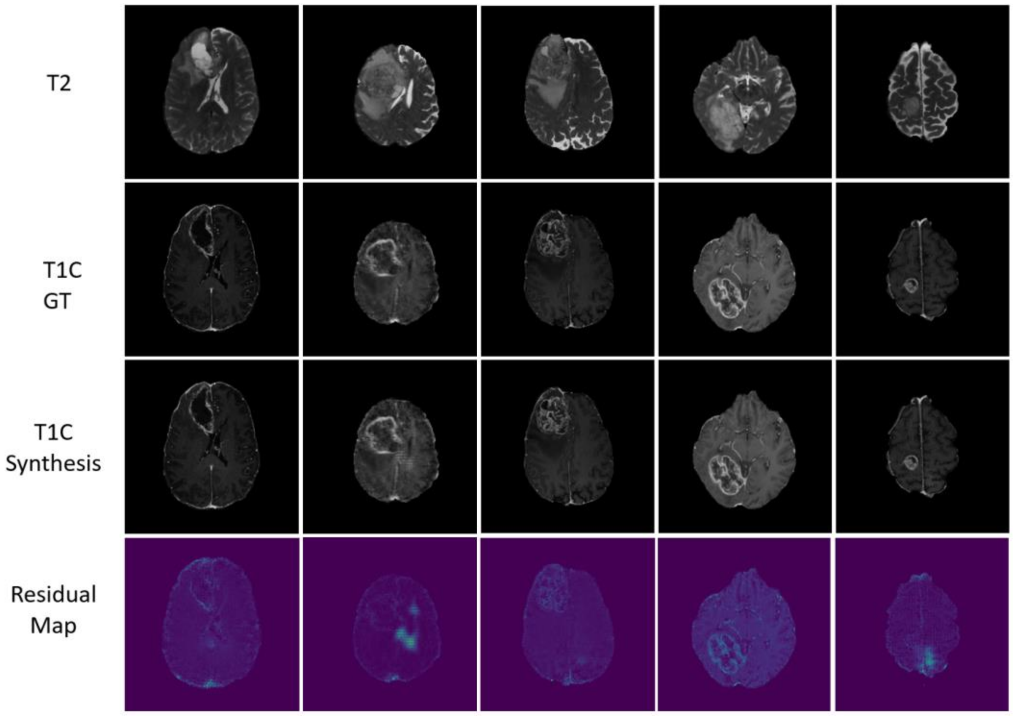


Fig. 9. Basic flow of the experiment. The synthesis has two directions, including T2 to T1C and its reverse. In this study we mainly focus on the synthesis of T2 to T1C due to its clinical practice value.

Quantity Results. We evaluated the WCycle-GAN, Cycle-GAN, Pix2Pix, and LDM, by calculating the l1 error and psnr between the ground truth T1CE and synthesised T1CE. The performance was assessed on test set. As shown in Table 4, the WCycle-GAN achieves better performance with lower L1 error and higher PSNR.

Table. 4. Quantity results of Wcycle-GAN

Metrics	WCycle-GAN	Cycle-GAN	Pix2Pix	LDM
PSNR	**43.69**	35.19	34.67	38.53
L1	**0.55**	0.97	1.57	0.81



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952172.

Fig. 10. Visualization results of Brain MR harmonisation by WCycle-GAN.

The Visualization results (Fig. 10) also show the satisfied results of WCycle-GAN. Interestingly, it can be found that there exist some checkerboard artifacts and some variances in the tumour region, which can be further improved in the future. Overall, the proposed WCycle-GAN achieves better performance than Pix2Pix model and conventional CycleGAN approach, which can be better used in MR sequence harmonisation.

3.4 Data harmonisation in Downstream tasks

3.4.1 Data harmonisation in airway segmentation

Table. 5. Segmentation performance before and after harmonisation

	IoU	Branch length	Branch ratio
Without harmonisation	0.7836	0.7048	0.5840
WCycle-GAN Harmonised	0.8345	0.7607	0.6673
Plain-CNN Harmonised	0.8608	0.8212	0.7353

To investigate the best model for data harmonisation on lung cancer CT scans, we compared the performance of a downstream application (airway segmentation) on data before and after harmonisation, as shown in Table 5 and Fig. 11.

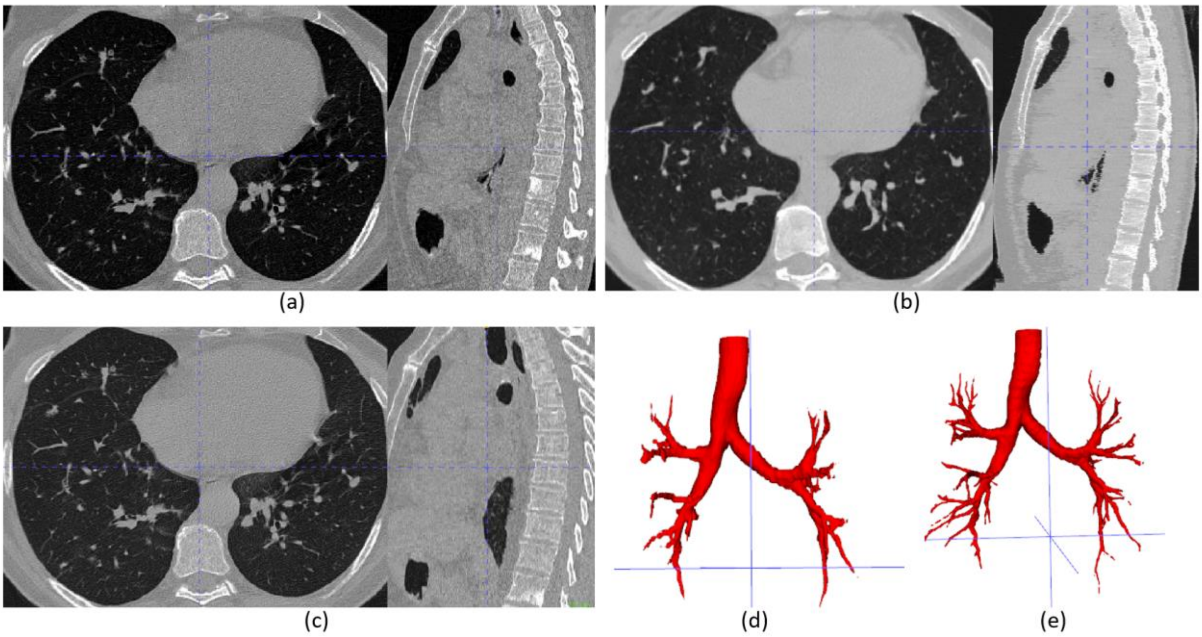


Fig. 11. (a) Original lung CT scans (b) harmonised by WCycle-GAN (c) harmonised by 3D Plain-CNN (d) The airway extracted on raw CT scan and (e) WCycle-GAN harmonised scan.

It can be seen that both Plain-CNN and WCycle-GAN improved the segmentation performance, with 5-6% gain in IoU, 6-12% in detected branch length, and 8-15% in detected branch ratio. Additionally, WCycle-GAN are more likely to remove noises in the original CT scan, leading to an unsmooth harmonisation effect as in Fig. 7(b). This was mainly because of the limitation of 2D generative models. Compared with the WCycle-GAN, 3D Plain-CNN achieved more stable and real results. To better understand the differences before and after harmonisation, the extracted airway trees were visualized in Fig. 7 (e-d). This illustrates the effectiveness of the proposed WCycle-GAN to harmonise low-resolution CT scans into high-resolution ones.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952172.

3.4.2 Data harmonisation in prostate cancer segmentation

The following Table. 6 shows final Dice Score mean on the 5 models trained for comparison.

Table 6. Segmentation performance before and after harmonisation

	Mean Dice Score
Original Images	0.86
Harmonised Images	0.91

Both scores show a high dice score and perform in general pretty well, qualitative differences are mostly seen in those cases where the repetition time is the furthest in comparison to those who have been harmonized around (repetition time > 6000) and show a lower contrast and intensity as well mean.

Below are some examples of 2D slices where the harmonized model performs visually better than the original one, it is important to highlight that as the dice score shows, the performance is very similar for both models and it only shows differences in some specific slices inside a case, in those cases the harmonized model is usually giving better results.

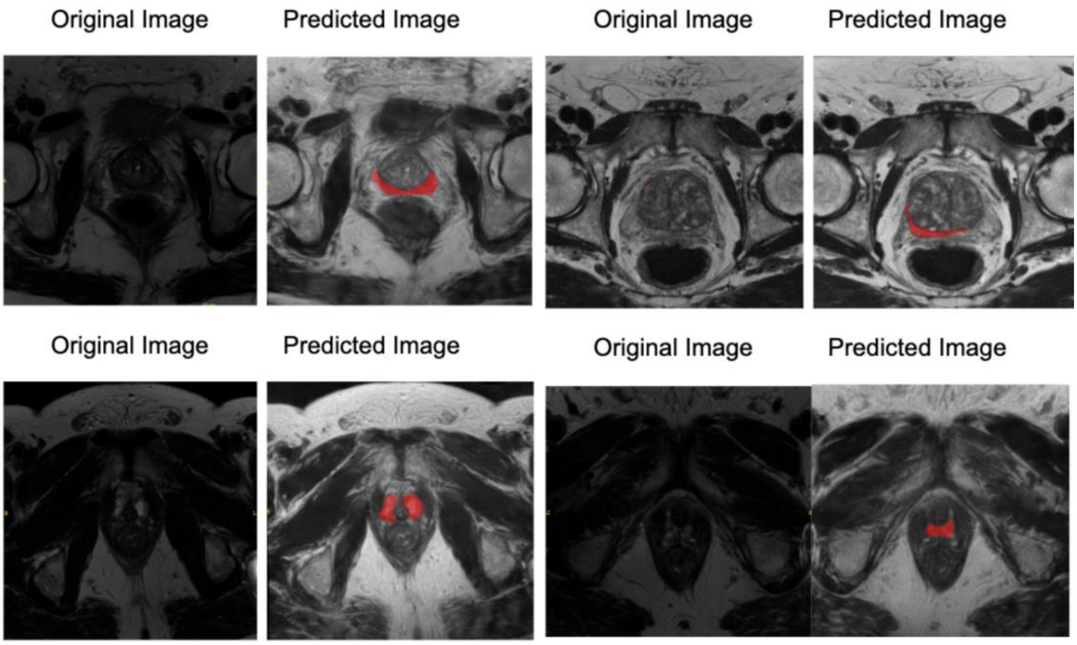


Fig. 12 Visualized segmentation results of original images and harmonised images

To better isolate the positive results of the harmonization method in the prostate, a more specific experiment has been carried out regarding the control of outliers, separating those images with the lowest contrast, either due to an extremely low repetition time together with an extremely high echo time, such as the cases shown above, such as cases where a sonar artifact has crept in, leaving the rest of the contrasts badly damaged.

To validate the operation in these cases, an algorithm has been trained excepting these outlier images, to later test the same model trained with the outlier samples before and after being harmonized, a visual example of the above statement is find below.

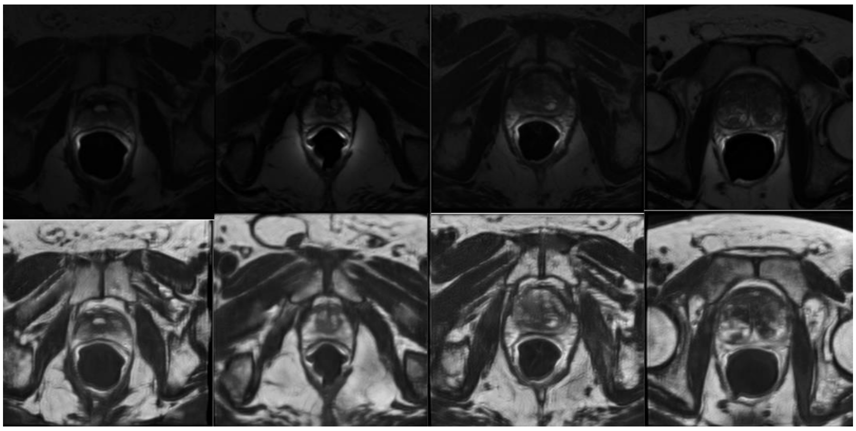


Fig 13. Harmonization results (bottom row) with heavy artifact samples

Dice score results under these conditions are found in table below.

Table 7. Dice Score for the performance on the outliers before and after harmonization

	Original	After harmonization
model #1	0.12	**0.60**
model #2	0.09	**0.58**
model #3	0.31	**0.63**
model #4	0.24	**0.66**

4. Conclusion

The results showed the feasibility of using GAN-based method for data harmonisation. The image quality such as resolution can be significantly improved through visualization. Meanwhile, the performance of computational modules was improved which indicated the effectiveness of harmonisation strategy.

Due to the GPU memory, current WCycle-GAN based method could only produce good performance on 2D images and may introduce artifacts for 3D images when stacking all harmonised 2D image slices into a 3D one. We will continue to develop novel harmonisation modules in further studies.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No. 952172.

5. References

[1] H. Mirzaalian *et al.*, "Inter-site and inter-scanner diffusion MRI data harmonization," *NeuroImage,* vol. 135, pp. 311-323, 2016.

[2] T. Zhu *et al.*, "Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study," *Neuroimage,* vol. 56, no. 3, pp. 1398-1411, 2011.

[3] J. Jovicich *et al.*, "Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects," *Neuroimage,* vol. 101, pp. 390-403, 2014.

[4] P. Leo, G. Lee, N. N. Shih, R. Elliott, M. D. Feldman, and A. Madabhushi, "Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images," *Journal of medical imaging,* vol. 3, no. 4, p. 047502, 2016.

[5] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136-144.

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

[7] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill,* vol. 1, no. 10, p. e3, 2016.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 2015, pp. 234-241: Springer.

[9] E. C. de Farias, C. Di Noia, C. Han, E. Sala, M. Castelli, and L. Rundo, "Impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features," *Scientific reports,* vol. 11, no. 1, pp. 1-12, 2021.

[10] D. Nie *et al.*, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering,* vol. 65, no. 12, pp. 2720-2730, 2018.

[11] P. Zhang, F. Wang, W. Xu, and Y. Li, "Multi-channel generative adversarial network for parallel magnetic resonance image reconstruction in k-space," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 180-188: Springer.

[12] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems,* vol. 27, 2014.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223-2232.

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

[15] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *arXiv preprint arXiv:1701.04862,* 2017.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952172.