Guiding Unsupervised MRI-to-CT synthesis using Content and style Representation by an Enhanced Perceptual synthesis (CREPs) loss*

Cédric Hémon[1 [0009−0003−6669−5108]], Valentin Boussot[1 [0009−0003−2465−5458]], and Blanche Texier[1 [0009−0000−9961−4958]]

Univ Rennes, CLCC Eugène Marquis, INSERM, LTSI - UMR 1099, F-35000 Rennes, France
https://ltsi.univ-rennes.fr/

Abstract. The goal of this research was to propose an unsupervised learning technique for producing synthetic CT (sCT) images from MRI data. For model training, a dataset consisting of 180 pairs of brain CT and MR scans, as well as 180 pairs of pelvis scans was used.
The devised methodology incorporates a 3D conditional Generative Adversarial Network (cGAN) training in an unsupervised way. To tackle challenges associated with unsupervised learning convergence, a novel ConvNext-based perceptual loss (CREPs loss) was developed to guide in the 3D cGAN-based MR-to-CT generation process.

Keywords: synthetic CT · MRI · Generation · Perceptual Loss

1 Introduction

While computed tomography (CT) is commonly used for radiation therapy (RT) planning, magnetic resonance imaging (MRI) offers higher contrast [2] and enables more accurate target volume visualization and delineation [4]. Combine these complementary imaging modalities requires the fusion of MRI and CT images. This fusion process involves a deformable registration of these 2 images. However, this procedure introduces uncertainties and registration errors, as highlighted in prior studies [12][3], with deviations of up to 2mm observed in prostate.
With the increasing use of MR-linac devices (combining MRI with a linear accelerator), the MR-only RT workflows have grown attention due to their potential to eliminate the CT acquisition and so the need for multimodal registration.
However, MRI lacks the necessary electron density information essential for accurate dose calculation [10]. This limitation has led to the proposal to generate synthetic CTs (sCTs) from MRI data, as proposed in recent studies by Boulanger et al. [1] and Spadea and Maspero et al. [11].

* Supported by CominLabs CEMMTAUR 2022 and by a PhD scholarship Grant from Elekta AB.

Using labeled data within supervised generation, which involves training from source-target image pairs, requires precise multimodal registration. Nevertheless, achieving accurate multimodal registration can be challenging, often resulting in limited accuracy such in pelvis region.
Furthermore, achieving precise registration is crucial for the efficacy of generation based on supervised learning, a challenge exacerbated by the inherent impossibility of acquiring MRI and CT scans under perfectly identical anatomical conditions [3]. Frequently, the time gap between these two acquisitions leads to anatomical variations such as in the pelvic region compared to the brain region, primarily attributed to differences in rectal gas and bladder filling. This inherent uncertainty highlights the need for unsupervised learning, which leverages unlabeled data to extract valuable information and patterns, providing a promising solution to address this issue.
The objective of this study was therefore to generate sCT image from MR by a 3D cGAN in an unsupervised learning context based on a new definition of the perceptual loss function [7] to guide the generation process.

2 Material and methods

2.1 Material

A model was generated per anatomical location (pelvis and brain), with one trained using 180 pairs of brain CT and MR scans, and the other using 180 pairs of pelvis CT and MR scans. These MR/CT pairs of images acquired from three distinct medical centers for brain and two for pelvis, each center employing different CT and MR imaging systems.

2.2 Method

Pre-processing: For each individual patient, the CT and MR images underwent resampling to achieve an uniform spacing of 1x1x2.5 mm^3 for pelvis and 1x1x1 mm^3 for brain through B-Spline interpolation. The CT intensity range was normalized within the range [-1000, 3000], subsequently scaled down by a factor of 1000 to narrow the value range to [-1, 3] prior to the training process. The MR intensity range was scaled down by a factor of 1000 to narrow the value range to [0, 2] prior to the training process.
The objective of this preprocessing step is to normalize the data and enhance training stability.

cGAN: The proposed algorithm was a customized unsupervised 3D cGAN (with a 6-block ResNet+PatchGAN) [5]. Using a customized 3D generator, our approach involves a 6-block residual network positioned between a single downsampling block and an individual upsampling block. Notably, no biases are incorporated throughout the network architecture. Following each convolutional

operation, batch normalization is applied, subsequently followed by the activation function. Specifically, a LeakyReLu activation with a fixed slope of 0.2 is employed for the ResNet-Block, while a ReLu activation is used for other blocks, except for the final convolution, where no activation function is used. The discriminator is a PatchGAN [6] network with size 70*70*70. Following each convolution, batch normalization is applied, succeeded by the LeakyReLU activation function with a consistent slope of 0.2.

Despite the fact that the discriminator helps to adjust the generation, it happens that the learning suffers from non-convergence especially in unsupervised training [9]. To ensure constrained output, a novel cGAN-based approach was introduced by incorporating a new perceptual loss function between sCT and CT, referred to as Content and style Representation for Enhanced Perceptual synthesis (CREPs) loss. The use of CREPs loss improves the robustness of the synthesis process of sCT from MRI in unsupervised learning. The perceptual loss (Figure 1) inspired by human visual perception was based on the idea of simulating independence between style and content of an image. Perceptual loss [7] was originally based on the VGG-16 network and is the most common and widely used in the literature in image-to-image translation.

In order to impose constraints on the generator output, a style loss between the sCT generated from MRI and a CT image from the learning cohort was incorporated. The study exclusively used the style loss (without content term) since there are discernible differences in content between the MR and CT modalities, highlighting the importance of capturing and preserving the stylistic characteristics during sCT generation. This loss was based on the ConvNext-Tiny network [8] which has fewer parameters (than VGG-16 based), allowing us to reduce the computational cost. The layers were determined empirically and defined as follows:

$$\text{Gram}_j^{\phi}(I) = \frac{\psi\psi^T}{C_j * H_j * W * j} \tag{1}$$

$$\psi \text{ being the flattened matrix } \phi_j(I) \text{ of size: } C_j * H_j W_j \tag{2}$$

$$l_{style}^{\phi,j}(sCT, CT) = ||\text{Gram}_j(sCT) - \text{Gram}_j(CT)||_1 \tag{3}$$

The use of style in feature maps differs from their content; instead of directly employing the feature maps, the 1-norm ($||...||_1$) between the Gram matrix Gram_j^{ϕ} of the output $\hat{I}_1$ (sCT) and the target I_2 (CT) were computed. The gram matrix (Equation.1) captures information about characteristics that tends to activate together. To calculate the perceptual loss, images are normalized with the same mean and standard deviation as the ImageNet pre-trained ConvNext. The medical input images, initially single-channel, are replicated across all three channels (RGB), aligning them with the network's natural image format used for training. Subsequently, these images are resampled to dimensions of 224×224, ensuring compatibility with the size of the training images. As part of the training process, the magnitudes of different loss functions are adjusted to match with those of conventional loss functions, such as Binary Cross Entropy (BCE) used for both the generator and discriminator.

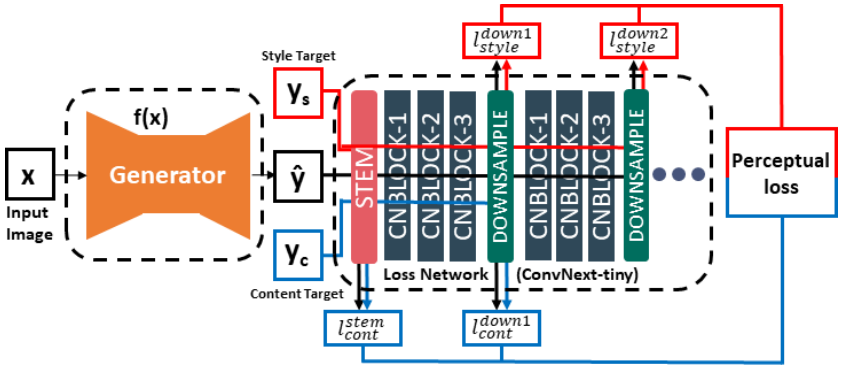

Fig. 1. ConvNext-based perceptual loss definition diagram. The new perceptual loss is the weighted sum of the content (in blue) and style (in red) loss functions defined from different layers of the pre-trained ConvNext-tiny network [8].

The sum operation contributes to the reduction of the batch of the Style loss, whereas the mean operation is employed to reduce the Content loss. In our case, the style reconstruction loss is weighted by 20 for ConvNext-based perceptual loss (Eq.5).

$$l^{\phi,j}_{perceptual}(\text{sCT}, \text{CT}) = l^{\phi,j}_{style}(\text{sCT}, \text{CT}) * 20 \tag{4}$$

The summation of the perceptual loss with the binary cross-entropy of the generator is used for training the generator. The generator complete loss function is defined as follows:

$$l_{generator}(sCT, CT) = l^{\phi,j}_{perceptual}(sCT, CT) + BCE(D(sCT), 1) \tag{5}$$

Training: The deep learning method was implemented in Python3.8 using Pytorch 1.12 with CUDA 11.3. The model was trained and tested on an NVIDIA RTX A6000 with 49 GB of VRAM. The model was trained for 200 epochs with a batch size of 8 without data augmentation. The model was trained using the AdamW optimizer with a fixed learning rate of 0.0001 and a weight decay of 0.001.
The model processes 3D patch from the entire image, with size 32x224x224 for the pelvic region and 66x168x168 for the brain, employing a stride shape equal to half of the patch.

Inference: During the inference process for sCT generation from MRI, the MRI undergoes a reshaping procedure to align its dimensions with a multiple of the patch size. This alignment is achieved by padding the original MRI image with zero values. Subsequently, the MRI is partitioned into distinct patches of the specified training size. This partitioning procedure is designed to include an overlap ranging from a minimum of 2 patches to a maximum of 9 for brain regions, and 8 for the pelvic region. This deliberate overlap is implemented to

enhance the subsequent reconstruction process.
The final value of sCT corresponds to the median value obtained from the diverse predictions generated by multiple patches. After experimenting with several aggregation techniques, it was determined that a straightforward median approach yielded the most optimal outcomes.
In instances where the length of the series is even, the series is extended by appending the value 3000. This modification aims to choose the highest value that resides in close proximity to the center of the series.
After obtaining the sCT and cropped it to its original dimensions, any intensity values beyond the range of [-1024, 3000] are adjusted to the nearest boundary of the interval.

References

1. Boulanger, M., Nunes, J.C., Chourak, H., Largent, A., Tahri, S., Acosta, O., De Crevoisier, R., Lafond, C., Barateau, A.: Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review. Physica Medica **89**, 265–281 (Sep 2021). https://doi.org/10.1016/j.ejmp.2021.07.027
2. Dirix, P., Haustermans, K., Vandecaveye, V.: The value of magnetic resonance imaging for radiotherapy planning **24**(3), 151–159 (2014). https://doi.org/http://dx.doi.org/10.1016/j.semradonc.2014.02.003
3. Florkow, M.C., Zijlstra, F., Kerkmeijer, L.G., Maspero, M., van den Berg, C.A., van Stralen, M., Seevinck, P.R.: The impact of mri-ct registration errors on deep learning-based synthetic ct generation **10949**, 831–837 (2019)
4. Gao, Z., Wilkins, D., Eapen, L., Morash, C., Wassef, Y., Gerig, L.: A study of prostate delineation referenced against a gold standard created from the visible human data. Radiotherapy and oncology **85**(2), 239–246 (2007). https://doi.org/https://doi.org/doi.org/10.1016/j.radonc.2007.08.001
5. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B.: Generative Adversarial Networks (Jun 2014)
6. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-Image Translation with Conditional Adversarial Networks (Nov 2018), http://arxiv.org/abs/1611.07004, arXiv:1611.07004 [cs]
7. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution (Mar 2016), arXiv:1603.08155 [cs]
8. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s (Mar 2022), http://arxiv.org/abs/2201.03545, arXiv:2201.03545 [cs]
9. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
10. Seco, J., Evans, P.: Assessing the effect of electron density in photon dose calculations. Medical physics **33**(2), 540–552 (2006). https://doi.org/doi/pdf/10.1118/1.2161407
11. Spadea, M.F., Maspero, M., Zaffino, P., Seco, J.: Deep learning based synthetic-ct generation in radiotherapy and pet: a review. Medical physics **48**(11), 6537–6566 (2021)
12. Ulin, K., Urie, M.M., Cherlow, J.M.: Results of a multi-institutional benchmark test for cranial ct/mr image registration. International Jour-

nal of Radiation Oncology* Biology* Physics **77**(5), 1584–1589 (2010). https://doi.org/doi:10.1016/j.ijrobp.2009.10.017