Guiding Unsupervised CBCT-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 1, 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 CBCT data. For model training, a dataset consisting of 180 pairs of brain CT and CBCT scans, as well as 180 pairs of pelvis scans was used.
The devised methodology incorporates a 2D conditional Generative Adversarial Network (cGAN) training under unsupervised conditions. To tackle challenges associated with unsupervised learning convergence, a novel ConvNext-based perceptual loss (CREPs loss) was developed to provide guidance in the CBCT-to-CT generation process.

Keywords: synthetic CT · CBCT · Generation · Perceptual Loss

1 Introduction

CBCT scans are widely used to obtain daily anatomical information to correctly position the patient for irradiation. They are also useful for monitoring morphological changes and estimating the patient delivered dose. However, the presence of a large number of HU inconsistencies and artifacts hinders CBCT-based dose planning.
In the literature [5], supervised learning from paired data is well studied for sCT generation due to its convergence speed and stability. Currently, the performance in 2D exceeds that in 3D. However, supervised learning using paired data suffers from overfitting and adds an additional bias due to CT/CBCT registration inaccuracies.
Unsupervised learning from unpaired data has better robustness but less accurate generation. However, the generation performance does not depend on the accuracy of the registration.
The purpose of this study was therefore to generate sCT image from CBCT

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

by a 2D cGAN in an unsupervised learning context using the paired unregistered images. We proposed a new definition of the perceptual loss based on the ConvNext network [3] to to stabilize and refine the generation process.

2 Material and methods

2.1 Material

For each anatomical location, a set of two models was generated, with one trained using 180 pairs of brain CT and CBCT scans, and the other using 180 pairs of pelvis CT and CBCT scans. These pairs of data originated from three distinct medical centers, each employing diverse CT and CBCT imaging systems.

2.2 Method

Pre-processing For each individual patient, the CT and CBCT images underwent resampling to achieve a uniform spacing of 1x1x2.5 mm^3 for pelvis and 1x1x1 mm^3 for brain through B-Spline interpolation. CBCT intensities were adjusted using histogram matching using their corresponding reference CT scans. The 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 objective of this preprocessing step is to normalize the data and enhance training stability.

Method The proposed algorithm was a customized unsupervised 2D GAN (with a 6-block ResNet+PatchGAN)[1].
Using a customized 2D 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 [2] network with size 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. To guide the generation, the sCT generation problem was considered as a style transfer problem. The perceptual loss function was based on the idea of being able to simulate independence between the style and content of an image. Assuming this property, the output image (sCT) was reconstructed with the content of an image (CBCT) by applying the style of an image (CT). Due

to the limitations posed by the VGG network, a novel perceptual loss was formulated using the ConvNext-tiny architecture referred to as Content and style Representation for Enhanced Perceptual synthesis (CREPs) loss.

CREPs loss The CREPs loss was derived from the perceptual loss introduced by Johnson et al. [3]. The ConvNext-Tiny network [4] which has fewer parameters (than VGG-16) was used to decrease the computational cost.
In the context of medical images, the optimization of feature map selection for computing style and content loss of the CREP loss has been performed empirically on medical image generation and registration problems. Unlike the VGG-based approach, the novel style loss computation is performed across multiple layers of the network.
The perceptual loss is complementary to the discriminator but compared to the latter it is based on a pre-trained network. The output image $\hat{I}_1$ (sCT) can be reconstructed with the content of an image I_1 (CBCT) by applying the style of an image I_2 (CT). This image $\hat{I}_1$ is reconstructed from two types of loss functions:

- Content loss function: The feature maps of the two images extracted at different layers of the network are compared. They are compared by the mean absolute error (Equation.1).

$$l_{cont}^{\phi,j}(\hat{I}_1, I_1) = ||\phi_j(\hat{I}_1) - \phi_j(I_1)||_1 \tag{1}$$

where $\phi_j(I_1)$ is the activation at the jth layer for the input I_1

- Style loss function: For style, the feature maps are not directly compared but through the 1-norm ($||...||_1$) between the Gram matrix Gram_j^{ϕ} of the output $\hat{I}_1$ (sCT) and the target I_2 (CT) are computed. (Equation.4). The gram matrix (Equation.2) capture information about characteristics that tend to activate together.

$$\text{Gram}_j^{\phi}(I_2) = \frac{\psi\psi^T}{C_j * H_j * W * j} \tag{2}$$

$$\psi \text{ being the flattened matrix } \phi_j(I_2) \text{ of size: } C_j * H_j W_j \tag{3}$$

$$l_{style}^{\phi,j}(\hat{I}_1, I_2) = ||\text{Gram}_j(\hat{I}_1) - \text{Gram}_j(I_2)||_1 \tag{4}$$

For the reconstruction, the perceptual loss $l_{perceptual}^{\phi}$ which is the weighted sum of these two functions ($l_{cont}^{\phi,j}$ and $l_{style}^{\phi,j}$) is optimized:

$$l_{perceptual}^{\phi}(\hat{I}_1, I_1, I_2) = \alpha * l_{cont}^{\phi}(\hat{I}_1, I_1) + \beta * l_{style}^{\phi}(\hat{I}_1, I_2) \tag{5}$$

where α and β are the weighting factors

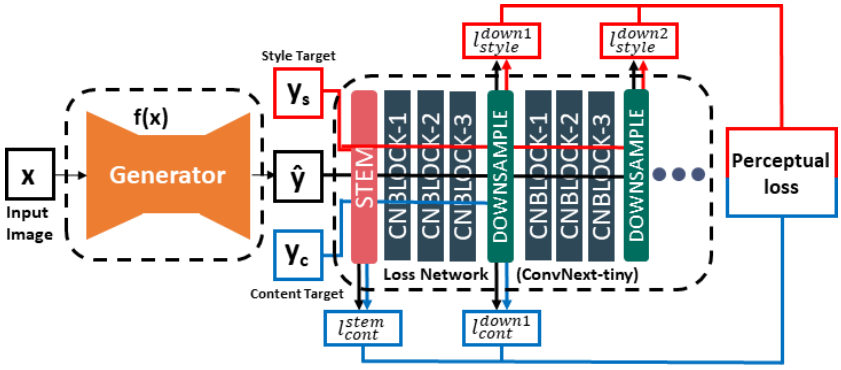

Fig. 1. ConvNext-based perceptual loss definition diagram. The new perceptual loss is the weighted sum of the content (in blue) and style (in red) loss functions (Equation.6) defined from different layers of the pre-trained ConvNext-tiny network [4].

$l^{\phi}_{perceptual}$ is used to generate the sCT from the content of the CBCT and the style of the CT of the same patient. To calculate the perceptual loss, the input images are normalized with the same mean and standard deviation as the ImageNet pre-trained ConvNext. The input images are repeated on all 3 channels (RGB) to match the natural image format and are resampled to $224 * 224$ size from the initial size. As part of the training process, the magnitudes of different loss functions are adjusted to align with those of conventional loss functions, such as Binary Cross Entropy (BCE) used for both the generator and discriminator.
The sum operation contributes to the reduction of the Style loss, whereas the mean operation is employed to reduce the Content loss. In our case, style loss and content loss are weighted as noted in the following equation:

$$l^{\phi}_{perceptual}(\text{sCT}, \text{CBCT}, \text{CT}) = 50 * l^{\phi}_{cont}(\text{sCT}, \text{CBCT}) + 20 * l^{\phi}_{style}(\text{sCT}, \text{CT}) \quad (6)$$

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 100 epochs with a batch size of 128 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 2D patch from the entire image, with size 224x224 for the pelvic region and 168x168 for the brain, employing a stride shape equal to half of the patch.

Inference During the inference process, CBCT intensities are adjusted through histogram matching between the CBCT and the CT scans from the atlas, specifically selecting CT scans with the highest mutual information from the training set.

The CBCT undergoes a reshaping procedure to align its dimensions with a multiple of the patch size. This alignment is achieved by padding the original CBCT image with the value -1000. Subsequently, the CBCT slices are 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. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B.: Generative Adversarial Networks (Jun 2014)
2. 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]
3. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution (Mar 2016)
4. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s (Mar 2022), arXiv:2201.03545 [cs]
5. 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). https://doi.org/10.1002/mp.15150