Synthrad 2023 - MRI-to-sCT generation to facilitate MR-only Radiotherapy

Thomas Helfer[1][0000-1111-2222-3333], Walter Hugo Lopez Pinaya[2][0000-0003-3739-1087], Francisco Pereira[4][0000-0003-2773-3426], Adam G. Thomas[3][0000-0002-2850-1419], and Jessica Dafflon[3,4][0000-0003-2540-0927]

[1] IACS, Stony Brook University, Stony Brook NY 11794, USA
[2] Department of Biomedical Engineering, King's College London, London, UK
[3] Data Science and Sharing Team, Functional Magnetic Resonance Imaging Facility, National Institute of Mental Health, Bethesda, USA
[4] Machine Learning Team, Functional Magnetic Resonance Imaging Facility National Institute of Mental Health, Bethesda, USA
`jessica.dafflon@nih.gov`

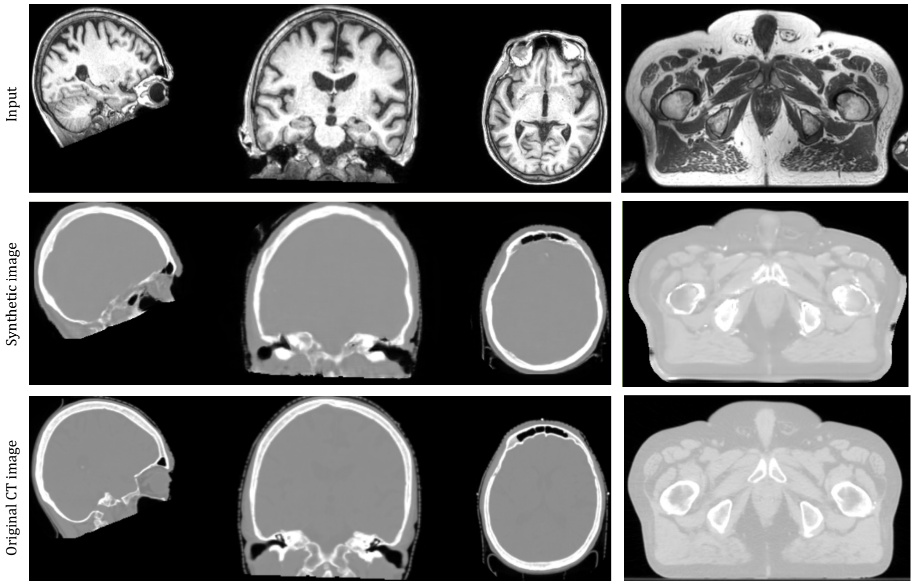

Fig. 1. Example of two synthetic CT images (sCT) from the MRI image (input). (Left) Generated brain image; (Right) generated pelvis. We can see that for both tissues the algorithms are able to generate synthetic images that resample the original CT image.

Keywords: Generative Models · Medical Imaging · Diffusion Models · Generative Adversarial Networks

1 Introduction

Medical imaging plays a crucial role in both the diagnosis and treatment of cancer patients, especially within the context of radiotherapy (RT). While conventional practice leans on Computed Tomography (CT) scans to identify treatment areas and determine optimal radiotherapy dosage, the advent of Magnetic Resonance Imaging (MRI)-only based RT has emerged as an attractive proposition due to reduced radiation exposure and enhanced tissue contrast. This approach can potentially diminish overall treatment expenses, workload, and even rectify residual registration discrepancies inherent in using both CT and MRI. Nevertheless, a significant challenge delaying the implementation of MRI-only RT is the absence of tissue attenuation information necessary for precise dose calculations. Specifically, bones are responsible for significant attenuation, and are not directly visible in MR. Various methodologies have been proposed to bridge this gap, including techniques to convert MRI data into CT-equivalent images (i.e., synthetic CT (sCT)) for precise treatment planning and dose computation. In response to this methodological challenge, the first task of the Synthrad 2023 challenge consisted of generating synthetic CT from given MRI scans. This report provides an overview of the two methods we employed in generating sCT: a latent diffusion model [6] with Controlnet [9] for brain images and a Pix2Pix model [3] for pelvis images.

2 Methods

We used the provided dataset by the Synthrad2023 competition [7], consisting of pelvis and brain images. Within this dataset, the competition organizers provided aligned CT and MRI images that we used to train the models. For a detailed description of the dataset, refer to [8].

Our method used different models; while a latent diffusion model (LDM) was used to generate brain images, a Pix2Pix model was used for the pelvis. We opted for distinct models due to variations in image sizes. As many pelvis images were excessively large to be processed effectively on the competition's provided hardware, we employed a more lightweight model for them.

2.1 Brain: Latent Diffusion Model

Diffusion Models [2] leverage an iterative denoising process to transform Gaussian noise into samples from a learned data distribution. However, training diffusion models can be very memory- and computationally intensive, therefore, applying diffusion models to latent representations allows for a better scalability and enables their use for high-dimensional data tasks, like 3D medical images.

This is why for this challenge, we used (1) an autoencoder to compress the input data into a lower-dimensional latent representation, (2) used the generative modeling properties of diffusion models in the latent space, and (3) conditioned the generative process using another image with a ControlNet (Figure 2). We trained the autoencoder with a combination of L1 loss, perceptual loss [10], a patch-based adversarial objective [1], and a KL regularization of the latent space. The encoder maps the brain image to a latent representation with a size of 48 × 48 × 48, which was then used as the diffusion model input. We also conditioned the generated CT with an MRI image during training using a ControlNet [9]. The ControlNet acted as a tool to guide and adjust the synthesis based on specific spatial and structural constraints. This facilitated the generation of images that closely adhered to desired spatial arrangements and structural configurations, allowing for a finer level of control over the final output.

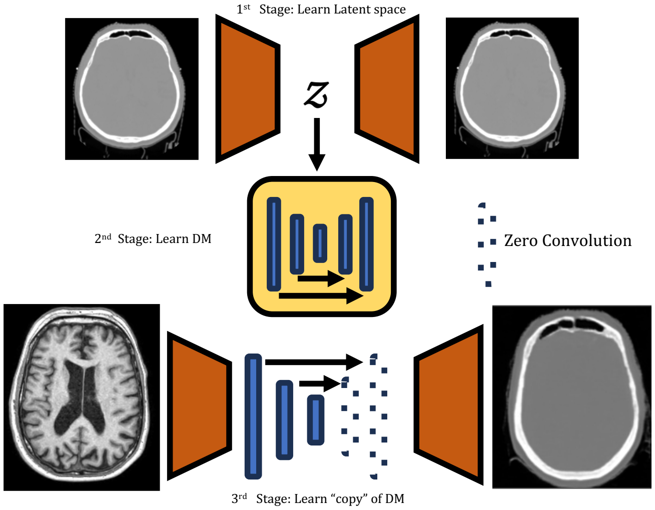

Fig. 2. The latent diffusion model consists of three steps: (1) we train an autoencoder with KL-loss to obtain a latent representation of the image; (2) then we train a Diffusion model using the latent space (z) obtained from the autoencoder; (3) finally, we train a ControlNet on top of the Diffusion model to control the generation of the fine grain details, by mapping the initially provided MRI image to the desired CT image.

We used the models and losses provided by the new MONAI Generative Models framework [5], an open-source platform that allows researchers to quickly train, develop, and evaluate generative models for medical imaging purposes. Our experiments followed the work from [4], as we used an LDM to generate high-resolution synthetic images.

2.2 Pelvis: Pix2Pix

In computer vision imaging, the Pix2Pix technique [3] stands as a powerful tool for translating between image domains. Employing a conditional generative adversarial network architecture, Pix2Pix has gained prombachinence for its ability to learn the intricate mapping between input and output images in paired datasets. By training the model on a dataset of 3D aligned pairs, each consisting of an MRI and its corresponding CT image, we enable Pix2Pix to learn the intricate relationships between these two modalities.

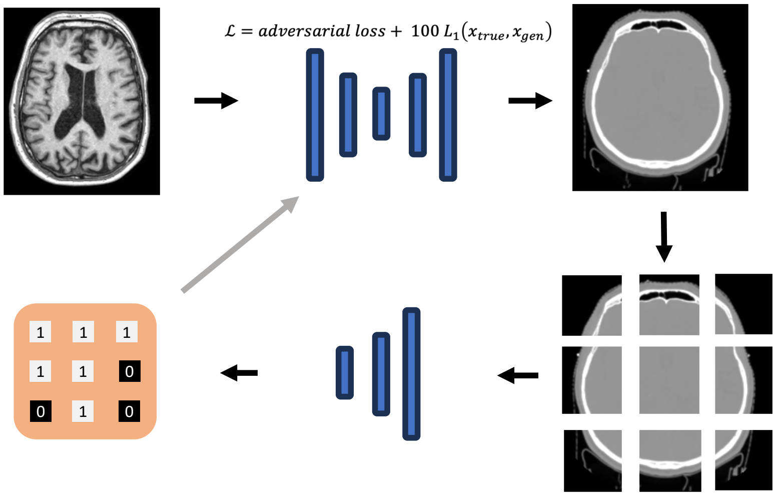
$\mathcal{L} = adversarial\ loss + 100\ L_1(x_{true}, x_{gen})$
1 1 1
1 1 0
0 1 0

Fig. 3. A paired MRI-CT dataset is used to train a Pix2Pix model to transform MRI into CT scans. During training, the input image is processed through a U-Net architecture. The training process involves utilizing an L1 loss function, measuring the pixel-wise disparity between the synthesized CT scan and the actual ground truth CT scan. Simultaneously, an adversarial loss function is incorporated to ensure that the generated CT scan is realistic and indistinguishable frombach a real CT scan. In the image, the orange square represents which parts of the patch the discriminator identifies as real or fake, this information is then used to update the weights of the network.

The heart of the Pix2Pix architecture lies in the generator-discriminator interplay. The generator, a U-Net network, learns to convert the structural features captured in MRI scans into the characteristic intensity patterns found in CT scans. It takes as input an MRI slice and produces a synthetic CT image. To enforce the similarity between the generated CT images and the actual CT scans, we employ a L1 loss function during training. This loss function quantifies the pixel-wise differences between the generated and target images, prombachoting the generation of visually accurate CT representations. In particular, we used the following loss:

$$\mathcal{L} = \text{adversarial loss} + \beta\mathcal{L}_1(x_{\text{true}}, x_{\text{gen}}), \tag{1}$$

where we weight the L1 loss (using a scaling factor β - we choose β to be 100) relative to the adversarial loss.

The discriminator network, employing a patch-based approach with a patch size of 9 × 9, provides feedback to ensure the generated CT-like images look realistic. This patch-based discriminator analyzes local patches of the generated CT scan and the corresponding real CT scan to guide the generator's learning process. The synergy of the L1 loss and the discriminator's adversarial training contributes to improved results. Notably, while the L1 loss ensures global structural coherence, the discriminator improves fine details and texture synthesis, aspects that are often lacking when using the L1 loss in isolation.

Through an iterative adversarial training process with a batch size of one, chosen due to the large memory requirement for 3D pictures, and a learning rate of 1e-3, the model refines its ability to generate high-quality CT-like representations frombach MRI scans.

References

1. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
2. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
3. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
4. Pinaya, W.H.L., Tudosiu, P.D., Dafflon, J., Costa, P.F.D., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: Deep Generative Models, pp. 117–126. Springer Nature Switzerland (2022)
5. Pinaya, W.H., Graham, M.S., Kerfoot, E., Tudosiu, P.D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P.F., Patel, A., et al.: Generative ai for medical imaging: extending the monai framework. arXiv preprint arXiv:2307.15208 (2023)
6. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
7. Thummerer, A., van der Bijl, E., Galapon, A., Verhoeff, J.J.C., Langendijk, J.A., Both, S., van den Berg, C.N.A.T., Maspero, M.: Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. Medical Physics **50**(7), 4664–4674 (Jun 2023). https://doi.org/10.1002/mp.16529, http://dx.doi.org/10.1002/mp.16529
8. Thummerer, A., van der Bijl, E., Galapon Jr, A., Verhoeff, J.J., Langendijk, J.A., Both, S., van den Berg, C.N.A., Maspero, M.: Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. Medical Physics (2023)
9. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
10. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)