Diffusion-Based Synthetic CT Generation from MRI: A Comparative Evaluation with Deep Learning Baselines

Hanyuan Su*

International School, Jinan University, Guangzhou, Guangdong, China

Abstract. Generating synthetic CT images based on magnetic resonance imaging (MRI) is an important research direction in radiotherapy. MRI has good soft tissue contrast but lacks electron density information needed for dose calculation, while CT images can provide Hounsfield Unit information and are therefore often used for radiotherapy planning. This paper compares and analyzes four deep learning methods, Diffusion, U-Net, GAN, and TransformerUNet, for the MRI-to-CT image synthesis task. Experiments are conducted using paired brain MRI-CT data from the SynthRAD2023 Grand Challenge dataset, and the generated results are evaluated using MAE, PSNR, SSIM, and NCC. The results show that the Diffusion model performs best overall, outperforming other models in MAE, PSNR, and NCC. The generated synthetic CT is closer to the real CT and can better preserve anatomical structures. However, the Diffusion model has a slow inference speed and high GPU memory requirements, and further optimization of computational efficiency is needed.

1 Introduction

Computed tomography (CT) images are widely used in radiotherapy planning because they provide Hounsfield unit information, which can be converted into electron density for dose calculation. In contrast, magnetic resonance imaging (MRI) provides better soft-tissue contrast and is valuable for delineating tumors and organs at risk. Therefore, generating a synthetic CT from MRI is an important task, especially for MRI-only radiotherapy workflows.

MRI-to-CT synthesis is a challenging cross-modality image translation problem. Although MRI and CT images can represent the same anatomical structures, their imaging mechanisms and intensity distributions are very different. MRI signals mainly reflect tissue relaxation properties, while CT intensities are related to X-ray attenuation and electron density. This difference makes the mapping from MRI to CT highly nonlinear, particularly in bone regions where MRI signals are usually weak but CT values are high. As a result, an effective synthesis model should preserve anatomical structures while generating CT-like intensity information.

In recent years, deep learning methods have been widely used for synthetic CT generation. U-Net-based models can preserve spatial information through encoder-decoder structures and skip connections. GAN-based methods can improve image realism through adversarial training, but they may suffer from unstable training and local artifacts. Transformer-based models, such as TransformerUNet, can capture long-range dependencies using self-attention mechanisms. However, these methods usually treat MRI-to-CT synthesis as a direct mapping task, which may limit their ability to model complex cross-modality uncertainty.

Diffusion models provide a different solution by formulating image generation as a progressive denoising process. Instead of generating CT images in a single forward pass, diffusion models gradually recover the target image from noise under conditional guidance. In this study, four models, including Diffusion, U-Net, GAN, and TransformerUNet, are compared for MRI-to-CT synthesis. The Diffusion model is based on conditional Improved DDPM and uses SwinVIT as the denoising network. All models are trained and tested on the brain subset of the SynthRAD2023 Grand Challenge dataset and evaluated using qualitative visualization and quantitative metrics, including MAE, PSNR, SSIM, and NCC.

2 Method

The aim of this study is to explore the efficacy of diffusion-based models in MRI-to-CT synthesis and to contrast their performance with that of representative image translation networks, such as U - Net, GAN, and TransUNet[1]. Given a patient's magnetic resonance image, the proposed framework intends to generate a synthetic computed tomography image that is anatomically consistent with the input MR image and quantitatively approximates the corresponding real CT image. In comparison with direct regression - based methods, diffusion models formulate image synthesis as a progressive denoising process, which offers a more stable and expressive generative mechanism for cross - modality medical image translation[2].

*hysu0508@stu2023.jnu.edu.cn

To evaluate the effectiveness of diffusion-based MRI-to-CT synthesis, three representative baseline models are selected for comparison. U-Net is used as a CNN-based direct mapping baseline due to its widely used encoder-decoder structure with skip connections[3]. GAN is included as an adversarial generative baseline, which improves image realism through generator-discriminator training but may suffer from unstable optimization and structural artifacts[4]. TransUNet is selected as a Transformer-based direct mapping baseline to evaluate the effect of global feature modeling in supervised image translation.

These benchmark models represent CNN regression, adversarial generation, and Transformer-based mapping, respectively, providing a comparison for validating the effectiveness of the diffusion model in MRI-to-CT synthesis[2][3][4].

2.1 Problem Formulation

Let x^{MR} denote the input MR image and x^{CT} denote the corresponding real CT image from the same patient. This study takes the three-dimensional medical image, so both images can be represented as volumetric data:

$$x^{MR}, x^{CT} \in R^{H\times W\times D} \quad (1)$$

where H, W, and D denote the height, width and depth of the image respectively. The goal of the MRI-to-CT synthesis task is let the model learn a mapping function F_θ from MR domain to CT domain:

$$\hat{x}^{CT} = F_\theta(x^{MR}) \quad (2)$$

where $\hat{x}^{CT}$ denotes the generated synthetic CT image and θ denotes the hyperparameters learned by the model. It is expected that the generated synthetic CT image preserves anatomical structures from MRI while maintaining CT intensity information close to the real CT image.

For supervised MRI-to-CT synthesis, the paired training dataset can be defined as:

$$D = \{(x_i^{MR}, x_i^{CT})\}_{i=1}^{N} \quad (3)$$

where N is the number of paired MRI-to-CT training samples. Each sample contains an MR image and its corresponding CT image from the same patient. All the compared models in this study are trained to solve the same MRI-to-CT synthesis task, but they differ in their generation mechanism.

For U-Net and TransUNet, the synthesis task is treated as a one-step deterministic mapping[5][6]:

$$\begin{cases} \hat{x}_{UNet}^{CT} = F_\theta^{UNet}(x^{MR}) \\ \hat{x}_{TransUNet}^{CT} = F_\theta^{TransUNet}(x^{MR}) \end{cases} \quad (4)$$

For GAN-based synthesis, the generator maps the MR image to a synthetic CT image[7]:

$$\hat{x}_{GAN}^{CT} = G(x^{MR}) \quad (5)$$

where G denotes the generator network.

For diffusion-based synthesis, the task is formulated as a conditional denoising process. Given the MR image as the condition, the model learns the reverse denoising distribution[8]:

$$p_\theta(x_{t-1}^{CT} \mid x_t^{CT}, x^{MR}) \quad (6)$$

where x_t^{CT} represents the noisy CT image at time step t, and the x_{t-1}^{CT} represents the CT image with less noise at the previous time step. Through iterative conditional denoising, the diffusion model gradually generates the final synthetic CT image.

2.2 Data

The research uses the brain subset of Task 1 from the SynthRAD2023 Grand Challenge dataset[9] for MRI-to-CT synthesis. The dataset includes 180 paired samples and they are preprocessed by the organizers through format conversion, resampling, rigid registration, anonymization, body mask generation, and cropping. In order to reduce the influence of image background, the final MR-CT pairs have a voxel size of $100 \times 100 \times 100$. The patch size is chosen as $64 \times 64 \times 4$ for training. Before training, the voxel intensities of CT scans are cut to $[-1007,3000]$, and jointly normalized

to the interval $[-1,1]$. For MR images, their voxel intensities are independently normalized to $[-1,1]$. The research uses the top 75% of patients in the dataset to train our model, the remaining 20% for testing, and 5% for validation.

2.3 Implementation Detail And Performance Evaluation

2.3.1 Baseline Model

This study established three benchmark models for MRI-to-CT image synthesis. First, the 3D U-Net is used as the convolutional neural network benchmark model. This model consists of a three-layer encoder, a bottleneck layer, and a three-layer decoder. Each encoder and decoder module includes two 3D convolutions, instance normalization, and a LeakyReLU activation function, and shallow spatial details are fused with deep semantic features through skip connections. The model's input and output are both single-channel 3D medical images, with a base number of 32 channels, varying to 32, 64, 128, and 256 channels. Finally, a normalized CT image is generated using a $1 \times 1 \times 1$ convolution and a tanh activation function.

Second, a conditional generative adversarial network based on 3D PatchGAN is used as the generative benchmark model. The model uses a 3D U-Net as the generator to generate corresponding CT images based on MRI scans. The discriminator takes the channel-stitched result of the MRI and the real or generated CT scan as input, and outputs patch-level real/fake images through multiple layers of 3D convolution. The discriminator has a base of 32 channels, which increases to 256 layers. During training, the generator loss consists of adversarial loss and L1 reconstruction loss, with the L1 loss weight set to 100 to ensure that the synthesized CT scan closely approximates the real CT scan in terms of structure and grayscale distribution.

Finally, TransformerUNet 3D is used as the Transformer enhancement model. This model maintains the encoder-decoder structure of U-Net, but introduces a Transformer module at the bottleneck layer.

Specifically, after the convolutional bottleneck layer outputs 256 channels of features, the 3D features are unfolded into a sequence and input into a global modeling module consisting of four Transformer Blocks. Each Transformer Block contains a LayerNorm, a multi-head self-attention mechanism, and a feedforward neural network, with 8 attention heads and an MLP expansion ratio of 4. This design allows the model to retain the local feature extraction capabilities of U-Net while further enhancing its ability to model long-range dependencies in 3D medical images.

2.3.2 Diffusion Model

This study employs a conditional diffusion model based on Improved DDPM[10] as the primary method. The model uses MRI images as conditional input and utilizes a SwinVIT network, combining Swin Transformer and residual convolution as a denoiser. During training, real CT images are progressively denoised over 1000 steps, and the network is trained to learn the inverse diffusion process for recovering the original CT from the noisy CT. During inference, starting from random Gaussian noise, a synthetic CT is generated through 50 iterative denoising steps under MRI conditional constraints. The network adopts a four-layer hierarchical coding structure with a basic number of 64 channels and channel expansion coefficients of $(1,2,3,4)$. A Swin Transformer attention mechanism is introduced in the multi-scale feature layers to enhance long-range dependency modeling capabilities, thereby achieving high-quality cross-modal image conversion from MRI to CT. The whole process is shown in Figure 1.

The AdamW optimizer is used to update the parameters of the SwinVIT network in the diffusion model. The learning rate is set to 1×10^{-5}, and the weight decay coefficient is set to 1×10^{-4}, to alleviate overfitting and improve training stability. The first and second moment estimation coefficients in the AdamW optimizer use the default settings, i.e., $\beta_1 = 0.9$, $\beta_2 = 0.999$ and the numerical stability term $\epsilon = 1 \times 10^{-8}$. Furthermore, automatic mixed precision and GradScaler are used for gradient scaling during training to reduce memory usage while maintaining numerical stability during training.

2.3.3 Synthetic CT Evaluation

To evaluate the quality of the synthetic CT (sCT) images generated by the four models, the study compares each sCT result with the corresponding real CT image. Four quantitative metrics are used, including mean absolute error (MAE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and normalized cross-correlation (NCC). Specifically, MAE is used to measure voxel-wise intensity differences, PSNR to assess peak signal similarity, SSIM to evaluate structural and perceptual similarity, and NCC to quantify the correlation between the generated and real CT images. Lower MAE values indicate smaller synthesis errors, whereas higher PSNR, SSIM, and NCC values suggest better sCT image quality.

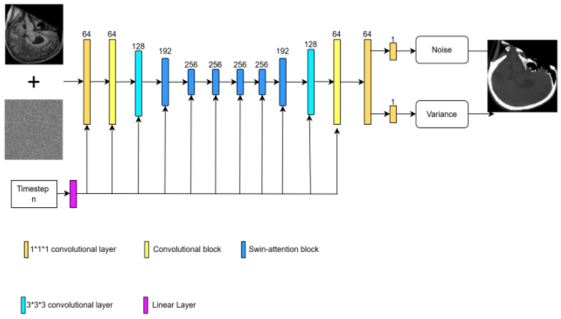

Figure 1: The diagram of the Diffusion Model Process (Picture credit: Original)

3 Result

Figure 2 shows the visual results of the four models. The first row shows the real CT and the corresponding MRI image. The following rows present the results of the diffusion model, U-Net, GAN, and TransformerUNet.

The results are shown in the X, Y, and Z views. This makes it easier to compare the generated images from different anatomical directions. The CT images are displayed with an intensity window of -100 to 400 HU, and the difference maps are used to show the local errors between the generated CT and the real CT.

From the figure 2, all four models are able to generate CT-like images from MRI. The main brain structure is generally preserved. However, the quality of the generated images is different among the models. The diffusion model gives clearer boundaries and its overall shape is closer to the real CT. This is especially visible around the skull contour and some internal tissue regions. In the difference maps, the diffusion model also shows fewer large-error regions, showing that its generated CT is closer to the real CT at the voxel intensity level.

The other three models still have visible errors in some areas. U-Net keeps the main structure, but the generated image looks smoother and some details are weakened. GAN produces images with a relatively realistic appearance, but local artifacts can still be observed in some regions. One possible reason is the instability of adversarial training. TransformerUNet uses Transformer modules for global feature modeling, but errors remain around local boundaries and complex anatomical regions.

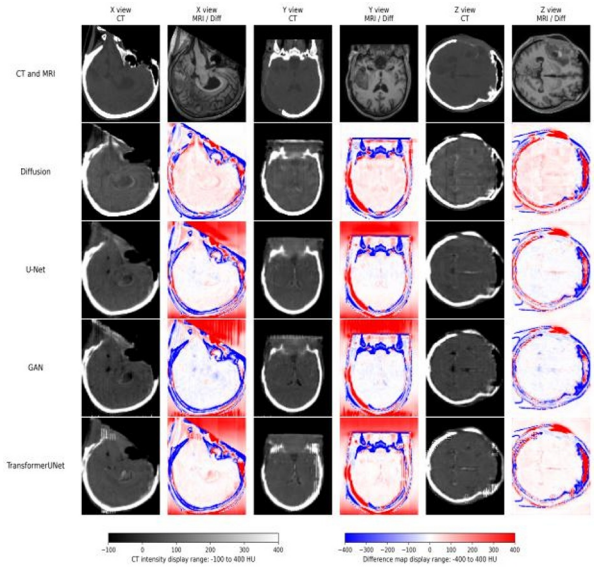

Figure 2: Qualitative Comparison of MRI-to-CT Synthesis Results (Picture credit: Original)

The quantitative results in Table 1 show a similar trend. The diffusion model gets the lowest MAE of 91.16 HU, the highest PSNR of 25.37 dB, and the highest NCC of 0.94. These values show that the generated CT from the diffusion model is closer to the real CT in terms of intensity error, signal similarity, and image correlation. Although the U-Net has a slightly higher SSIM than the diffusion model, the difference is small. Overall, the Diffusion model performs better on most evaluation metrics.

Table 1: Quantitative Comparison of MRI-to-CT Synthesis Performance

Model	MAE	PSNR	SSIM	NCC
Diffusion	91.16	25.37	0.74	0.94
GAN	163.04	23.21	0.65	0.90
TransfromerUnet	158.60	24.10	0.71	0.92
U-Net	150.73	24.45	0.74	0.93

4 Discussion

In this study, diffusion, U-Net, GAN, and TransformerUNet are compared for the MRI-to-CT synthesis task. Among the four models, the diffusion model shows the best overall result. It gets the lowest MAE and the highest PSNR and NCC, showing that its generated CT images have smaller intensity errors and stronger similarity with the real CT images. Although the difference between models is not the same for every metric, the diffusion model performed better in most of the main evaluation indicators.

The visual comparison also supports this result. In the three displayed views, the CT images generated by the diffusion model look closer to the real CT, especially around the skull boundary and some internal tissue regions. U-Net keeps the general brain shape, but the generated image is relatively smooth. Some details are weakened. GAN can produce a CT-like appearance, but local artifacts and intensity changes can still be found in some areas. TransformerUNet benefits from the Transformer structure, but it still has errors near local boundaries and high-density bone regions.

A possible reason for the better result of the diffusion model is its step-by-step denoising process. U-Net and TransformerUNet predict the CT image directly in one forward pass, while the diffusion model gradually recovers the CT image from noise. This process may be more suitable for the MRI-to-CT synthesis, because the relationship between MRI intensity and CT intensity is not simple. In addition, the SwinVIT denoising network combines convolutional features with attention-based modeling, which helps the model use both local image details and larger anatomical information.

There are still some limitations in this study. The generated CT images still have errors around bone boundaries and high-density regions. The accurate bone reconstruction from MRI remains difficult. The dataset used in this experiment is also limited. The suggestion is that the generalization ability of the models needs to be tested on more patients and data from different scanners or imaging protocols. Another limitation is the computational cost of the diffusion model. Since it requires multiple sampling steps during inference, it is slower than direct generation models such as the U-Net. Its larger model structure also requires more GPU memory during training and testing.

Future work can focus on improving both model accuracy and efficiency. A larger dataset could be used to improve the robustness of the models. The sampling strategy of the Diffusion model could also be optimized to reduce inference time while keeping image quality. In addition, since synthetic CT images are often used for radiotherapy dose calculation, dosimetric evaluation should be included in future studies to further check whether the generated CT images are reliable for clinical use.

5 Conclusion

This paper compares the performance of four models, diffusion, U-Net, GAN, and TransformerUNet, in the MRI-to-CT image synthesis task. Experimental results show that the diffusion model outperforms the other three models on most metrics. It gets the lowest MAE and the highest PSNR and NCC, showing that its generated CT images are closer to real CT scans. Visualization results also show that the diffusion model preserves skull boundaries and local tissue structures more clearly.

This result may be related to the generation method of the diffusion model. U-Net and TransformerUNet directly predict CT images, while the diffusion model starts with noise and gradually reconstructs the CT image. This gradual denoising process allows the model to more stably learn the complex relationship between MRI and CT. Simultaneously, the SwinVIT structure

helps the model extract both local details and large-scale anatomical information.

However, the diffusion model also has significant limitations. Due to the need for multi-step sampling during inference, its generation speed is slower than that of the direct generation model. Meanwhile, due to the more complex network structure, training and inference require more GPU memory. Therefore, although the diffusion model performs better in terms of image quality, its computational efficiency still needs further improvement. It can explore faster sampling methods, more lightweight network structures, and use larger-scale datasets to further validate the model's performance in the future.

References

1. Boulanger, M., Nunes, J. C., Chourak, H., Largent, A., Tahri, S., Acosta, O., de Crevoisier, R., Lafond, C., & Barateau, A. (2021). Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review. Physica Medica, 89, 265–281.
2. Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., Stegmaier, J., Kuhl, C., Nebelung, S., Kather, J. N., Truhn, D., & Kiessling, F. (2023). Denoising diffusion probabilistic models for 3D medical image generation. Scientific Reports, 13, Article 7303. https://doi.org/10.1038/s41598-023-34341-2
3. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015.
4. Lyu, Q., & Wang, G. (2022). Conversion between CT and MRI images using diffusion and score-matching models. arXiv preprint arXiv:2209.12104.
5. Li, X., Yadav, P., & McMillan, A. B. (2022). Synthetic computed tomography generation from 0.35T magnetic resonance images for magnetic resonance-only radiation therapy planning using perceptual loss models. Advances in Radiation Oncology, 7(1), 100767.
6. Phan, V. M. H., Xie, Y., Zhang, B., Qi, Y., Liao, Z., Perperidis, A., Phung, S. L., Verjans, J. W., & To, M.-S. (2024). Structural attention: Rethinking transformer for unpaired medical image synthesis. In Medical Image Computing and Computer Assisted Intervention — MICCAI 2024.Kuznetsova, E., Li, Y. F., Ruiz, C., & Zio, E. (2013). Reinforcement learning for microgrid energy management.Energy, 59, 133–146.
7. Kang, S. K., An, H. J., Jin, H., Ahn, S. H., Kim, J., Lee, J. S., & Chie, E. K. (2021). Synthetic CT generation from weakly paired MR images using cycle-consistent generative adversarial network for MR-guided radiotherapy. Medical Physics, 48(12), 7393–7403.
8. Lyu, Q., & Wang, G. (2022). Conversion between CT and MRI images using diffusion and score-matching models. arXiv preprint arXiv:2209.12104.
9. Thummerer, A., van der Bijl, E., Galapon, A., Verhoeff, J. J. C., Langendijk, J. A., Both, S., Lee, J. A., van den Berg, C. A. T., & Maspero, M. (2023). SynthRAD2023 Grand Challenge dataset: Generating synthetic CT for radiotherapy. Medical Physics, 50(7), 4664–4674. https://doi.org/10.1002/mp.16529
10. Pan, S., Abouei, E., Wynne, J., Wang, T., Qiu, R. L. J., Li, Y., Chang, C.-W., Peng, J., Roper, J., Patel, P., Yu, D. S., Mao, H., & Yang, X. (2024). Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model. Medical Physics.