A Hybrid Network with Multi-scale Structure Extraction and Preservation for MR-to-CT Synthesis in SynthRAD2023

Zeli Chen[1,2], Kaiyi Zheng[1], Chuanpu Li[1], and Yiwen Zhang[1]

[1] School of Biomedical Engineering, Southern Medical University, Guangzhou, 510515, China
[2] DAMO Academy, Alibaba Group

Abstract. Synthesis of Computed Tomography (CT) images from Magnetic Resonance (MR) images is clinical significance for MR-only treatment planning to eliminate the co-registration errors between MR and CT images. In this paper, we combined the advantages of CNN and Transformer to propose a hybrid network with multi-scale structure extraction and preservation, named MSEP, for MR-to-CT synthesis in SynthRAD2023 challenge. Experimental results with unseen validation and preliminary test from the leaderboard showed the effects of our methodology.

Keywords: MR-to-CT synthesis · Transformer · Multi-scale · CNN.

1 Introduction

Radiotherapy is a critical and effective component of comprehensive cancer treatment and care. Because of the superior soft tissue contrast for providing precise information of the tumor and the organs at risk (OAR), the magnetic resonance (MR) image has been commonly used as an auxiliary modality in radiotherapy [10]. The registration of CT and MR images is a routine clinical practice for transferring the tumor and OAR delineations since MR images can not provide electron density information for dose calculation. To reduce the unnecessary radiation exposure and the systematic errors in MRI/CT co-registration, the MR-to-CT synthesis is a clinically significant solution for MR-only treatment planning [9].

Many different methods have been proposed for the generation of synthetic CT from MR images [8,11,1,2,5]. Most of these methods [8,11,1] are based on the convolutional neural network (CNN), which pays more attention to local detailed information, making synthetic CT a better structural consistency with real CT. For example, Boni et al. [1] adopted a conditional GAN framework called pix2pixHD to create a robust model prone to multi-center data. However, due to the principles of local processing and inherent weight sharing in the convolutional layer, these methods intrinsically limit the effectiveness of long-range spatial dependence and the ability to learn global contextual information. They

pose dramatic limitations to distinguish cortical bone and air which show low signals in conventional MR images. Recently, ViT [6] adopted a global self-attention mechanism to capture long-range contextual information. Several studies [5,4] have developed methods using the high-dimensional features from the CNN encoder into the ViT module to combine the merits of CNN and Transformer and reduce the computational overhead. However, the quadratic complexity and the high computation cost of the global self-attention mechanism still constrain the flexibility of MR-to-CT synthesis in clinical practices.

In this paper, we combined the advantages of CNN and Transformer to propose a hybrid network with multi-scale structure extraction and preservation, named MSEP, for MR-to-CT synthesis in SynthRAD2023 challenge. During the multi-scale structure extraction stage, we used CS-Attention in the encoder to adaptively extract spatial position information of image tokens of different sizes. During the multi-scale structure preservation stage, we added a residual dilated swin transformer (RDSformer) to each skip connection in UNet to better preserve structural information in cross-modal features [7]. Finally, to enhance the detail information of the synthesized CT, we utilised a perceptual loss based on VGG19 during training. Experimental results with unseen validation and preliminary test from the leaderboard showed the effects of our methodology.

2 Method

The overview of our proposed MSEP developed on UNet (See Fig. 1 (a)) for MR-to-CT synthesis is shown in Fig. 1.

Multi-scale Structure Preservation Stage To better preserve structural information in cross-modal features from encoder and decoder, we added a residual dilated swin transformer (RDSformer) to each skip connection in UNet. Each RDS transformer block contains a smoothed dilated convolution [3] and a Swin Transformer block to capture complementary information between the two different modules (See Fig. 1 (b)). The smoothed dilated convolutions in front of the Swin Transformer blocks use separable and shared convolutions before the dilated convolutions to increase the receptive field. In order to keep the network stable, we add two long residual lines between the RDS transformer blocks.

Local attention and shift window in Swin Transformer are the keys to significantly reduce the computational complexity. The only difference between the two Swin Transformer layers in a complete Swin Transformer block is whether the window multi-headed attention (W-MSA) module contains a shifted window. We use a 3D local window to replace the 2D local window for self-attention computation. The process of a RDSfomer block is formulated as:

$$\begin{aligned} \hat{\mathcal{X}}_s^l &= (\mathrm{S})\,\mathrm{W\text{-}MSA}\left(\mathrm{LN}\left(\mathrm{SDConv}\left(\mathcal{X}_s^{l-1}\right)\right)\right) + \mathcal{X}_s^{l-1}, \\ \mathcal{X}_s^l &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{\mathcal{X}}_s^l\right)\right) + \hat{\mathcal{X}}_s^l, \end{aligned} \tag{1}$$

where, l stands for the layer index and (S) is for using a shift window. $\mathrm{SDConv}\,(\cdot)$ denotes smoothed dilated convolution function.

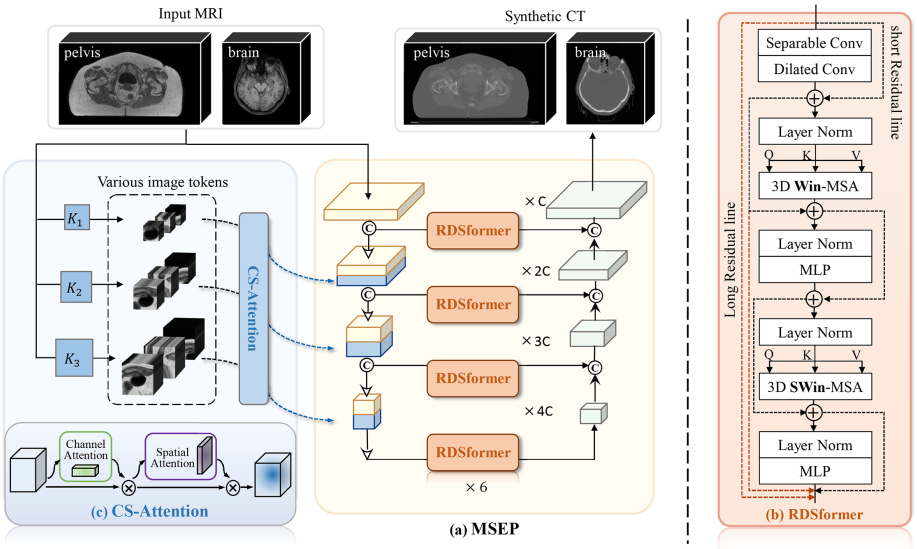

Fig. 1. Framework of MSEP.

Multi-scale Structure Extraction Stage To overcome the limitation of CNN-based approaches in extracting only local information, we propose multi-scale structure extraction branches (MSE branches) to adaptively capture the multi-scale global spatial information and long-range dependencies. The MSE branch, which embeds input images with various sizes of tokens in each down-sampling branch of a CNN-based encoder, contains an image tokenization and a token interaction. The proposed MSE branches with i-th scale can be expressed as:

$$\mathbf{M}_i(\mathcal{X}) = \mathbf{CS\text{-}Atten}(\mathbf{P}(\mathcal{X}, k_i)) \tag{2}$$

where $\mathbf{P}(\cdot)$ denotes patch embedding through a 3D convolution of kernel size k_i and a reshape operation. In the image tokenization step, we use the convolutions with different kernel sizes to slice the input image into 3D non-overlapping multi-scale tokens. In the token interaction, the 3D image tokens are reshaped into a sequence of flattened tokens and are fed to CS-Attention (See Fig. 1 (c)) to adaptively capture the long-range and global spatial information. Specifically, CA-Attention block consists of a channel attention and a spatial attention. To integrate local detailed features from the CNN-based encoder and multi-scale structure-aware information from MSE branches, the fused features can be formulated as:

$$\begin{aligned} \zeta_0 &= \mathbf{F}_0(\mathcal{X}), \\ \zeta_j &= \mathbf{M}_j(\mathcal{X}) + \mathbf{F}_j(\zeta_{j-1}), \quad j = 1, 2, 3, \end{aligned} \tag{3}$$

where $\mathbf{F}_j$ is the j-th convolution layer in the CNN-based encoder.

Objective Functions In MSEP, the objective function is mainly composed of two parts, including the L1 distance loss, and the VGG-based perceptual loss:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{L1} + \lambda_2 \mathcal{L}_p \tag{4}$$

where λ_1 and λ_2 are the scaling factors controlling the relative importance of the loss terms, and the perceptual loss $\mathcal{L}_p$ adopts the pre-trained VGG-19 network to improve visual quality.

3 Experiments and Results

3.1 Implementation Details

Due to the significant size differences between the head and pelvis data, we trained them separately. For the head data, we divided the training and validation sets by proportionately splitting them from three centers, with a ratio of 171 for training and 9 for validation. For the pelvis data, we divided them with a ratio of 174 for training and 6 for validation. The image inputs for the head and pelvis models are $48 \times 160 \times 160$ and $32 \times 192 \times 360$, respectively. The normalization method used is z-score and the data augmentation including random translation and reversal was carried out to enhance the model training. All networks were by Adam optimizer with a batch size of 2, an initial learning rate 2×10^{-4}, and 200 epochs. The trade-off parameters settings are set empirically to $\lambda_1 = 20$ and $\lambda_2 = 1$. At the testing stage, an input MR image is first partitioned into overlapping patches, and then average the whole synthetic CT patches to obtain the final synthetic CT image. The experiments were implemented by Pytorch and trained on a Tesla V100 GPU with 32 GB of memory.

3.2 Results

As illustrated in Table. 1, we conducted an ablation study to evaluate the impact of each of the proposed components on MSEP in the head region. Especially, the MSE branches and the RDSformer block had a significant improvement, which proved the effectiveness of multi-scale structure extraction and preservation. Meanwhile, the visualized results of MSEP are shown in Fig. 2.

Table 1. Ablation study for the proposed components on MSEP in head region.

Methods	MAE	SSIM	PSNR
A. Baseline (UNet)	70.221	0.786	28.456
B. A + MSE branches	67.143	0.812	28.841
C. B + RDSformer	62.621	0.846	29.021
D. C + Perceptual loss	**61.985**	**0.858**	**29.114**

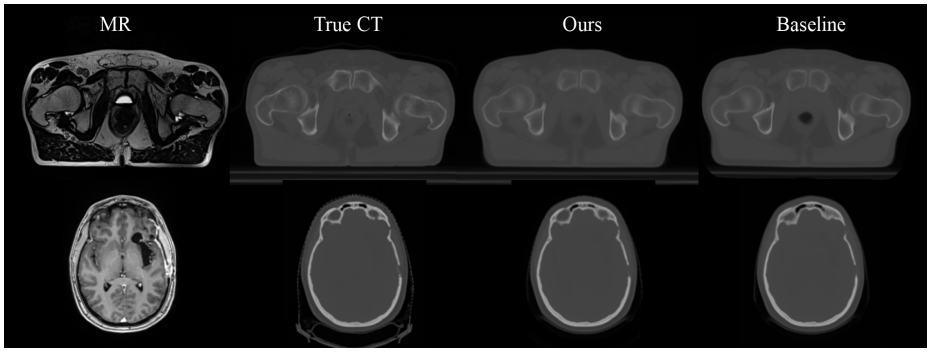

Fig. 2. Visual comparison of synthetic CT images from our MSEP and the baseline.

Table 2. Quantitative comparisons of different models for MR-to-CT synthesis on head region.

Methods	MAE	SSIM	PSNR
A. 2.5D ResNet34	72.101	0.766	28.214
B. SwinUNETR	63.817	0.839	28.912
C. MSEP	61.985	0.858	29.114
D. 0.1A+0.45B+0.45C	**59.893**	**0.871**	**29.307**

Table 3. Quantitative comparisons of different models for MR-to-CT synthesis on pelvis region. MSEP† represents the image input size used is $16 \times 224 \times 224$.

Methods	MAE	SSIM	PSNR
A. 2.5D ResNet34	55.452	0.786	29.508
B. MSEP†	51.213	0.804	29.833
C. MSEP	49.695	0.838	30.068
D. 0.2A+0.3B+0.5C	**48.455**	**0.846**	**29.469**

Ensembling models with significant differences can effectively yield more accurate results. In addition to MSEP, we integrated the 3D SwinUNETR and 2.5D resnet34 models and weighted them according to specific ratios for ensemble integration. The specific results are shown in Table. 2 and Table. 3.

4 Conclusion

In this paper, we propose a hybrid network for MR-to-CT synthesis via multi-scale structure extraction and preservation from Transformer and CNN, named MSEP. We validate the effectiveness of the proposed modules through ablation studies. In unseen validation and preliminary test from the leaderboard showed, ours proposed MSEP achieved remarkable performance in MR-to-CT synthesis on head and pelvis dataset.

References

1. Boni, K.N.B., Klein, J., Vanquin, L., Wagner, A., Lacornerie, T., Pasquier, D., Reynaert, N.: Mr to ct synthesis with multicenter data in the pelvic area using a conditional generative adversarial network. Physics in Medicine & Biology **65**(7), 075002 (2020) 1
2. Brou Boni, K.N., Klein, J., Gulyban, A., Reynaert, N., Pasquier, D.: Improving generalization in mr-to-ct synthesis in radiotherapy by using an augmented cycle generative adversarial network with unpaired data. Medical Physics **48**(6), 3003–3010 (2021) 1
3. Chen, D., He, M., Fan, Q., Liao, J., Zhang, L., Hou, D., Yuan, L., Hua, G.: Gated context aggregation network for image dehazing and deraining. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1375–1383. IEEE (2019) 2
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021) 1
5. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: Residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022). https://doi.org/10.1109/TMI.2022.3167808 1
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 1
7. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 1
8. Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., Shen, D.: Medical image synthesis with context-aware generative adversarial networks. In: International conference on medical image computing and computer-assisted intervention. pp. 417–425. Springer (2017) 1
9. Schmidt, M.A., Payne, G.S.: Radiotherapy planning using mri. Physics in Medicine & Biology **60**(22), R323 (2015) 1
10. Violas, P., Estivalezes, E., Briot, J., de Gauzy, J.S., Swider, P.: Objective quantification of intervertebral disc volume properties using mri in idiopathic scoliosis surgery. Magnetic resonance imaging **25**(3), 386–391 (2007) 1
11. Xiang, L., Wang, Q., Nie, D., Zhang, L., Jin, X., Qiao, Y., Shen, D.: Deep embedding convolutional neural network for synthesizing ct image from t1-weighted mr image. Medical image analysis **47**, 31–44 (2018) 1