Swin UNETR Based MRI-to-CT Synthesis

Fuxin Fan[1], Jingna Qiu[2], and Yixing Huang[3]

[1] Pattern Recognition Lab,
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
fuxin.fan@fau.de
[2] Department of Artificial Intelligence in Biomedical Engineering,
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
jingna.qiu@fau.de
[3] Department of Radiation Oncology, University Hospital Erlangen,
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
yixing.yh.huang@fau.de

Abstract. This paper utilizes a state-of-the-art framework, named Swin UNETR, for CT image synthesis from MRI. According to the leaderboard of SynthRad2023, the evaluation metrics MAE, PSNR and SSIM for our model are 62.90 HU, 28.64 and 0.875, respectively, in the validation data set.

Keywords: Swin ViT · CT Synthesis

1 Method and Material

According to the description of SynthRad2023, 540 pairs of MRI and CT volumes from three different institutes are included in this challenge [1, 2]. 360 pairs of them are released for model training. 60 MRI cases without corresponding CT labels are used as the validation set. The rest unreleased 120 pairs are used for the final test. Each volume from brain and neck region has been preprocessed by the organizers to ensure the same voxel size of $1\,mm \times 1\,mm \times 1\,mm$. The voxel size of volumes from pelvis region is rescaled to $1\,mm \times 1\,mm \times 2.5\,mm$. In addition, MRI and CT pairs are registered by organizers. Binary masks for all cases are also provided to give the patient outline segmentation, and the regions within the segmentation are used for evaluation.

1.1 Network Structure

In this work, we use a state-of-the-art network, Swin UNETR [3], for MRI-to-CT synthesis. The implementation of the Swin UNETR is available under the open-source framework MONAI [4]. The architecture of the Swin UNETR is shown in Fig. 1. This network consists of Shift window (Swin) vision transformer (ViT)-based encoder and CNN-based decoder.

A subvolume of size $32 \times 96 \times 96$ is randomly selected from an MRI volume and fed into the network. The Swin UNETR split the subvolume into a sequence

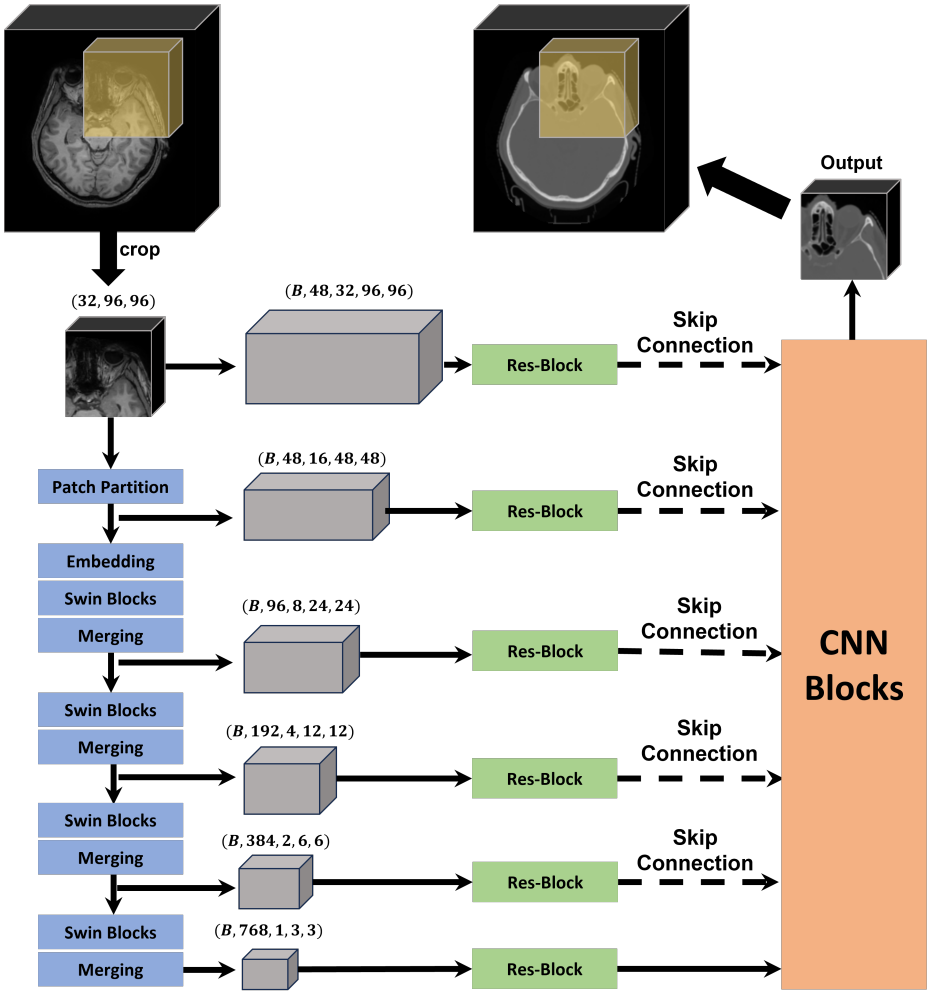

Fig. 1. The network structure of Swin UNETR [3].

of patches, with the size of $2 \times 2 \times 2$. Each patch is embedded into a vector with the feature dimension of 48. The patch sequence then goes through four stages, and each stage has 2 Swin blocks followed by a patch merging operation. After patch merging, the side length of one patch is doubled. At the same time, the output dimension is also doubled. The output from each stage is reshaped and forwarded into a residual block before concatenating with CNN-based blocks. The residual block consists of two $2 \times 2 \times 2$ convolutional layers followed by an instance normalization layer. In each CNN block, the concatenated features are fed into another residual block and a deconvolutional layer. The feature size gets halved after the deconvolutional layer. The final outputs with single channel are computed by using a $1 \times 1 \times 1$ convolutional layer.

1.2 Training Process

Training Data Processing All MRI intensity values are divided by 1000. CT values are subtracted by the minimum value of each volume (e.g., mostly -1024) to be nonnegative, and then divided by 2000. Afterwards, MRI subvolumes with the size of $32 \times 96 \times 96$, together with their corresponding binary masks and CT subvolumes are randomly selected to construct the data for network training. Afterwards, MRI subvolumes are pixel-wisely multiplied with their binary masks as the input of the network.

Network Training The network is trained using an NVIDIA A100 GPU with 80 GB memory. Two models with the same structure in Fig. 1 are trained with data from different body regions separately (i.e., pelvis and brain regions). Each model is trained on all 180 patient cases in each epoch. For each case, 20 subvolumes are randomly selected for each epoch. The predictions and labels are pixel-wisely multiplied with their corresponding binary masks before loss calculation. The L1 loss function and the Adam optimizer are used. The values for β_1 and β_2 are 0.9 and 0.999. The models are trained for 4000 epochs and the learning rate has stepwise decay from 0.0005 to 0.00005.

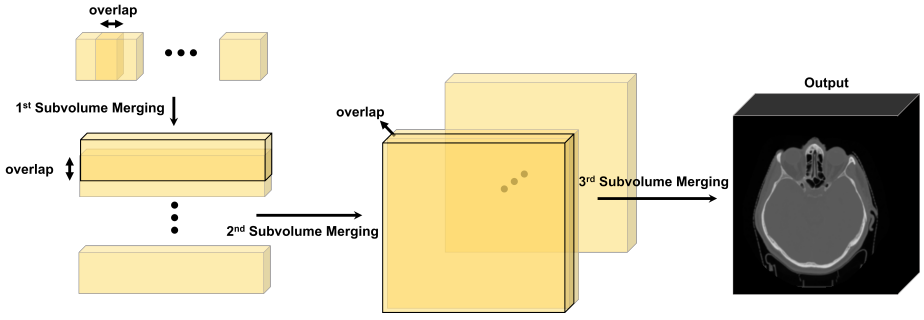

Fig. 2. Full CT volume synthesis by subvolume merging.

Prediction Post-processing To reduce the inference time, only subvolumes within binary masks are predicted. Afterwards, the whole CT volume is constructed by merging adjacent subvolumes. The merging process is shown in Fig. 2. The smallest subvolumes construct long cuboids. Then long cuboids are connected with each other to build flat cubes. The CT volume is then obtained by merging all flat cubes together. The overlap areas of adjacent subvolumes are multiplied with two weight maps to keep smooth intensity transition. The weight for the former subvolume decreases from 1 to 0 along the merging direction, whereas the weight for the latter increases from 0 to 1 complementarily. For CT synthesis in the brain region, the overlapping lengths are set to 28, 72,

and 72 in three dimensions, which correspond to the dimensions in the smallest subvolumes with the size of $32 \times 96 \times 96$.

After subvolume merging, the intensity of the CT volume is scaled back to HU. The intensity multiplies with 2000 and then minors 1023. Meanwhile, the values larger than 3000 are set to 3000, according to the evaluation instructions provided by organizers.

2 Results

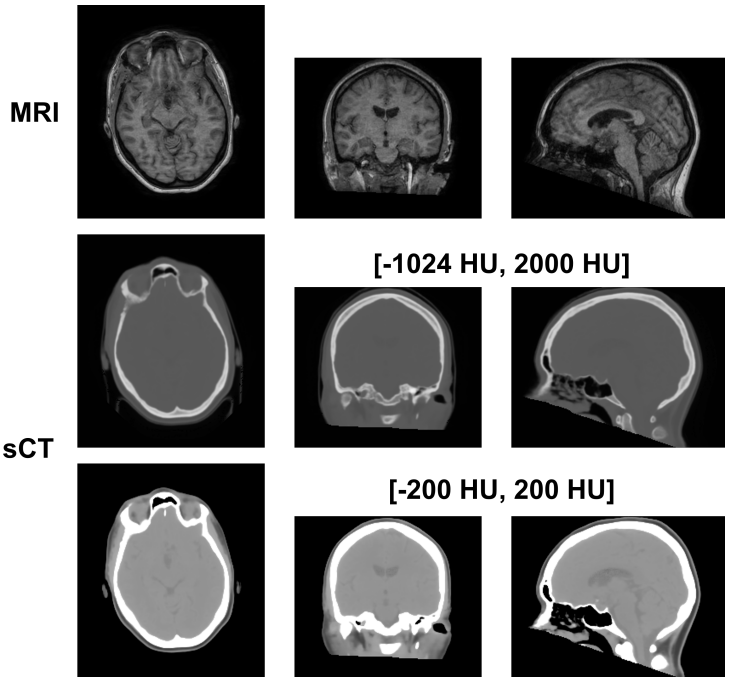

Fig. 3. MRI and synthetic CT (sCT) in brain region. The intensity window for sCT: [-1024, 2000] HU (second row) and [-200, 200] HU (third row).

The prediction results for two cases from the validation dataset are displayed in Fig. 3 and Fig. 4 in three orthogonal views, respectively. Since the label data is not released, we do not include the difference images between the prediction and the ground truth. According to the leaderboard, the MAE, PSNR and SSIM metrics for the sCT in Fig. 3 are 67.88 HU, 28.12 and 0.871, respectively. They are 61.79 HU, 28.47 and 0.868 for Fig. 4, respectively. All in all, for the 60 cases in the leaderboard, the mean MAE, mean PSNR, and mean SSIM are 62.90 HU, 28.64 and 0.875, respectively.

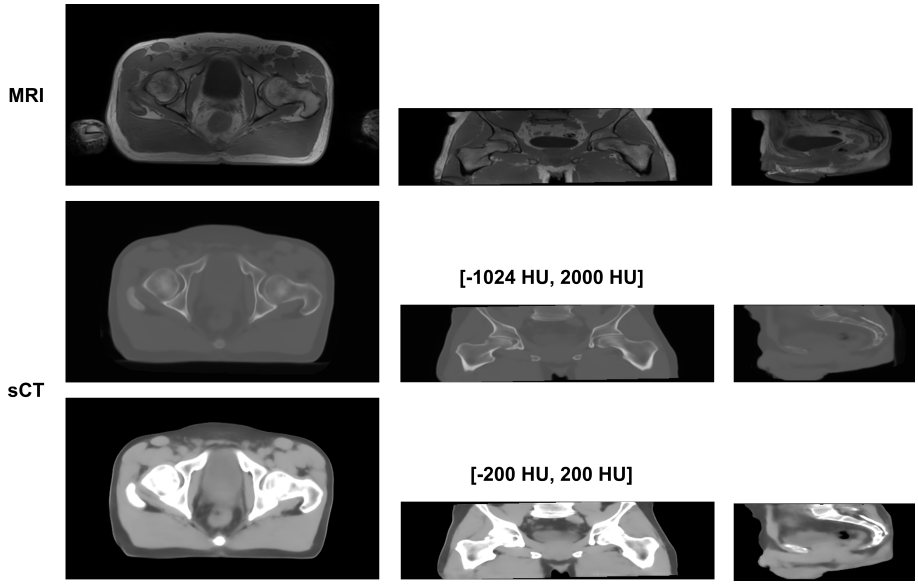

Fig. 4. MRI and synthetic CT (sCT) in pelvis region. The intensity window for sCT: [-1024, 2000] HU (second row) and [-200, 200] HU (third row).

Acknowledgements Thanks for the computational support from Erlangen Regional Computing Center and Erlangen Center for National High Performance Computing.

References

1. Thummerer, A., Bijl, E., Galapon Jr, A., Verhoeff, J., Langendijk, J., Both, S., Berg, C. & Maspero, M. SynthRAD2023 Grand Challenge dataset: Generating synthetic CT for radiotherapy. *Medical Physics*. (2023)
2. Thummerer, A., Huijben, E., Terpstra, M., Gurney-Champion, O., Afonso, M., Pai, S., Koopmans, P., Eijnatten, M., Perko, Z. & Maspero, M. SynthRAD2023 Challenge design: Synthesizing computed tomography for radiotherapy. (Zenodo,2023,4), https://doi.org/10.5281/zenodo.7781049
3. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. & Xu, D. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in mri images. *International MICCAI Brainlesion Workshop*. pp. 272-284 (2021)
4. Cardoso, M., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D. & Others Monai: An open-source framework for deep learning in healthcare. *ArXiv Preprint ArXiv:2211.02701*. (2022)