Synthetic CT generation from CBCT images: Short Paper for SynthRAD 2023

Pengxin Yu[1] (✉)

Infervision Medical Technology Co., Ltd. Beijing, China
ypengxin@infervision.com

Abstract. Medical imaging, especially in radiotherapy, is pivotal for oncological diagnoses and treatments. Traditionally, X-ray imaging has been crucial in radiotherapy (RT) for patient positioning and monitoring at various stages of dose delivery. Cone-beam computed tomography (CBCT) is integral for image-guided adaptive radiation therapy (IGART) in both photon and proton therapies. Yet, artifacts like shading, streaking, and cupping due to scatter noise and truncated projections, hinder CBCT's suitability for precise dose calculations. To address this, "synthetic CT" (sCT) has been introduced, enhancing CBCT quality to CT levels. Transitioning from CBCT to CT permits accurate dose computations, refining adaptive CBCT-based RT and elevating IGART quality. Lately, deriving sCT from CBCT using artificial intelligence, including machine or deep learning, has gained traction. We introduce a deep learning approach with multi-scale residual modules for CBCT-to-sCT generation. In this work, we propose a deep learning method based on multi-scale residual modules for generating sCT from CBCT. In the quantitative evaluation of the MICCAI 2023 SynthRAD Challenge public validation cases, this method achieves the average PSNR of xxxx (Brain PSNR of xxxx, Pelvis PSNR of xxxx), the average SSIM of xxxx (Brain SSIM of xxxx, Pelvis SSIM of xxxx) and the average MAE of xxxx (Brain MAE of xxxx, Pelvis MAE of xxxx).

Keywords: CBCT · Synthetic CT · Deep learning

1 Introduction

Medical imaging, especially in radiotherapy, is pivotal for oncological diagnoses and treatments. Traditionally, X-ray imaging has been crucial in radiotherapy (RT) for patient positioning and monitoring at various stages of dose delivery. Cone-beam computed tomography (CBCT) is integral for image-guided adaptive radiation therapy in both photon and proton therapies. Yet, artifacts like shading, streaking, and cupping due to scatter noise and truncated projections, hinder CBCT's suitability for precise dose calculations. To address this, "synthetic CT" (sCT) has been introduced, enhancing CBCT quality to CT levels. Transitioning from CBCT to CT permits accurate dose computations, refining adaptive CBCT-based RT and elevating IGART quality. Lately, deriving sCT from CBCT using artificial intelligence, including machine or deep learning, has gained traction. In this work, we propose a deep learning method based on multi-scale residual modules for generating sCT from CBCT. Inspired by Ge et al. [1],

we used the consecutive 3D multi-scale residual blocks to richly extract the multi-scale stereo feature for fine-granted and latent spatial structure mining from the CBCT noisy volume. Then, a creative stereo-correlation constraint is used by elegantly penalizing the gradient deviation in the voxel adjacent region (i.e., 3D 26-neighborhoods) for guiding the structural detail. Further, a image-expression constraint on the perceptual feature representations that are transformed from the pretrained deep convolution autoencoder, was added to maintain the scene content. The main contributions of this work are summarized as follows:

1) We proposed a multi-scale residual-based deep learning method for automated generating synthetic CT from CBCT.
2) A stereo-correlation constraint was used to guide the structural details of the generated results, and a image-expression constraint was used to maintain the scene content.
3) The effectiveness of the proposed method are demonstrated on SynthRAD2023 challenge public validation cases, where we achieve the competitive results in terms of quantitative image quality.

2 Dataset and Evaluation Metrics

2.1 Dataset

Description of the dataset. The SynthRAD2023 challenge dataset contains imaging data of patients who underwent radiotherapy in the brain or pelvis region. Overall, the population is predominantly adult and no gender restrictions were considered during data collection. The inclusion criteria was the acquisition of a CT and CBCT, used for patient positioning, were required. Data was collected at 3 Dutch university medical centers: Radboud University Medical Center, University Medical Center Utrecht and University Medical Center Groningen. For anonymization purposes, from here on institution names are substituted with A, B and C, without specifying which institute each letter refers to.

Table 1. Dataset.

	Brain			Pelvis		
	A	B	C	A	B	C
Training	60	60	60	60	60	60
Validation	10	10	10	10	10	10
Testing	20	20	20	20	20	20

Details of data split. For brain and pelvis, each center provides 60 patients for a total amount of 180 training patients per anatomy. For validation and testing, an additional

30/60 patients are available for each anatomy. In total, for all anatomies combined, 540 image pairs (360 training, 60 validation and 120 testing) are available in this dataset, as shown in Table 1. For each anatomy we train separately to obtain the corresponding model. The training dataset (180 cases) was randomly divided into training (165 cases, 55 cases for each center) and internal validation (15 cases, 5 cases for each center) set, where internal validation set is used to model selection.

2.2 Evaluation Metrics

Image similarity will be evaluated by ranking (equal weights) mean absolute error (MAE), peak-signal-to-noise (PSNR), and (structural similarity index) SSIM between sCT and CT. Dosimetric evaluation will be performed globally and locally by comparing photon and proton dose calculations between reference CT and sCT. Relative dose difference, dose-volume histogram, and gamma index will be used to rank the dosimetric evaluation.

3 Method

As mentioned in Figure 1, the input CBCT is first normalized based on the center and anatomical structure, then generated in a sliding window manner, and finally the generated results are concatenated to obtain the final synthetic CT prediction.

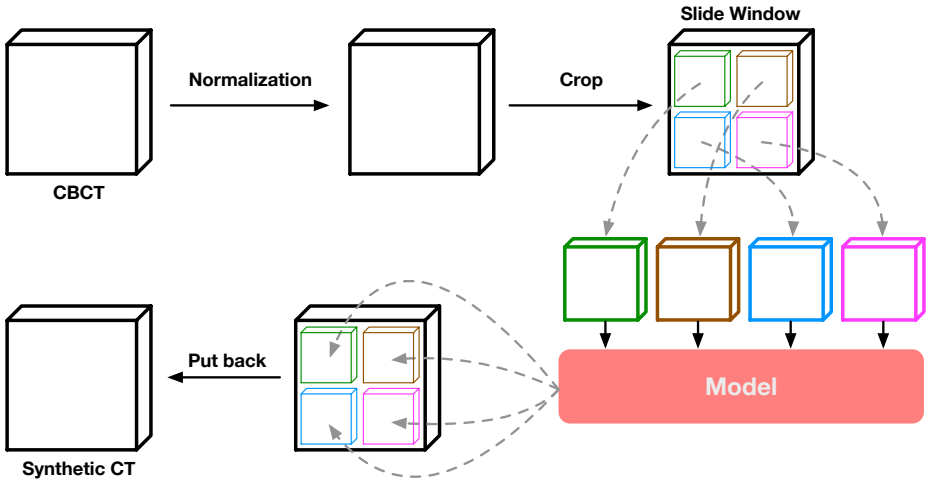

Fig. 1. Framework pipeline

3.1 Normalization

After analyzing the dataset, we found that different centers/anatomies have different intensity distributions. Therefore, we set different windows for different centers/anatomies to perform intensity normalization as follows:

- Brain Center A: [0, 3000];
- Brain Center B / C: [-1000, 2000];
- Pelvis Center A: [0, 2000];
- Pelvis Center B / C: [-1000, 1000];

After windowing, the data is scaled to [0,1] using the max-min normalization method.

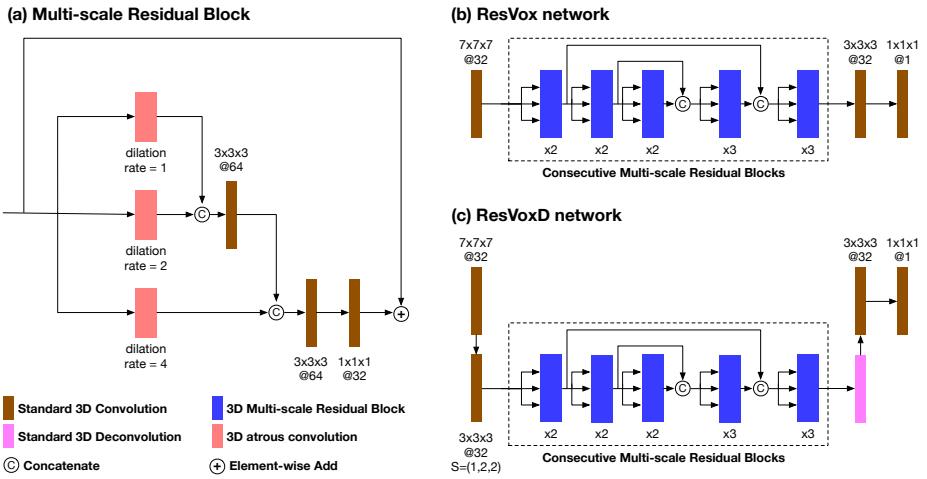

Fig. 2. (a) Multi-scale residual block is constructed by collateral 3D atrous convolutions, hierarchical fusion and residual connection, for fine-granted and latent spatial structure in sparse and noisy volume; (b) ResVox is composed multi-scale residual blocks and two standard 3D convolution layers for the final sCT generation; (c) ResVoxD performs a 2x downsampling after the first convolution layer of ResVox, then passes through consecutive multi-scale residual blocks, then goes through a deconvolution to restore the original size, and finally performs the final sCT generation through two standard 3D convolution layers.

3.2 Proposed Method

Network architecture details. Inspired by Ge et al. [1], we used the consecutive 3D multi-scale residual blocks to richly extract the multi-scale stereo feature for fine-granted and latent spatial structure mining from the CBCT noisy volume. As shown in Figure 2(a), the multi-scale residual block innovatively deploys collateral 3D atrous convolutions, the hierarchical fusion and the residual connection, to mine the fine-granted

and latent spatial structure in the sparse and noisy volume. We designed two architectures, the architecture without downsampling is called ResVox, and the architecture with one downsampling is called ResVoxD, as shown in Figure 2 (b) and (c).

Stereo-Correlation Constraint and Image-Expression Constraint. The stereo-correlation constraint and image-expression constraint are added into the optimization process of our method, to naturally guide the structural detail and scene content in the generated image and improve the generalization of the framework. Specifically, the stereo-correlation constraint is proposed to measures the similarity of the inter-voxel changes, namely stereo gradient, between the generated image and the corresponding CT. Analogously, the image-expression constraint is performed on the high-level feature representations space that encodes the perceptual and semantic information.

1) For the stereo-correlation constraint, the stereo gradient vector is constructed on each voxel $p_{i,j,k}$ with its 26 adjacent voxel $p_{i+a,j+b,k+c}(a,b,c=0,\pm1,and\ a^2+b^2+c^2\neq 0)$ to characterize the voxel correlation of the local changes in the neighborhood region, as:

$$g_{i,j,k}=[p_{i,j,k}-p_{i+a,j+b,k+c}],for\ a,b,c=0,\pm1,and\ a^2+b^2+c^2\neq 0 \quad (1)$$

Then this penalty item on gradient deviation is implemented as Eq.(2), to guide the structural detail. In Eq.(2), $g_{i,j,k}$ and $g'_{i,j,k}$ are from the ground-truth CT $y\in\mathbb{R}^{M_p\times N+p\times D_p}$ and generated results $y'\in\mathbb{R}^{M_p\times N+p\times D_p}$, respectively, and $\|\cdot\|_2$ is L2 norm.

$$L_{SteCor}=\frac{1}{M_pN_pD_p}\sum_{i,j,k}||g_{i,j,k}-g'_{i,j,k}||_2^2 \quad (2)$$

2) For the image-expression constraint, the high-level feature representations is transformed by the pretrained model (PM). Such remarkable reconstruction from the PM feature and the large perceptive field of its each element obviously indicate the intrinsic expression for the global content of the scene. This extracted feature ($\mathbb{R}^{M_f\times N_f\times D_f\times C_f}$) from the PM is used for the image-expression constraint, as:

$$L_{ImgExp}=\frac{1}{M_fN_fD_fC_f}\sum||PM(y)-PM(y')||_2^2 \quad (3)$$

Loss function L1 loss is used as the reconstruction loss combined with the two constraints, and the final loss function is formulated as:

$$Loss=\frac{1}{M_pN_pD_p}\sum_{i,j,k}|y_{i,j,k}-y'_{i,j,k}|+\alpha*L_{SteCor}+\beta*L_{ImgExp} \quad (4)$$

3.3 Post-processing

Ensembling: we selected several models for ensemble based on the results of the internal validation set and the official validation set. Specifically, for each case, we average the prediction results of several models as the final prediction result.

3.4 Environments and requirements

The environments and requirements of the method is shown in Table 2.

Table 2. Environments and requirements.

Windows/Ubuntu version	Ubuntu 18.04.6 LTS
CPU	Intel(R) Xeon(R) Silver 4210R CPU@2.40GHz
RAM	4×32GB
GPU	Nvidia A6000
CUDA version	11.4
Programming language	Python 3.8.10
Deep learning framework	Pytorch (Torch 1.8.1, torchvision 0.9.1)

3.5 Training protocols

The training protocols of the method is shown in Table 3.

Table 3. Training protocols.

Data augmentation methods	Crop, Resize, mirroring.
Initialization of the network	"he" normal initialization
Batch size	1
Patch size	8×180×180
Hyper-parameters in loss	$\alpha = 50, \beta = [0.1, 1]$
Max epochs	1000
Optimizer	AdamW
Initial learning rate	0.0003
Learning rate decay schedule	ReduceLROnPlateau
Training time	72 hours

3.6 Testing protocols

Patch aggregation method: use the sliding window manner for prediction. Window size is the same as the input size, and the step is 2 of z-axis and 32/48 of y-axis and x-axis. Multiple predictions for each voxel are averaged to get the final prediction.

4 Results

The quantitative results of final ensemble model on the internal validation set are shown in Table 4. On the internal validation set, our final ensemble prediction achieved the average MAE of 52.4615, average PSNR of 30.6649, and the average SSIM of 0.9071. On the public validation set, our final ensemble prediction achieved the average MAE of 53.0297 (Brain MAE of 47.7308, Pelvis MAE of 58.3285), average PSNR of 30.7014 (Brain PSNR of 31.8243, Pelvis PSNR of 29.5784), and the average SSIM of 0.9028 (Brain SSIM of 0.9269, Pelvis SSIM of 0.8787).

Table 4. Quantitative results in terms of MAE, PSNR, and SSIM.

	Internal			Public		
	MAE	PSNR	SSIM	MAE	PSNR	SSIM
Brain	48.6591	31.4869	0.9291	47.7308	31.8243	0.9269
Pelvis	56.2639	29.8429	0.8851	58.3285	29.5784	0.8787
All	52.4615	30.6649	0.9071	53.0297	30.7014	0.9028

5 Discussion and Conclusion

In this work, we propose a deep learning method based on multi-scale residual modules for generating sCT from CBCT. During method development we found that anatomy had a strong influence on the results, with brains showing significantly better synthetic quality than pelvis. On the other hand, various noises inherent in CBCT can also affect the synthesis quality. Although the performance is not perfect on some data, our method achieves competitive synthesis performance. We believe that by increasing the number and diversity of training data, better deep learning methods can be developed for synthesizing CT from CBCT.

References

1. Ge, R., Yang, G., Xu, C., Chen, Y., Luo, L., Li, S.: Stereo-correlation and noise-distribution aware resvoxgan for dense slices reconstruction and noise reduction in thick low-dose ct. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 328–338. Springer (2019)