Appendix 1

The architecture of CBCT image content encoder E^{c}_{CBCT} and CT image encoder E_{CT} are shown in *Figure S1*. Each box in the figure corresponds to a multi-channel feature map. The number of channels is denoted on top of the box and the size of 2D feature output is provided at lower left edge of the box. Notably, the light gold box represents a residual blocks group (RBG) composed of three residual blocks. The number below the box is the dilation rates of residual blocks in RBG. Both CBCT image content encoder E^{c}_{CBCT} and CT image encoder E_{CT} have a stack of RBGs. The CBCT image attribute encoder E^{a}_{CBCT} has similar architecture as encoder E^{c}_{CBCT} except that it does not contain any RBG.

The architecture of the feature pyramid decoding of the CBCT image generator G_{CBCT} is shown in *Figure S2*. The generator G_{CBCT} employed the feature pyramid decoding (36) to effectively combine CBCT attribute component with CT anatomical information (or CBCT content component). Feature fusion is performed before the first two up-sampling layers and the final series of convolution layers. Notably, the CT image generator G_{CT} does not use this feature pyramid decoding and CT anatomical information (or the CBCT content component) is the only input at its decoding phase.

The architecture of the RBG is shown in *Figure S3*, where the number in each box is the number of features for corresponding map and C is the number of features for the input feature map. Notably, a RBG consists of three succeeding residual blocks with different dilation rates. The dilation rate of entire residual block is represented by the dilation rate of dilated convolution in its middle convolution layer. The different dilation rates (1, 2, 3) is applied in three residual blocks of a RBG to meet the requirements of hybrid dilated convolution (HDC) (34). The HDC enables deeper layers of the network to access information from a larger range of pixels while keep anatomical information for each pixel. The pre-activation architecture of the residual unit is introduced by (37).

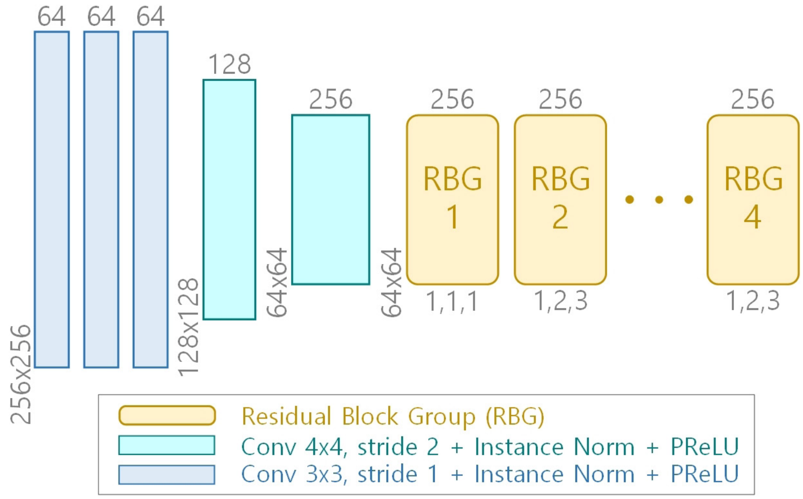
64
64
64
128
256
256
256
256
RBG 1
RBG 2
• • •
RBG 4
256x256
128x128
64x64
64x64
1,1,1
1,2,3
1,2,3
Residual Block Group (RBG)
Conv 4x4, stride 2 + Instance Norm + PReLU
Conv 3x3, stride 1 + Instance Norm + PReLU

Figure S1 The architecture of CBCT image content encoder E^{c}_{CBCT} and CT image encoder E_{CT}.

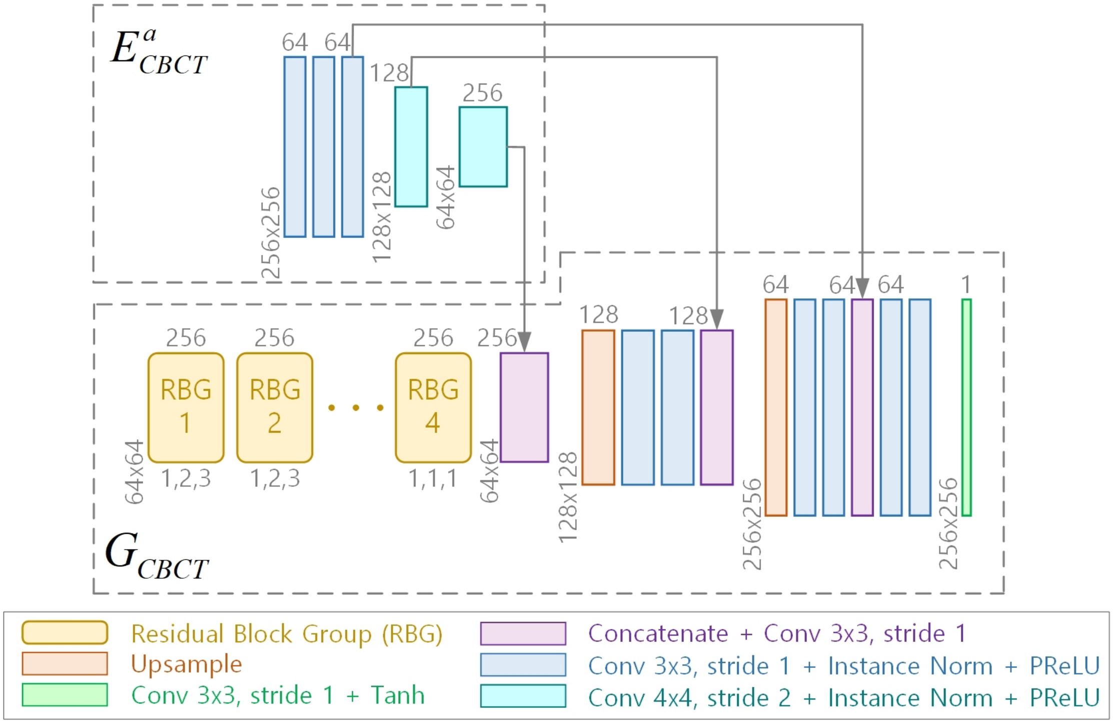
E^a_{CBCT}
64
64
128
256
256x256
128x128
64x64
256
256
256
256
RBG 1
RBG 2
RBG 4
1,2,3
1,2,3
1,1,1
64x64
64x64
128
128
128x128
64
64
64
1
256x256
256x256
G_{CBCT}
Residual Block Group (RBG)
Upsample
Conv 3x3, stride 1 + Tanh
Concatenate + Conv 3x3, stride 1
Conv 3x3, stride 1 + Instance Norm + PReLU
Conv 4x4, stride 2 + Instance Norm + PReLU

Figure S2 The architecture of the feature pyramid decoding of the CBCT image generator G_{CBCT}.

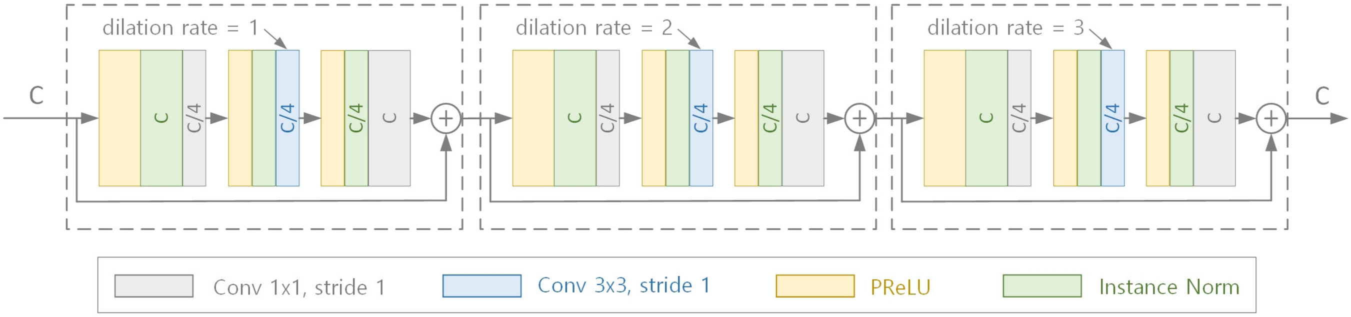

Figure S3 The architecture of the residual block group in encoders and generators.

https://dx.doi.org/10.21037/qims-20-1056

Appendix 2

The mean structural similarity index (SSIM) is defined by:

$$SSIM(x,y)=\frac{\left(2\mu_x\mu_y+c_1\right)\left(2\sigma_{xy}+c_2\right)}{\left(\mu_x^2+\mu_y^2+c_1\right)\left(\sigma_x^2+\sigma_y^2+c_2\right)} \qquad [1]$$

Where x and y are the sCT image and dpCT image, respectively. μ_x and μ_x^2 are the average and variance of image x, while μ_y and μ_y^2 are the average and variance of image y. σ_{xy} is the covariance of images x and y. c_1 and c_2 are constants used to avoid instability. The side-length of the sliding window used in the computation was set to 7.

The peak signal-to-noise ratio (PSNR) is defined by:

$$PSNR(x,y)=20\log_{10}\frac{MAX_I}{\sqrt{MSE(x,y)}} \qquad [2]$$

Where *MSE*(*x*,*y*) represents the mean square error between sCT image x and dpCT image y. MAX_1 represents the maximum value of the quantization of the pixel values of images x and y. This value was set to 3000, as all samples were clipped to [–1000, 2000] HU.

MAE is the mean value of absolute errors. It measures the magnitude of the difference between two images. MAE is defined by:

$$MAE(x,y)=\frac{1}{mn}\sum_{i,j}^{mn}\left|\left(x(i,j)-y(i,j)\right)\right| \qquad [3]$$

Where *x*(*i*,*j*) and *y*(*i*,*j*) are the value of pixels in sCT and dpCT images, respectively. *mn* is the total number of pixels.

RMSE is the square root of mean value of the squared deviations between the observed and the true values. It reflects the deviation between two images. RMSE is defined by:

$$RMSE(x,y)=\sqrt{\frac{1}{mn}\sum_{i,j}^{mn}\left(x(i,j)-y(i,j)\right)^2} \qquad [4]$$

Where *x*(*i*,*j*) and *y*(*i*,*j*) are the values of pixels in sCT and dpCT images, respectively, *mn* is the total number of pixels.

References

36. Lin TY, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. 2017 The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017:2117-25.

37. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. 2016 The European Conference on Computer Vision (ECCV) 2016:630-45.