frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 08 August 2024
DOI 10.3389/fonc.2024.1440944

Check for updates

OPEN ACCESS

EDITED BY
Xiaodong Wu,
The University of Iowa, United States

REVIEWED BY
Jianxin Chen,
Nanjing University of Posts and Telecommunications, China
Pranjal Sahu,
Siemens Healthineers, United States

*CORRESPONDENCE
Zhiwen Liang
✉ liangzhiwen@hust.edu.cn

†These authors have contributed equally to this work

RECEIVED 30 May 2024
ACCEPTED 19 July 2024
PUBLISHED 08 August 2024

CITATION
Hu Y, Cheng M, Wei H and Liang Z (2024) A joint learning framework for multisite CBCT-to-CT translation using a hybrid CNN-transformer synthesizer and a registration network.
Front. Oncol. 14:1440944.
doi: 10.3389/fonc.2024.1440944

A joint learning framework for multisite CBCT-to-CT translation using a hybrid CNN-transformer synthesizer and a registration network

Ying Hu[1,2†], Mengjie Cheng[3†], Hui Wei[4] and Zhiwen Liang[5,6*]

[1]School of Mathematics and Statistics, Hubei University of Education, Wuhan, Hubei, China, [2]Bigdata Modeling and Intelligent Computing Research Institute, Hubei University of Education, Wuhan, Hubei, China, [3]Nutrition Department, Renmin Hospital of Wuhan University, Wuhan, China, [4]Department of Radiotherapy, Affiliated Hospital of Hebei Engineering University, Handan, China, [5]Cancer Center, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [6]Hubei Key Laboratory of Precision Radiation Oncology, Wuhan, China

Background: Cone-beam computed tomography (CBCT) is a convenient method for adaptive radiation therapy (ART), but its application is often hindered by its image quality. We aim to develop a unified deep learning model that can consistently enhance the quality of CBCT images across various anatomical sites by generating synthetic CT (sCT) images.

Methods: A dataset of paired CBCT and planning CT images from 135 cancer patients, including head and neck, chest and abdominal tumors, was collected. This dataset, with its rich anatomical diversity and scanning parameters, was carefully selected to ensure comprehensive model training. Due to the imperfect registration, the inherent challenge of local structural misalignment of paired dataset may lead to suboptimal model performance. To address this limitation, we propose SynREG, a supervised learning framework. SynREG integrates a hybrid CNN-transformer architecture designed for generating high-fidelity sCT images and a registration network designed to correct local structural misalignment dynamically during training. An independent test set of 23 additional patients was used to evaluate the image quality, and the results were compared with those of several benchmark models (pix2pix, cycleGAN and SwinIR). Furthermore, the performance of an autosegmentation application was also assessed.

Results: The proposed model disentangled sCT generation from anatomical correction, leading to a more rational optimization process. As a result, the model effectively suppressed noise and artifacts in multisite applications, significantly enhancing CBCT image quality. Specifically, the mean absolute error (MAE) of SynREG was reduced to 16.81 ± 8.42 HU, whereas the structural similarity index (SSIM) increased to 94.34 ± 2.85%, representing improvements over the raw CBCT data, which had the MAE of 26.74 ± 10.11 HU and the SSIM of 89.73 ± 3.46%. The enhanced image quality was particularly beneficial for organs with low contrast resolution, significantly increasing the accuracy of automatic segmentation in these regions. Notably, for the brainstem, the mean Dice similarity coefficient (DSC) increased from 0.61 to 0.89, and the MDA decreased from 3.72 mm to 0.98 mm, indicating a substantial improvement in segmentation accuracy and precision.

Conclusions: SynREG can effectively alleviate the differences in residual anatomy between paired datasets and enhance the quality of CBCT images.

KEYWORDS

CBCT, synthetic CT, deep learning, hybrid transformer, adaptive radiotherapy

1 Introduction

During radiotherapy, weight loss, tumor shrinkage and anatomical deformation may cause unwanted dose distribution and degrade the precision of dose delivery (1). Cone-beam computed tomography (CBCT) seems to be the most convenient way to obtain 3D anatomical information on the day of treatment. The role of recalculating the dose distribution and evaluating the necessity of replanning during CBCT is essential for adaptive radiation treatment (ART). However, cone beams generate a large amount of scatter in projection images, which results in severe artifacts, including cupping, shading, streaks, and inhomogeneities, hence reducing Hounsfield unit (HU) accuracy (2).

Many traditional methods have been introduced to improve the quality of CBCT images, including antiscatter accessories (3), scatter correction (4) and iterative reconstruction (5). In recent years, a commercial algorithm named Acuros CTS was proposed by Varian Medical Systems for clinical applications (6, 7). They corrected scatter by calculating primary and scatter images in the projection domain, followed by performing FDK-based reconstruction and statistical iterative reconstruction, and obtained clearer images and more accurate HU values. However, the direct use of CBCT in the adaptive pathway is still limited by the fact that the image quality of CBCT is considered significantly inferior to that of planning CT (pCT) in terms of the contrast-to-noise ratio and imaging artifacts (8, 9).

Recently, researchers have focused on improving the quality of CBCT images via convolutional neural networks (CNNs). Jiang et al. (10) proposed a deep residual CNN (DRCNN), which uses a residual U-Net framework, to learn the mapping function between scatter CBCT and scatter-free CBCT. Li et al. (11) utilized the DRCNN to convert CBCT images to synthetic CT (sCT) images for nasopharyngeal carcinoma (NPC) patients, maintaining the anatomical structure information of the CBCT images while correcting the HU distribution, similar to pCT images. Liang et al. (12) introduced a cycle-consistent generative adversarial network (cycleGAN) to generate sCT images for head and neck patients. Subsequently, cycleGANs have been used for patients with pelvic and/or prostate cancer (13, 14).One common deficiency of the abovementioned CNN-based models is their disregard of the global pixel relationships within images, which is primarily due to the limited receptive fields. These global relationships play crucial roles in achieving high-quality image restoration (15). To address this problem, the transformer architecture has recently been introduced to computer vision (16), offering the ability to model long-range dependencies and nonlocal information. Vision transformers (ViTs) (17) divide images into patches and employ multihead self-attention (SA) mechanisms to capture the relationships among patches. Chen et al. (18) obtained superior performance to that of a cycleGAN in the CBCT-to-CT translation task by using a transformer-based network. Nevertheless, the SA mechanisms of ViTs lead to quadratic computational complexity with respect to the image size, which poses challenges for low-level tasks that typically handle high-resolution images. Moreover, while ViTs excel in terms of capturing the global context and long-range dependencies, they may struggle to capture fine-grained local details and high-frequency components such as image edges. Additionally, a ViT typically requires larger datasets and a more extensive training process than other methods do for optimal generalization (19).

Another challenge is the local structural misalignment in paired datasets used for supervised learning. Rossi et al. (20) reported that the supervised learning approach can obtain better quantitative evaluation results but produces more blur and artifacts in qualitative evaluations, which is due to the higher sensitivity of the supervised training process to the pixelwise correspondence contained in the loss function. In practice, limited by the utilized scanning system or ethics, we usually cannot obtain paired images with perfect pixelwise matches from two modalities. To minimize the differences between paired images, previous studies (21, 22) have applied deformable image registration (DIR) to compensate for the anatomical mismatches resulting from patient position differences and potential internal anatomical changes. However, limited by the ability of DIR, the resulting datasets do not represent ideal pixelwise paired images and may introduce uncertainties in the training and evaluation processes of the constructed networks.

In this paper, we introduce SynREG to address the challenges encountered in sCT generation scenarios. Our approach combines a hybrid CNN-transformer synthesizer to capture both local and global information and a U-Net-based registration network to correct residual anatomy mismatches in the training pairs. By utilizing a supervised learning strategy, SynREG is trained on diverse anatomical datasets, which allows it to produce high-quality sCT images across multiple sites.

2 Materials and methods

2.1 Data collection and processing

Data from 135 patients with abdominal cancer, chest cancer or head and neck cancer were collected for training purposes in this study. Planning CT (pCT) and CBCT images were obtained from a CT simulator (Philips Medical Systems, Cleveland, OH, USA) and a kV CBCT system integrated on the Halcyon 2.0 system (Varian Medical Systems, Palo Alto, CA, USA), respectively. All CBCTs were scanned with a half-bowtie filter and reconstructed by the traditional filtered backprojection method with a 2-mm slice thickness, followed by our clinical scanning protocol. Detailed information about the scanning parameters is listed in Table 1. Deformable registration was implemented using MIM software (v.7.0.1, MIM Software Inc., Cleveland, OH, USA) to pair the pCT images with the CBCT images. The deformed CT volumes were resampled to the corresponding CBCT voxel spacing and then cropped to the CBCT dimensions and number of slices. Finally, a large dataset with 10,084 image pairs was used for training the model. In addition, data from an additional 23 patients with image pairs were collected for independent testing.

2.2 SynREG framework

We present the overall framework of our proposed SynREG algorithm in Figure 1. In our setup, each training sample consists of a pair of CBCT and pCT images, both with dimensions of 256x256. The CBCT image is initially passed through a synthesizer to generate an sCT image. Subsequently, a registration subnetwork (Reg-net) is employed to calculate the deformation vector field (DVF) between the sCT and pCT images. This allows for the manipulation of sCT to align with pCT. The synthesizer and Reg-Net are trained together using batches of paired CBCT and pCT images, ensuring optimal performance. In the following sections, we provide more detailed information on our model and the implemented loss functions.

2.2.1 Hybrid CNN-transformer synthesizer

Due to similar physical processes, CBCT can be viewed as a potentially degraded version of a CT image. Hence, choosing the most critical features while eliminating undesirable features in the channel dimension is crucial for noise suppression and artifact removal. Inspired by Restormer (23), we employ SA across the feature dimension instead of the spatial dimension to construct the fundamental transformer block. Consequently, we introduce a hybrid CNN-transformer synthesizer that incorporates a stack of nine depth convolution-based transformer blocks (DTBs) organized in a UNet architecture (24) (Figure 2).

Given a CBCT image $I_{CBCT} \in \mathbb{R}^{H\times W\times 1}$, the synthesizer first applies a 1×1 convolution to obtain low-level feature maps $F_0 \in \mathbb{R}^{H\times W\times 32}$, where $H \times W$ represents the spatial resolution. Subsequently, the encoding path of UNet extracts these shallow features F_0 through four consecutive layers of convolution and downsampling. The features extracted at each layer are relayed to the corresponding layers of the decoding path via skip connections, whereas the bottom-level features are passed to the stack of DTBs. With this design, the skip connections effectively facilitate the high-frequency features to the decoder, whereas the DTB bottleneck serves as an effective approach for learning pairwise relationships among low-frequency features.

A DTB consists of two fundamental components: a multihead depthwise convolution-transposed attention (MDTA) module and a multiscale feedforward network (MSFN), as shown in Figure 3. Within the architecture, the MDTA module applies SA across channels to compute the cross covariance across the channels and generate an attention map that implicitly encodes global context information. This attention map is then used to weight the feature maps, allowing the model to focus on the most relevant information. Figure 3A illustrates the architecture of a single-head DTA, which initially encodes channelwise context through 1×1 convolutions, followed by 3×3 depthwise convolutions to capture spatial local context within each channel. The MDTA extends this foundation by utilizing multiple parallel heads, each of which independently focuses on distinct parts of the input. Then, SA across the channels is applied to generate attention. MDTA has linear complexity, hence reducing the temporal and memory complexity of the network. The attention mechanism is generally formulated as Equation (1).

$$Att(Q, K, V) = softmax(\frac{QK^T}{\lambda})V \tag{1}$$

where λ is a learnable scaling parameter that controls the magnitude of the dot product of Q and K.

The MSFN module (Figure 3B) is applied after the MDTA module, and its effectiveness has been verified by Chen et al. (25). It consists of two multiscale local information extraction operations. After performing layer normalization, a 1x1 convolution is applied to expand the channel dimensionality. Then, the expanded features are fed into two parallel branches, in which 3x3 and 5x5 depthwise convolutions are employed to enhance the multiscale local information extraction process. The extracted features derived from both branches are subsequently concatenated. After another multiscale information extraction operation, a 1x1 convolution is used to keep the size of the output tensor matched with that of the tensor that was initially fed into the MSFN.

TABLE 1 CBCT scanning parameters used for head, thorax and pelvis patients.

CBCT Mode	Energy (kV)	Exposure (mAs)	CTDIvol (mGy)	DLP (mGy*cm)	Scan time (Sec)	Scan diameter (cm)
Head	100	126	3.33	49.9	16.6	28.2
Thorax	125	294	5.88	88.2	30.8	49.2
Pelvis	125	1080	21.6	324	36.7	49.2

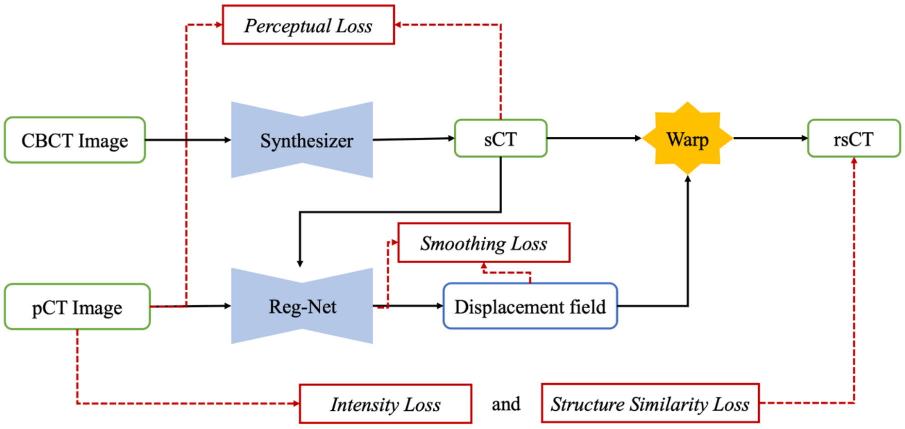


FIGURE 1
The overall framework of the proposed SynREG approach mainly includes a synthesizer and a registration network for achieving enhanced image quality and correcting residual anatomical mismatches.

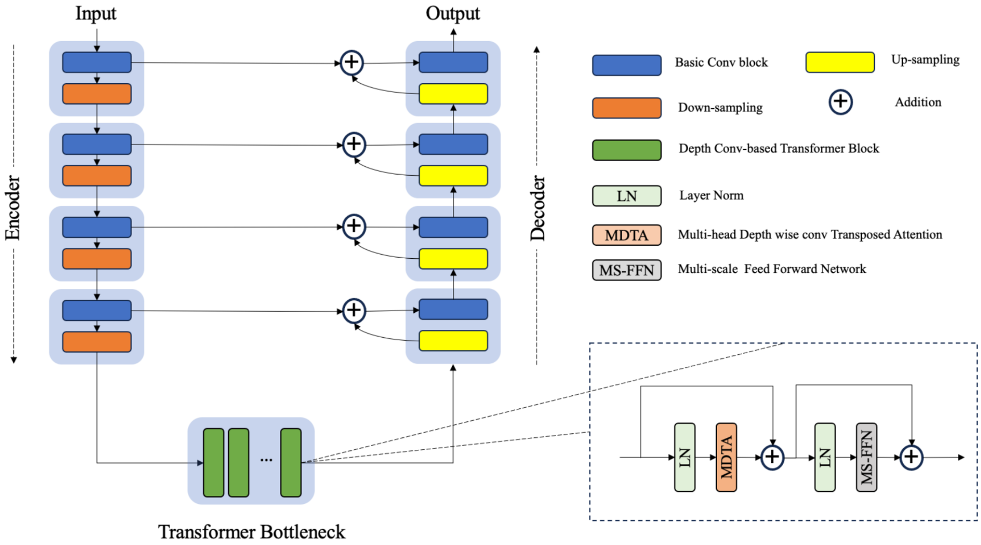
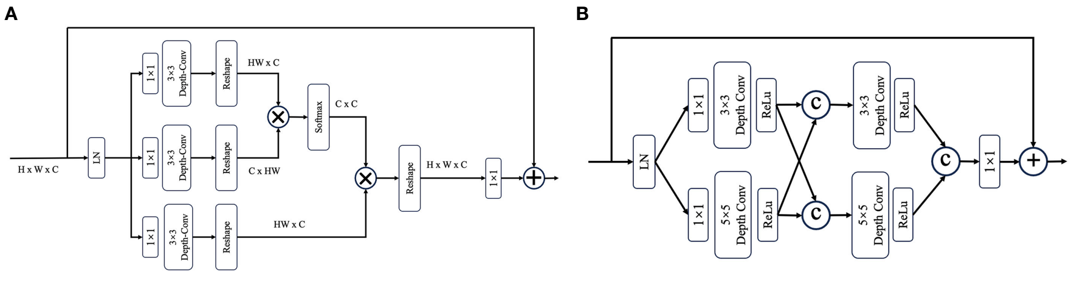


FIGURE 2
The architecture of the hybrid CNN-transformer synthesizer. The synthesizer was constructed on the basis of a U-net structure composed of an encoder, a transformer bottleneck and a decoder. Each transformer block contains a multihead depthwise convolution transposed attention module (MDTA) and a multiscale feed forward network module (MS-FFN).



Depthwise Convolution Transposed Attention

Multi-Scale Feed Forward Network

FIGURE 3
The architecture of **(A)** the depthwise convolution transposed attention (DTA) module and **(B)** the multiscale feed forward modules (MS-FFN). Where ⊗ and © refer to the multiplication and concatenation operations, respectively.

2.2.2 Registration network

The Reg-Net employed in this study is based on the work of Kong et al. (26). Its objective is to acquire prior knowledge about the DVF from the input sCT and pCT images. The DVF represents the displacement of each pixel, and by warping the sCT using the calculated DVF, the resulting registered sCT (referred to as rsCT) can be optimized to minimize its differences from the pCT via the pixelwise intensity loss function.

Reg-Net is a modified version of the U-Net architecture that consists of seven downsampling blocks, three residual blocks, and seven upsampling blocks. In each downsampling block, features are extracted at various levels with different numbers of filters, namely, 32, 64, 64, 64, 64, 64, 64 and 64. The upsampling process is the counterpart of the downsampling process and incorporates skip connections to collect the corresponding blocks at each level. Finally, Reg-Net outputs DVFs across the horizontal and vertical dimensions, ensuring accurate reconstruction of a high-resolution DVF representation.

2.2.3 Loss functions

The synthesizer employs a perceptual loss for computing the feature similarity between sCT and pCT images at multiple levels. To extract deep multilevel features and structural information, we introduce the deep image structure and texture similarity (DISTS) index as the perceptual loss because it unifies texture similarity and structural similarity into a single index. The loss function is formulated as Equation (2).

$$\mathcal{L}_{perceptual} = D(x, y; \alpha, \beta) \\ = 1 - \sum_{i=0}^{m} \sum_{j=1}^{n_i} \left(\alpha_{ij} l\left(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}\right) + \beta_{ij} k(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) \right) \quad (2)$$

where x and y represent the sCT and pCT images, respectively. i represents the convolution layers, and j represents the channel in the i th convolution layer. α_{ij} and β_{ij} are positive weights, which are pretrained via a variant of the visual geometry group (VGG) network. $l(\overset{\ddot{A}}{n})$ and $k(\overset{\ddot{A}}{n})$ are the defined texture similarity and structure similarity, respectively. The details of the DISTS index were described by Ding et al. in 2020 (27).

Reg-Net has three loss functions, including intensity loss, structural similarity loss and smoothing loss. Here, we use the Charbonnier loss (28) as the intensity loss, which compares the intensity difference between the rsCT and pCT images (referred to as x and y, respectively) and is formulated as Equation (3).

$$\mathcal{L}_{intensity} = \sqrt{||y - x||^2 + e^2} \quad (3)$$

where e is a constant that is set to 10^{-3}.

Structural similarity is measured using locally normalized cross-correlation (LNCC) (29), which emphasizes the anatomical similarity between rsCT and pCT images, and defined as Equation (4).

$$\mathcal{L}_{structure} = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - \mu_{x_i})(y_i - \mu_{y_i})}{\sigma_{x_i} \sigma_{y_i}} \quad (4)$$

where N is the number of samples and where $(\mu_{x_i}, \ \mu_{y_i})$ and $(\sigma_{x_i}, \ \sigma_{y_i})$ denote the means and standard deviations of x_i and y_i, respectively.

The smoothing loss is defined in Equation 5 to evaluate the smoothness of the deformation field and minimize its gradient.

$$\mathcal{L}_{smooth} = \mathbb{E}_{x,\ y}[\| \nabla R(x, \ y) \|^2] \quad (5)$$

The total loss of the proposed SynREG approach is Equation (6).

$$\mathcal{L} = \mathcal{L}_{perceptual} + \lambda_1 \mathcal{L}_{intensity} + \lambda_2 \mathcal{L}_{structure} + \lambda_3 \mathcal{L}_{smooth} \quad (6)$$

3 Experiments

3.1 Implementation details

The CBCT/CT image pairs obtained from 136 patients were randomly divided into a training set and a validation set at a ratio of 0.9 to 0.1. The training set comprised 9,076 pairs of images, whereas the validation set consisted of 908 pairs. Both the CBCT and CT images had an HU value threshold range set to [-1000, 2200], with any values outside this range being set to the nearest threshold values. The HU values were subsequently normalized and mapped to the range of (-1, 1). During the training process, a patch with 256x256 dimensions was randomly cropped from each processed image and used as a network input. Additionally, data augmentation techniques such as random flipping and rotation were applied with a probability of 0.3.

The adaptive moment estimation (Adam) optimizer was employed for optimization with the momentum parameters set to $\beta 1 \ = \ 0.5$ and $\beta 2 \ = \ 0.999$. A superconvergence cosine annealing strategy with a warm-up learning rate was implemented during training (30). Initially, the learning rate was set to 0.0001, and it gradually increased to a maximum of 0.1 at epoch 50. Then, it gradually decreases to zero by epoch 200 following a cosine function.

During the training process, λ_1, λ_2 and λ_3 in the loss function were empirically set to 5, 1 and 1, respectively. The intensity loss between the pCT and rsCT images was calculated for the validation data every 10 epochs. The model that achieved the minimum intensity loss was saved as the best model.

3.2 Image quality evaluation metrics

To quantitatively evaluate the image quality of the images generated by each model in comparison with the reference pCT images, we employed commonly used metrics such as the mean absolute error (MAE), root mean square error (RMSE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM). These metrics are defined by Equations (7)–(10).

$$MAE(I_1, I_2) = \frac{1}{n_i n_j} \sum_{x,y}^{n_i n_j} |I_1(x, y) - I_2(x, y)| \quad (7)$$

$$RMSE(I_1, I_2) = \sqrt{\frac{1}{n_i n_j} \sum_{x,y}^{n_i n_j} |I_1(x,y) - I_2(x,y)|^2} \quad (8)$$

$$PSNR(I_1, I_2) = 10 \times log_{10}(\frac{P^2}{RMSE(I_1, I_2)^2}) \quad (9)$$

$$SSIM(I_1, I_2) = \frac{(2\mu_{I_1}\mu_{I_2} + c_1)(2\sigma_{I_1,I_2} + c_2)}{(\mu_{I_1}^2 + \mu_{I_2}^2 + c_1)(\sigma_{I_1}^2 + \sigma_{I_2}^2 + c_2)} \quad (10)$$

where I_1 and I_2 represent two different images used for comparison purposes; $I(x, y)$ is the HU value of pixel (x, y) in image I; $n_i n_j$ is the total number of pixels in image I; P is the maximum HU range of the image; and μ σ, c_1 and c_2 are the same as those defined above.

3.3 Segmentation evaluation

Automatic segmentation is an important aspect of clinical work that can improve the efficiency of the ART workflow. In this study, a commercial AI-based autocontour module of UIH TPS (v.1.0, United Imaging Healthcare Co., Shanghai, China) and a well-known open-source tool, TotalSegmentator (TS) (31), whose accuracy and robustness have been tested on diverse datasets, were adopted to evaluate the segmentation results. Considering the limited field of view (FOV) of CBCT, we selected the brainstem and parotids from the head cases and the bladder and rectum from the pelvis cases for testing. We generated automatic segmentations on the pCT, CBCT and sCT images. The segmented pCT contours were regarded as the ground truths, and the contours from the other image modalities were compared. The Dice similarity coefficient (DSC) and mean distance to agreement (MDA) were used to evaluate the segmentation accuracy. A higher DSC and lower MDA indicate better consistency between the segmented contours and the ground truths.

3.4 Statistical analyses

To determine if the data from the two groups were significantly different, we adopted the paired t test if the data were normally distributed; otherwise, the Wilcoxon signed-rank test, a nonparametric test for paired samples, was adopted. A statistical significance level of $p < 0.05$ was used.

4 Results

4.1 Comparison with other benchmark models

SynREG was compared with three other image benchmark models: pix2pix (32), cycleGAN (33) and SwinIR (34). The results demonstrate that our method outperforms these benchmarks, exhibiting a remarkable ability to generate high-quality sCT images that capture intricate textures and faithfully preserve anatomical structures. As evidenced by the yellow arrows in Figure 4, the sCT images generated by SynREG show the detailed texture of the bronchi and the precise shape of the tumor, both of which are crucial for accurate clinical diagnosis and tumor delineation.

Table 2 presents the quantitative results on the test dataset, revealing significant improvements in both the MAE and SSIM, with p values less than 0.01. Furthermore, Figure 5 illustrates the performance metrics for individual sites, demonstrating substantial reductions in the MAE and RMSE, along with notable increases in the SSIM for all sites. This finding reveals the generalizability of our model.

Figure 6 highlights the visual improvements achieved by SynREG for multisite cases. The original CBCT image exhibits severe noise, spatial nonuniformity and various artifacts, including beam hardening artifacts and streak artifacts. However, the sCT images generated by our method exhibit remarkable visual

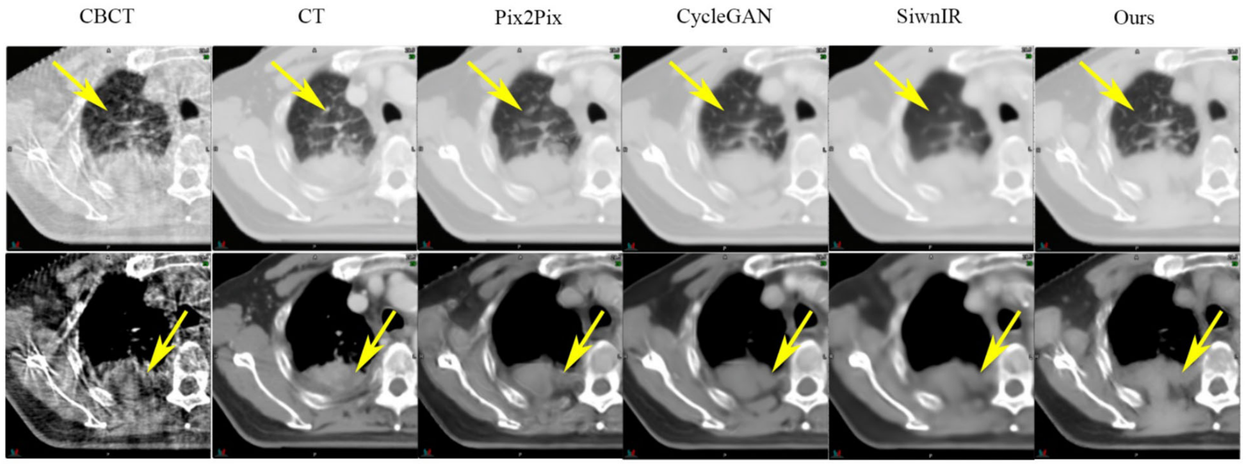


FIGURE 4
Comparison of the sCT images generated by the SynREG model and other benchmark models. The yellow arrows highlight the areas with apparent visual differences. The upper row shows an example slice in the lung window/level, whereas the lower row shows the same slice in the soft tissue window/level.

TABLE 2 Quantitative comparison of the test dataset among different benchmark models.

	MAE (HU)	RMSE (HU)	PSNR (dB)	SSIM (%)
CBCT	26.74 ± 10.11	87.17 ± 27.43	33.86 ± 2.79	89.73 ± 3.46
Pix2Pix	18.17 ± 8.29	64.45 ± 27.23	35.53 ± 3.07	91.97 ± 2.89
CycleGAN	18.32 ± 8.66	66.84 ± 27.64	36.18 ± 3.06	93.60 ± 2.97
SwinIR	17.97 ± 8.52	66.49 ± 28.31	36.40 ± 3.18	93.65 ± 2.98
Ours	16.81 ± 8.42	64.10 ± 27.82	36.59 ± 3.12	94.34 ± 2.85

The reported values are the average ± STD results.

performance, effectively reducing noise and eliminating artifacts. This finding demonstrates the robustness and effectiveness of our proposed method in generating clinically relevant sCT images.

4.2 Ablation experiments

To investigate the impact of Reg-Net and varying loss combinations on model performance, we evaluated four distinct configurations: M1, M2, M3 and M4. M1 serves as a baseline, employing only the synthesizer subnetwork and the L1 loss; M2 incorporates the SynReg architecture in conjunction with the L1 loss; M3 incorporates the SynReg with perceptual loss; and M4, our proposed method, leverages both perceptual loss and L1 loss within the SynReg framework. The quantitative results from these experiments are presented in Table 3. Reg-Net contributes to improving the MAE and SSIM of the model by mitigating local structural misalignments. The perceptual loss enhances the PSNR and SSIM by generating high-fidelity images, though it may not directly contribute significantly to reducing the MAE. Our proposed method, M4, integrates all of these components effectively, achieving superior results.

4.3 Training on an individual dataset versus the entire dataset

When focusing on the head and neck dataset alone, our model trained on this dataset achieved a mean MAE of 14.18 HU, significantly reducing the intensity error for those cases. However, when applied to the thorax and pelvis cases, no MAE reduction was observed, with values of 38.84 HU and 28.14 HU, respectively. This highlights the model's limited generalizability when trained on a single dataset. Conversely, the model trained on the entire dataset (synREG) consistently improved the MAE across anatomical sites. Notably, it achieved a lower MAE of 13.67 HU for head and neck cancer patients, emphasizing the importance of diverse and representative data for robust, generalizable models.

4.4 HU calibration

Figure 7A shows the HU calibration performance. By referring to the pCT image as the reference image, the HU difference relative to the sCT image was significantly improved. In the high-frequency areas (i.e., edges) of the sCT image, the HU differences were greater than those in other areas, indicating intrinsic anatomical differences. The HU profiles of the yellow line in Figure 7A

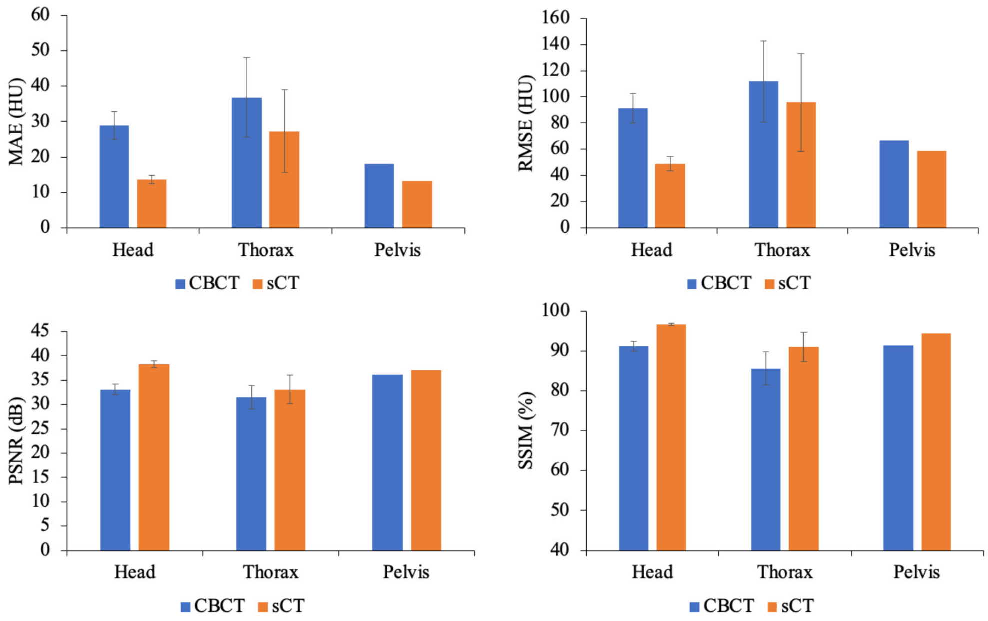


FIGURE 5
Comparison of the MAEs, RMSEs, PSNRs and SSIMs produced for the CBCT and sCT images at individual sites, including the head, thorax and pelvis, in the test dataset.

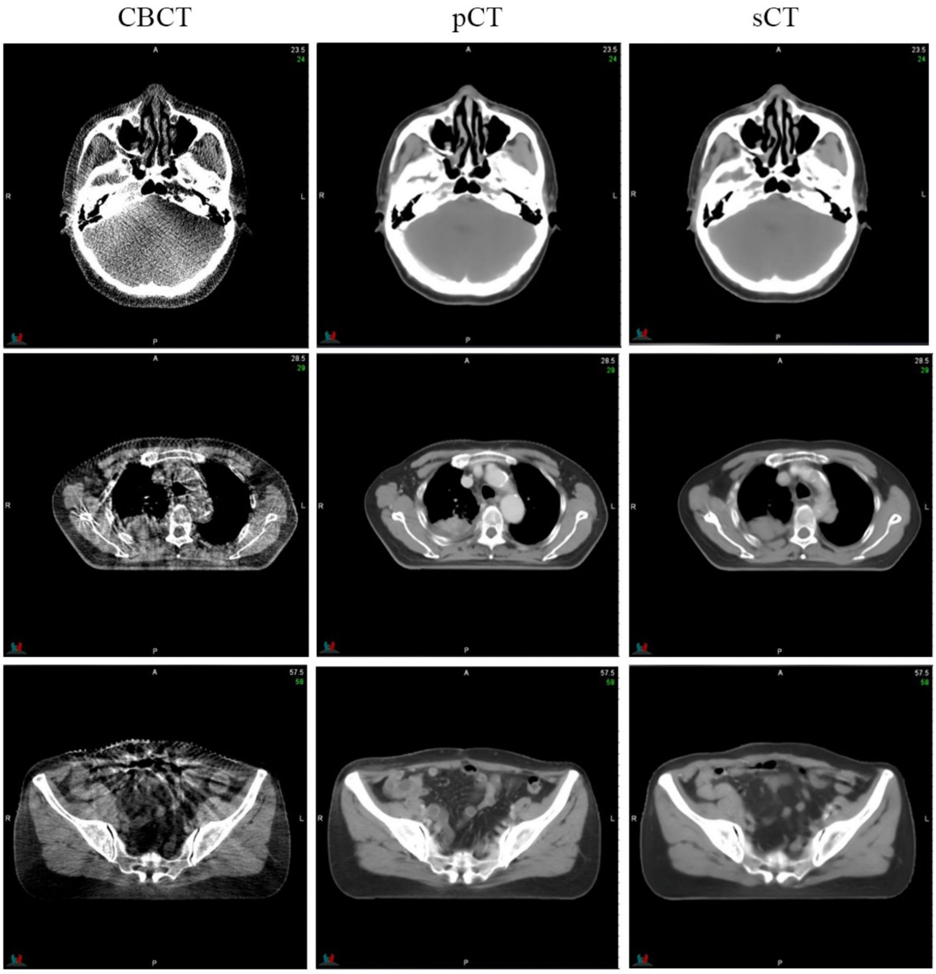


FIGURE 6
Examples of image slices obtained for the head, thorax and pelvis cases. The display window ranged from -400 HU to 400 HU.

obtained across bone, soft tissue and air are shown in Figure 7B. Furthermore, the HU distributions of the example case are shown in Figure 7C. Our model effectively mapped the intensity distribution of the CBCT image to the pCT image.

4.5 ROI contouring

The DSC and MDA values obtained for 15 patients, including 8 head patients and 7 pelvis patients, are presented in Tables 4, 5, respectively. Thorax cases were excluded because of the limited FOV of the CBCT scanning system, preventing full organ scanning.

Table 4 presents the DSC outcomes achieved with the UIH and TS tools, revealing a consistent enhancement in segmentation accuracy across most regions when sCT images were used instead of CBCT. Notably, the brainstem mean DSC significantly improved from 0.61 to 0.89. Additionally, Table 5 shows that the mean MDA values of the brainstem significantly decreased from 3.72 mm when CBCT was used to 0.98 mm when sCT was used. For the parotid glands, we also observed positive trends, highlighting the role of sCT in improving contour accuracy through image quality enhancement. Although the DSC and MDA values for the bladder and rectum do not significantly change, sCT images still provide slightly higher segmentation accuracy in these regions.

TABLE 3 Quantitative results of the ablation experiment.

	MAE (HU)	PSNR (dB)	SSIM (%)
M1 (Syn Only + L1)	18.54 ± 8.81	34.55 ± 3.06	91.80 ± 2.77
M2 (SynReg + L1)	17.48 ± 8.09	35.87 ± 2.93	93.96 ± 2.84
M3 (SynReg + Lperceptual)	17.89 ± 8.44	36.32 ± 3.27	94.38 ± 2.96
M4 (SynReg + Lperceptual + L1)	16.81 ± 8.42	36.59 ± 3.12	94.34 ± 2.85

The compared models (M1–M4) are trained with different settings.

5 Discussion

In this study, we employed a deep learning approach to translate multisite CBCT images to sCT images. To utilize paired data with local structural misalignment for training, we proposed SynREG, which disentangled the sCT generation and anatomical

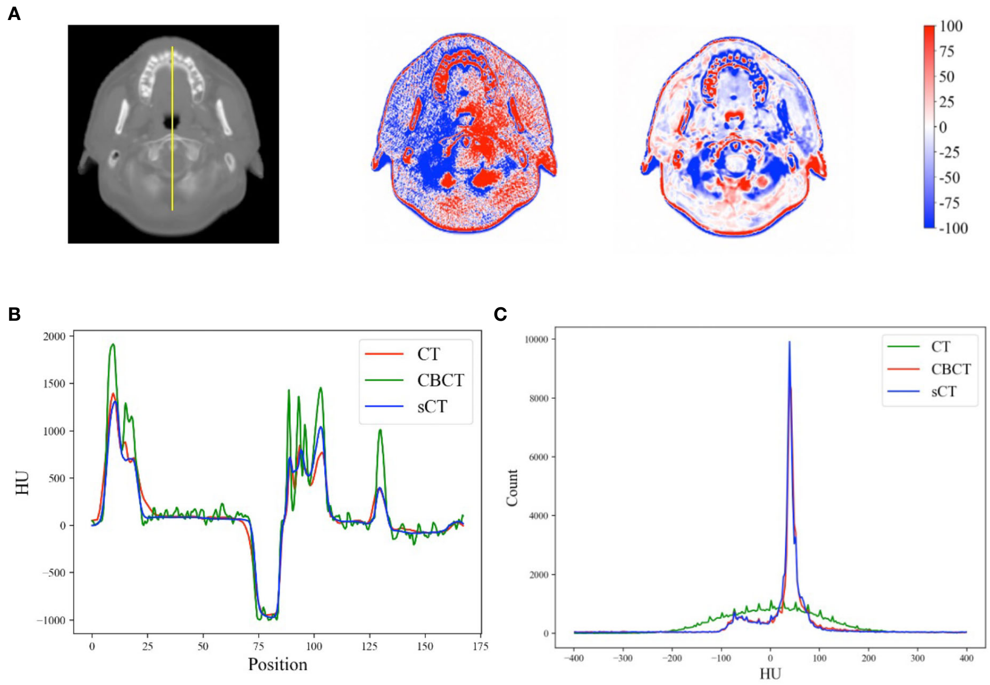


FIGURE 7
(A) HU differences between the CBCT, sCT and rsCT images and the pCT image for a head case example. **(B)** HU profiles of the yellow line in **(A)**. **(C)** HU distributions within the range of [-400, 400] for the example cases across all image modalities.

correction processes via a synthesizer and a Reg-Net, respectively. With this approach, we can train the model in a supervised manner, which has demonstrated the advantages of efficiency in data and computation due to its explicit learning objective (35). Moreover, the transfer learning ability of a supervised pretraining model can be further enhanced when models are trained on increasingly expansive datasets (36). In this study, we trained a singular model capable of generating sCT images for the head, thorax and pelvis. Figures 4, 5 demonstrate the efficacy of our proposed model in enhancing image quality across all sites through quantitative and qualitative evaluations. The quantitative results (Table 2) indicate that SynREG outperforms the other unsupervised benchmark models and outperforms the model trained solely on the head and neck dataset.

Although CBCT and pCT images share anatomical similarities, simple HU mapping is not sufficient for generating sCT images because noise and artifacts may easily introduce nonlinearity in the intensity profile mapping process. To capture more complete local and global relationships for high-quality image restoration and generation. We introduced a hybrid CNN-transformer model to enhance the representation ability of the model while saving computer resources. Traditional ViTs require many computational resources and large datasets. Various self-attention computation methods, such as local window attention (37, 38), channel dimension attention (39), and sparse self-attention (40), have been proposed to reduce model complexity. SwinIR (34) utilizes a swin transformer to perform image restoration. Compared with vanilla ViT, the swin transformer employs a shifted window mechanism to combine local attention with global attention, enabling the capture of global context information while maintaining computational efficiency. Despite its strengths, the patch embedding utilized in the Swin transformer encounters

TABLE 4 Comparison of the mean Dice coefficients for automatic segmentation of various organs between CBCT and sCT images.

		Brain Stem		Parotide_L		Parotide_R		Bladder		Rectum	
		CBCT	sCT	CBCT	sCT	CBCT	sCT	CBCT	sCT	CBCT	sCT
Dice	UIH	0.63	0.89	0.89	0.91	0.87	0.90	0.63	0.70	0.68	0.68
	TS	0.59	0.89	0.88	0.91	0.88	0.91	0.60	0.70	0.68	0.69
	Mean	0.61	0.89	0.88	0.91	0.88	0.91	0.62	0.70	0.68	0.68

TABLE 5 Comparison of the mean distance to agreement (MDA) for automatic segmentation of various organs between CBCT and sCT images. (mm).

		Brain Stem		Parotide_L		Parotide_R		Bladder		Rectum	
		CBCT	sCT	CBCT	sCT	CBCT	sCT	CBCT	sCT	CBCT	sCT
MDA	UIH	3.51	0.98	1.00	0.77	0.98	0.79	3.45	3.41	3.60	3.07
	TS	3.93	0.98	1.05	0.77	0.98	0.79	3.53	3.42	3.66	3.05
	Mean	3.72	0.98	1.03	0.77	0.98	0.79	3.49	3.42	3.63	3.06

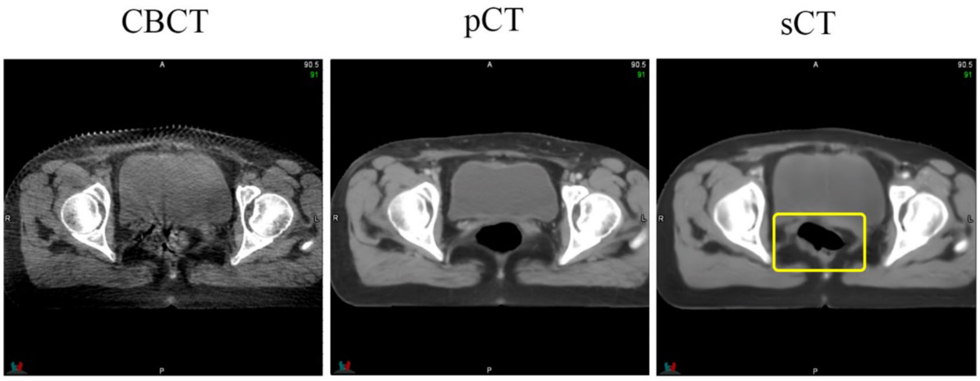


FIGURE 8
An instance of hallucinations on an sCT image. The area highlighted by the yellow box indicates the location of the hallucinations, which are not present in the CBCT image.

inherent limitations in capturing local details, particularly for the preservation of fine details (41), as shown in Figure 4. Hence, we did not opt for patch embedding and instead adopted a depthwise convolutional transposed attention mechanism, which proves to be more effective for enhancing the representation of fine details and for generating high-quality images. Additionally, this approach offers the advantage of linear computational complexity, making it a more efficient solution for our task.

The segmentation of target structures and organs at risk is a crucial component of the radiotherapy workflow. Most deep learning-based autosegmentation models applied in radiotherapy are trained with CT and/or MR images, significantly enhancing the efficiency and accuracy of the task (42). However, owing to the limited generalization ability of autosegmentation models, it is not advisable to train a model on one image modality and directly apply it to another modality, as this often leads to suboptimal performance.

Our results demonstrate that the DSC of CBCT is the lowest. This can be attributed to data distribution differences caused by different image modalities (Figure 7C), as well as the inferior quality of CBCT images. Following the conversion of CBCT to sCT images, both the data distribution disparity and image quality were enhanced, resulting in an increase in the DSC value. Notably, for the brainstem, which has low contrast and is overwhelmed by noise in CBCT images, the increase in DSC was primarily due to the essential image quality enhancement process. On the other hand, for the bladder and rectum, which have higher contrast and can be easily discriminated in CBCT images, the relatively lower DSC was primarily due to the structural mismatches caused by variations in bladder and rectum fullness. Therefore, the increase in DSC yielded by sCT was not significant for these regions. Similarly, the MDA results also supports these findings.

Supervised learning often relies on high-quality datasets for optimal model performance. Our method mitigates the need for precisely matched paired images by disentangling image generation from anatomical correction. This approach is capable of handling most scenarios and generating high-quality sCT images. However, we also encountered unreliable results in certain scenarios, especially for cavities. As shown in Figure 8, while the bladder structure of the CBCT image was effectively preserved, the appearance of a cavity on the sCT image was unreliable. This is because performing image registration in cases with large deformations remains a significant challenge, thereby compromising dataset quality. Exploring ways to enhance dataset quality and/or incorporating a more targeted loss function are potential approaches to address this limitation and achieve a more accurate clinical implementation model.

In this study, we used a joint learning framework called SynREG to address the challenge of training a model with imperfectly aligned CBCT–CT paired data. Our approach involved proposing a hybrid CNN-transformation model for sCT generation and a registration network for anatomical correction. Additionally, we explored the feasibility of training a singular model for generating multisite sCT images. Our quantitative and qualitative findings demonstrated the superior performance of our method and its potential application in ART.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

YH: Methodology, Writing – original draft. MC: Data curation, Formal analysis, Writing – original draft, Writing – review & editing. HW: Data curation, Writing – original draft. ZL: Conceptualization, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Noble DJ, Yeap P-L, Seah SYK, Harrison K, Shelley LEA, Romanchikova M, et al. Anatomical change during radiotherapy for head and neck cancer, and its effect on delivered dose to the spinal cord. *Radiother Oncol.* (2019) 130:32–8. doi: 10.1016/j.radonc.2018.07.009

2. Rührnschopf E-P, Klingenbeck K. A general framework and review of scatter correction methods in X-ray cone-beam computerized tomography. Part 1: Scatter compensation approaches: Scatter compensation approaches. *Med Phys.* (2011) 38:4296–311. doi: 10.1118/1.3599033

3. Stankovic U, Ploeger LS, van Herk M, Sonke J-J. Optimal combination of anti-scatter grids and software correction for CBCT imaging. *Med Phys.* (2017) 44:4437–51. doi: 10.1002/mp.12385

4. Trapp P, Maier J, Susenburger M, Sawall S, Kachelrieß M. Empirical scatter correction: CBCT scatter artifact reduction without prior information. *Med Phys.* (2022) 49:4566–84. doi: 10.1002/mp.15656

5. Inui S, Nishio T, Ueda Y, Ohira S, Ueda H, Washio H, et al. Machine log file-based dose verification using novel iterative CBCT reconstruction algorithm in commercial software during volumetric modulated arc therapy for prostate cancer patients. *Phys Med.* (2021) 92:24–31. doi: 10.1016/j.ejmp.2021.11.004

6. Maslowski A, Wang A, Sun M, Wareing T, Davis I, Star-Lack J. Acuros CTS: A fast, linear Boltzmann transport equation solver for computed tomography scatter - Part I: Core algorithms and validation. *Med Phys.* (2018) 45:1899–913. doi: 10.1002/mp.12850

7. Wang A, Maslowski A, Messmer P, Lehmann M, Strzelecki A, Yu E, et al. Acuros CTS: A fast, linear Boltzmann transport equation solver for computed tomography scatter - Part II: System modeling, scatter correction, and optimization. *Med Phys.* (2018) 45:1914–25. doi: 10.1002/mp.12849

8. Jarema T, Aland T. Using the iterative kV CBCT reconstruction on the Varian Halcyon linear accelerator for radiation therapy planning for pelvis patients. *Phys Med.* (2019) 68:112–6. doi: 10.1016/j.ejmp.2019.11.015

9. Hu Y, Arnesen M, Aland T. Characterization of an advanced cone beam CT (CBCT) reconstruction algorithm used for dose calculation on Varian Halcyon linear accelerators. *BioMed Phys Eng Express.* (2022) 8:025023. doi: 10.1088/2057-1976/ac536b

10. Jiang Y, Yang C, Yang P, Hu X, Luo C, Xue Y, et al. Scatter correction of cone-beam CT using a deep residual convolution neural network (DRCNN). *Phys Med Biol.* (2019) 64:145003. doi: 10.1088/1361-6560/ab23a6

11. Li Y, Zhu J, Liu Z, Teng J, Xie Q, Zhang L, et al. A preliminary study of using a deep convolution neural network to generate synthesized CT images based on CBCT for adaptive radiotherapy of nasopharyngeal carcinoma. *Phys Med Biol.* (2019) 64:145010. doi: 10.1088/1361-6560/ab2770

12. Liang X, Chen L, Nguyen D, Zhou Z, Gu X, Yang M, et al. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. *Phys Med Biol.* (2019) 64:125002. doi: 10.1088/1361-6560/ab22f9

13. Kurz C, Maspero M, Savenije MHF, Landry G, Kamp F, Pinto M, et al. CBCT correction using a cycle-consistent generative adversarial network and unpaired training to enable photon and proton dose calculation. *Phys Med Biol.* (2019) 64:225004. doi: 10.1088/1361-6560/ab4d8c

14. Kida S, Kaji S, Nawa K, Imae T, Nakamoto T, Ozaki S, et al. Cone-beam CT to Planning CT synthesis using generative adversarial networks. *ArXiv* (2019) abs/1901.05773. Available online at: https://api.semanticscholar.org/CorpusID:58014127.

15. Ali AM, Benjdira B, Koubaa A, El-Shafai W, Khan Z, Boulila W. Vision transformers in image restoration: A survey. *Sensors.* (2023) 23:2385. doi: 10.3390/s23052385

16. Parvaiz A, Khalid MA, Zafar R, Ameer H, Ali M, Fraz MM. Vision Transformers in medical computer vision—A contemplative retrospection. *Eng Appl Artif Intell.* (2023) 122:106126. doi: 10.1016/j.engappai.2023.106126

17. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021). Available online at: http://arxiv.org/abs/2010.11929 (Accessed November 17, 2023).

18. Chen X, Liu Y, Yang B, Zhu J, Yuan S, Xie X, et al. A more effective CT synthesizer using transformers for cone-beam CT-guided adaptive radiotherapy. *Front Oncol.* (2022) 12:988800. doi: 10.3389/fonc.2022.988800

19. Peng J, Qiu RLJ, Wynne JF, Chang C-W, Pan S, Wang T, et al. CBCT-Based synthetic CT image generation using conditional denoising diffusion probabilistic model. *Med Phys.* (2023) 51:1847–59. doi: 10.1002/mp.16704

20. Rossi M, Cerveri P. Comparison of supervised and unsupervised approaches for the generation of synthetic CT from cone-beam CT. *Diagnostics.* (2021) 11:1435. doi: 10.3390/diagnostics11081435

21. Chen L, Liang X, Shen C, Jiang S, Wang J. Synthetic CT generation from CBCT images via deep learning. *Med Phys.* (2020) 47:1115–25. doi: 10.1002/mp.13978

22. Zhang Y, Yue N, Su M, Liu B, Ding Y, Zhou Y, et al. Improving CBCT quality to CT level using deep learning with generative adversarial network. *Med Phys.* (2021) 48:2816–26. doi: 10.1002/mp.14624

23. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H. Restormer: Efficient Transformer for High-Resolution Image Restoration (2022). Available online at: http://arxiv.org/abs/2111.09881 (Accessed October 11, 2023).

24. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation (2015). Available online at: http://arxiv.org/abs/1505.04597 (Accessed May 25, 2024).

25. Chen X, Li H, Li M, Pan J. Learning A sparse transformer network for effective image deraining. *2023 IEEE/CVF Conference on Computer Vision and Pattern*

Recognition (CVPR). Vancouver, BC, Canada: IEEE (2023). pp. 5896–905. doi: 10.1109/CVPR52729.2023.00571

26. Kong L, Lian C, Huang D, Li Z, Hu Y, Zhou Q. Breaking the dilemma of medical image-to-image translation. *ArXiv E-Prints*. (2021) arXiv:2110.06465. doi: 10.48550/arXiv.2110.06465

27. Ding K, Ma K, Wang S, Simoncelli EP. Image quality assessment: unifying structure and texture similarity. *IEEE Trans Pattern Anal Mach Intell*. (2020) 44:2567–81. doi: 10.1109/TPAMI.2020.3045810

28. Lai W–S, Huang J-B, Ahuja N, Yang M-H. Fast and Accurate Image Super-Resolution with Deep Laplacian Pyramid Networks (2018). Available online at: http://arxiv.org/abs/1710.01992 (Accessed September 13, 2023).

29. Baig A, Chaudhry MA, Mahmood A. Local normalized cross correlation for geo-registration. *Proceedings of 2012 9th International Bhurban Conference on Applied Sciences & Technology (IBCAST)*. Islamabad, Pakistan: IEEE. (2012). pp. 70–4. doi: 10.1109/IBCAST.2012.6177529

30. Liu Z. Super convergence cosine annealing with warm-up learning rate. *2nd international conference on artificial intelligence, big data and algorithms*. (2022). p. 1–7.

31. Wasserthal J, Breit H-C, Meyer MT, Pradella M, Hinck D, Sauter AW, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol Artif Intell*. (2023) 5:e230024. doi: 10.1148/ryai.230024

32. Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-Image Translation with Conditional Adversarial Networks (2018). Available online at: http://arxiv.org/abs/1611.07004 (Accessed October 30, 2023).

33. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE international conference on computer vision (ICCV)*. (2017) 2242–51. doi: 10.1109/ICCV.2017.244

34. Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R. SwinIR: Image restoration using swin transformer. *2021 IEEE/CVF international conference on computer vision workshops (ICCVW)*. (2021). p. 1833–44. doi: 10.1109/ICCVW54120.2021.00210

35. Li W, Yuille A, Zhou Z. HOW WELL DO SUPERVISED 3D MODELS TRANSFER TO MEDICAL IMAGING TASKS? *The Twelfth International Conference on Learning Representations* (2024). Available online at: https://openreview.net/forum?id=AhizIPytk4.

36. Kolesnikov A, Beyer L, Zhai X, Puigcerver J, Yung J, Gelly S, et al. Big Transfer (BiT): General Visual Representation Learning (2020). Available online at: http://arxiv.org/abs/1912.11370 (Accessed May 26, 2024).

37. 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 (2022). Available online at: http://arxiv.org/abs/2201.01266 (Accessed October 30, 2023).

38. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. (2021). pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986

39. Maaz M, Shaker A, Cholakkal H, Khan S, Zamir SW, Anwer RM, et al. EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications (2022). Available online at: http://arxiv.org/abs/2206.10589 (Accessed November 8, 2023).

40. Zhao G, Lin J, Zhang Z, Ren X, Su Q, Sun X. Explicit Sparse Transformer: Concentrated Attention Through Explicit Selection (2019). Available online at: http://arxiv.org/abs/1912.11637 (Accessed November 8, 2023).

41. Li C, Zhang J, Wei Y, Ji Z, Bai J, Shan S. Patch Is Not All You Need (2023). Available online at: http://arxiv.org/abs/2308.10729 (Accessed September 13, 2023).

42. Harrison K, Pullen H, Welsh C, Oktay O, Alvarez-Valle J, Jena R. Machine learning for autoSegmentation in radiotherapy planning. *Clin Oncol*. (2022) 34:74–88. doi: 10.1016/j.clon.2021.12.003