Universiteit Utrecht

thirona

Cone-beam CT to CT harmonization by learning disentangled representations

Minor Research Project

MSc Medical Imaging

Raquel González López

Supervisors: Karan Malhotra and Bálint Hompot

Senior Deep Learning Engineers, Thirona B.v.

Examiner: Dr. ir. Koen L. Vincken

Associate Professor, UMC Utrecht

Nijmegen, 29th July 2024

Cone-beam CT to CT harmonization by learning disentangled representations

Raquel González López
MSc Medical Imaging
Universiteit Utrecht
r.gonzalezlopez@students.uu.nl

***Abstract*—Chest computed tomography (CT) is a widely used imaging modality for evaluating thoracic pathology, with fan-beam CT (FBCT) and cone-beam CT (CBCT) being the primary types. While CBCT reduces radiation exposure, it often results in lower image quality, limited field of view, and increased artifacts, restricting its clinical applications. Given these challenges and the predominance of FBCT data and automated models designed for FBCT, developing high-quality CBCT-to-CT image synthesis is essential for improving CBCT image quality and expanding its applications. One promising approach is image harmonization, which mitigates domain shifts in medical images by translating from a source image to a target image acquisition setting while preserving the underlying anatomy. Recent methods, such as HACA3 for magnetic resonance imaging, have used disentangled representations of anatomy, contrast, and artifacts to respect anatomical differences between contrasts. To this end, this study aimed to perform image synthesis harmonization between CBCT and FBCT to capture CBCT anatomy while preserving FBCT quality and resolution. The harmonized images were evaluated through downstream tasks requiring various detail levels: lobes, nodules, and airways segmentation. Results showed that the model effectively reduced artifacts and smoothed the HU distribution, leading to improved segmentation performance, particularly in lobes and airways, although some detail was lacking in segmented airways. External evaluations suggested the model's potential generalizability to other distributions. Overall, this harmonization approach enhances CBCT image quality, expanding its applicability across various imaging tasks.**

***Index Terms*—CT, Cone-beam CT, disentangled representation, image synthesis, harmonization**

I. Introduction

Chest computed tomography (CT) is a widely used imaging modality that allows for the evaluation of thoracic pathology [1, 2]. There are two main types of CT scanning methods: Fan-beam CT (FBCT) and Cone-beam CT (CBCT). In FBCT, X-rays are irradiated in a fan shape and imaged by a line sensor or X-ray flat panel. This method involves acquiring sequential, thin axial scans through the patient's volume, which are then processed using a computer. It is typically used in diagnosis or pre-operative planning since it produces high-quality images and soft tissue differentiation. On the contrary, CBCT is often conducted during interventional procedures, such as bronchoscopy, biopsy, or radiotherapy. CBCT uses diverging kV X-rays which are irradiated in a conical shape which enables the acquisition of images over a much larger volume in a single scan in comparison to FBCT. In this way, the scan can be done in a relatively shorter time [3].

The performance of CT systems is significantly influenced by the quality and specifications of the X-ray detectors and sources employed. Current CBCT acquisition modes aim to lower the radiation dosage to the patient. However, this reduction in dose may result in a reduction in the image quality and the accuracy of the assigned Hounsfield unit (HU) [4]. This HU helps distinguish different tissue types by value, which are independent of X-ray energy. Furthermore, the intrinsic issue of the large cone geometry produces noisier images and artifacts than FBCT. These limitations preclude CBCT images from various high-precision clinical applications.

Challenges in data variability between FBCT and CBCT further complicate their direct comparison and integration. It is common that both CTs are neither paired nor one-to-one mapped, especially because CBCTs have a more limited field of view (FOV). Structural differences might be present in the CBCT data, such as atelectasis or reduction in lung volume during intraoperative oxygenation (see Fig. 1) [5]. Other differences include variations in breathing volume, heart motion, and phantom airways—artifacts that mimic airways without existing physically. Additionally, the presence of intraoperative navigation devices such as bronchoscopes, which may contain metal components, can cause beam hardening artifacts [6]. Given these challenges and the predominance of FBCT data and automated models designed for FBCT, developing high-quality CBCT-to-CT image synthesis is essential for improving CBCT image quality and expanding its applications.

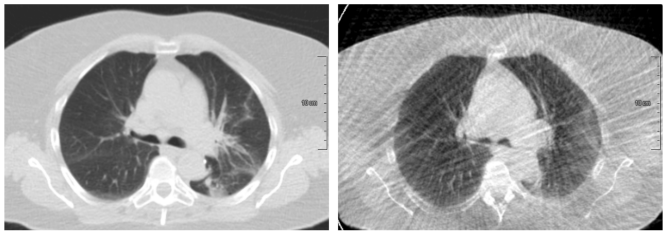

Fig. 1. An example of Fan-Beam (left) and Cone-Beam (right) chest CT.

One promising image synthesis approach is image harmonization, which aims to mitigate domain shift in medical images by translating from a source image to a target image acquisition setting while preserving the underlying anatomy. In recent years, image-synthesis harmonization techniques have emerged, which

are a special type of image-to-image (I2I) translation. This can be performed in supervised and unsupervised ways depending on the availability of training data [7]. On the one hand, supervised harmonization requires paired data, using pixel-to-pixel error for training, which lacks generalizability. On the other hand, unsupervised harmonization does not require paired training data but assumes a similarity between the two image domains. One of the most known I2I methods is CycleGAN by learning a translation model and a discriminator [8]. A common drawback in unsupervised models is the geometry shift, where structural changes in the image can lead to inconsistencies in anatomical features. This occurs because the harmonization process heavily depends on the discriminator's judgment.

Most of the unsupervised I2I methods rely on the assumption of identical anatomy between images, which is not the case for CT and CBCT images. This can be achieved through the cycle-consistency loss to encourage anatomy preservation. However, cycle-consistency loss alone is an insufficient constraint due to the geometry shift [9–11]. To overcome the challenges of unpaired data and differences in anatomy features between images to harmonize, recent advances in magnetic resonance imaging harmonization have focused on learning a disentangled representation of anatomy and contrast [11, 12]. As a result, this disentanglement of the latent space respects the inherent anatomical difference between contrasts. Zuo *et al.* [13] further explored this by proposing Harmonization with Attention-based Contrast, Anatomy, and Artifact Awareness (HACA3), which also disentangles artifact representations. Current state-of-the-art approaches for CBCT harmonization often use CycleGAN-based methods or a combination with disentangled representations [14–18]. These methods have shown promising results but are limited by the assumption of identical anatomy and potential structural differences in the data. Most of the disentangled representation approaches have been applied to magnetic resonance imaging and, to the best of our knowledge, have not been used for CBCT-to-CT image synthesis.

This study aims to perform I2I translation from CBCT to FBCT. The main objective is to capture the anatomy acquired in CBCT with the quality and resolution of FBCT by using the image synthesis harmonization approach. More specifically, we aim to adapt the HACA3 model [13] for CT data. To quantitatively assess the performance of the model, we evaluate the resulting harmonization on several downstream segmentation tasks including lobes, airways, and nodules structures. We further assess the model's adaptability by evaluating on external real CBCT data.

II. Methods

A. Dataset

The dataset used in this work is from the free publicly available Lung Image Database Consortium image collection (LIDC-IDRI), which consists of diagnostic and lung cancer screening thoracic CT scans [19]. It was initiated by the National Cancer Institute and further advanced by the Foundation for the National Institutes of Health. In total, it contains 1018 cases from seven different academic centers, with the collaboration of eight medical imaging companies. For this study, a subset of 95 cases was used, which were divided into 56 for training (70%), 9 for validation (10%), and 20 for testing (30%). All images have a resolution of 512 x 512 pixels in the x-y dimensions, with an in-plane pixel size ranging from 0.461 mm to 0.977 mm. The z-dimension and slice thickness vary between scans, with slice thicknesses ranging from 0.6 mm to 5.0 mm. Manual annotations from expert image analysts were used as ground truth during the downstream tasks evaluation.

To derive CBCT scans from the provided FBCT data, the ASTRA toolbox was employed, which is an open-source software package that supports tomographic reconstruction tasks [20, 21]. In this process, alignment of the simulated CBCT data with the FBCT was crucial. From this point forward, the simulated CBCT data will be referred to as genCBCT. To achieve this, a virtual Cone-beam projection is rotated 180 degrees around the FBCT volume to obtain the forward projection or sinogram. This is then reconstructed using the Feldkamp-Davis-Kress method [22]. To replicate a real CBCT scan setting, the same parameters as the Philips Allura FD20 Xper scanner were used: an 810 mm source-axis distance, an 1195mm source-detector distance, and a 30x40 cm^2 detector with a 0.775 mm pixel pitch [23]. Furthermore, a bronchoscope or scope was also simulated, as it is commonly present in interventional procedures for lung inspection. A bronchoscope typically consists of a thin, flexible outer tube with a working channel, where an inner tube can be inserted for procedures such as biopsy or ablation. This setup allows for detailed examination and intervention within the lung anatomy. Fig. 2 depicts the genCBCT with and without scope.

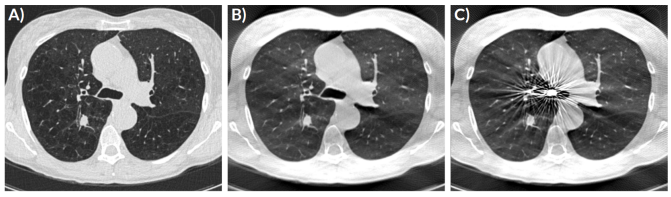

Fig. 2. An example of A) Fan-Beam CT with its correspondent ASTRA generated CBCT B) without the presence of scope and C) with scope.

To perform an external evaluation with real CBCT data, the Learn2Reg dataset was utilized [24, 25]. This dataset comprises 14 patients, each with a set of scans: prior FBCT and CBCT taken before treatment, and another FBCT and CBCT taken at the end of treatment. The FBCT scans were acquired at maximum inspiration, while both CBCT scans were taken at maximum expiration. All images had the same dimensions of 390x280x300 with a spacing of 1x1x1 mm. Due to the limited FOV, the CBCT images were zero-padded to match these dimensions. This dataset has been previously used in image registration challenges due to its inclusion of variations in breathing phases and time shifts between the beginning and end of treatment. It was employed during inference to evaluate the model's adaptability to real CBCT data.

B. Preprocessing

Both the FBCT and the genCBCT images were resized to 224x224x224, to obtain the same number of 2D slices in all orientations (axial, coronal, and sagittal) from the 3D volumes. In the case of the Learn2Reg data, the zero-padding present in the CBCTs was removed before the resizing operation. Then, the intensity values of the images were clipped between -1000 and 2500. The high maximum clipping parameters were chosen to account for the higher values of the bronchoscope and its artifacts since the metallic components usually have high HU units ($\geq$ 2000).

C. Network architecture

Following the image preprocessing, the next section delineates the employed network architecture. The adopted framework is based on the HACA3 architecture [13], which follows an "encoder-attention-decoder" structure. It consists of three major components: (1) encoding contrast, anatomy, and artifacts, (2) anatomy fusion with attention, and (3) decoding (see Fig. 3).

1) Learning anatomy, contrast, and artifact representations: To disentangle the different anatomy, contrast, and artifact representation, in HACA3 they introduced contrastive learning [26]. The idea behind it is learning discriminative features from query, positive and negative examples. In the case of the anatomy encoder, these examples are small image patches and can be denoted as p_q, p_+, and p_-. First, the query patch p_q is randomly selected at a random location of the anatomy representation β of one of the two contrasts, either 0 which represents FBCT, or 1 which represents CBCT. Then, the positive patch p_+ is selected from the same corresponding location in the other contrast. Meanwhile, the negative patches $p_-^{(n)}$ are sampled from the same locations as p_q as well as from random locations within different anatomy representations β. From this, with a contrastive loss function $\mathcal{L}_C$, p_q is encouraged to be more similar to p_+ than to $p_-^{(n)}$, but not identical. It can be defined as follows:

$$\mathcal{L}_C(p_q, p_+, p_-^{(n)}) = -\log\left[\frac{e^{p_q \cdot p_+}}{e^{p_q \cdot p_+} + \frac{1}{N}\sum_{n=1}^{N} e^{p_q \cdot p_-^{(n)}}}\right], \quad (1)$$

where $n = 1, ..., N$ and N is the total number of negative examples. In this way, the $p_-^{(n)}$ at the same location in different contrast encourages contrast information to be removed from β, and the random locations help to capture a broader anatomy representation.

The artifact encoder also uses contrastive learning with query, positive, and negative examples denoted as x_q, x_+, and $x_-^{(m)}$ respectively. In this case, the x_q and x_+ examples represent 2D slices from the same volumetric image. The negative examples $x_-^{(m)}$ are either slices from different volumes or augmented examples of x_q with simulated artifacts. These artifacts involved noise, motion, spikes, and ghosting, and were simulated using TorchIO [27]. This helps the artifact encoder to focus only on the artifact information since it is the only difference between those examples. Moreover, selecting slices from different volumes encourages different levels of artifacts to be learned. The final loss to train the artifact encoder is given by $\mathcal{L}_C(x_q, x_+, x_-^{(m)})$ in Eq. 1, where $m = 1, ..., M$ and N is the total number of negative examples.

2) Decoding with attention mechanism: To account for inherent anatomical differences across CT scans, the model learns to integrate anatomical features to generate harmonized images with high fidelity and improved robustness against imaging artifacts and poor quality. As illustrated in Fig. 3, fully connected networks are employed to learn keys $K = k_1, k_2$ and queries Q from the encoded representations of contrast θ and artifacts η [28]. The similarity between K and Q is measured to obtain the learned attentions α. These attentions highlight source images with contrast and image quality similar to the target image, guiding the decoder to use the corresponding β representations for harmonization. Then, these are used to obtain the optimal anatomical representation β^* by computing the weighted average with the attention α. Lastly, the decoder uses both β^* and the target contrast θ_t as inputs to generate the output synthetic image $\hat{x}_t$.

D. Loss function

The HACA3 framework is structured as a conditional variational autoencoder (CVAE), where θ serves as the latent variable and β^* acts as the condition. The CVAE loss used to train HACA3 is given by:

$$\mathcal{L}_{\text{CVAE}} = |\hat{x}_t - y_t|_1 + \lambda_1 \mathcal{D}_{\text{KL}}[p(\theta|y_t)\|p(\theta)], \quad (2)$$

where $|\hat{x}_t - y_t|_1$ is the L1 norm between the synthetic image $\hat{x}_t$ and target image y_t, $\mathcal{D}_{\text{KL}}$ is the KL divergence and $p(\theta)$ is the standard normal distribution. Furthermore, a cycle consistency loss between the contrast E_θ and artifact encoder E_{eta} is used to reanalyze the synthetic image and further regularize. It can be defined as:

$$\mathcal{L}_{cyc} = |E_\theta(\hat{x}_t) - \theta_t|_1 + |E_\eta(\hat{x}_t) - \eta_t|_1, \quad (3)$$

where θ_t and η_t are the target contrast and artifact representations, respectively. These two losses, $\mathcal{L}_{CVAE}$ and $\mathcal{L}_{cyc}$ are combined with the contrastive loss described in Eq. 1 for both the anatomy and artifact encoder as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CVAE}} + \lambda_2 \mathcal{L}_C(p_q, p_+, p_-^{(n)}) + \lambda_3 \mathcal{L}_C(x_q, x_+, x_-^{(m)}) + \lambda_4 \mathcal{L}_{\text{cyc}}, \quad (4)$$

where λ's are hyperparameters: $\lambda_1 = 10^{-5}$ and $\lambda_{2,3,4} = 0.1$, based on Zuo *et al.* [13].

E. Implementation details

To better align the genCBCT with actual CBCT, data augmentation techniques were applied. To mimic the restricted FOV, random cropping was used, such as isolating just one lung in the image. Additionally, jitter transformations, including adjustments to brightness, contrast, and sharpness, were applied

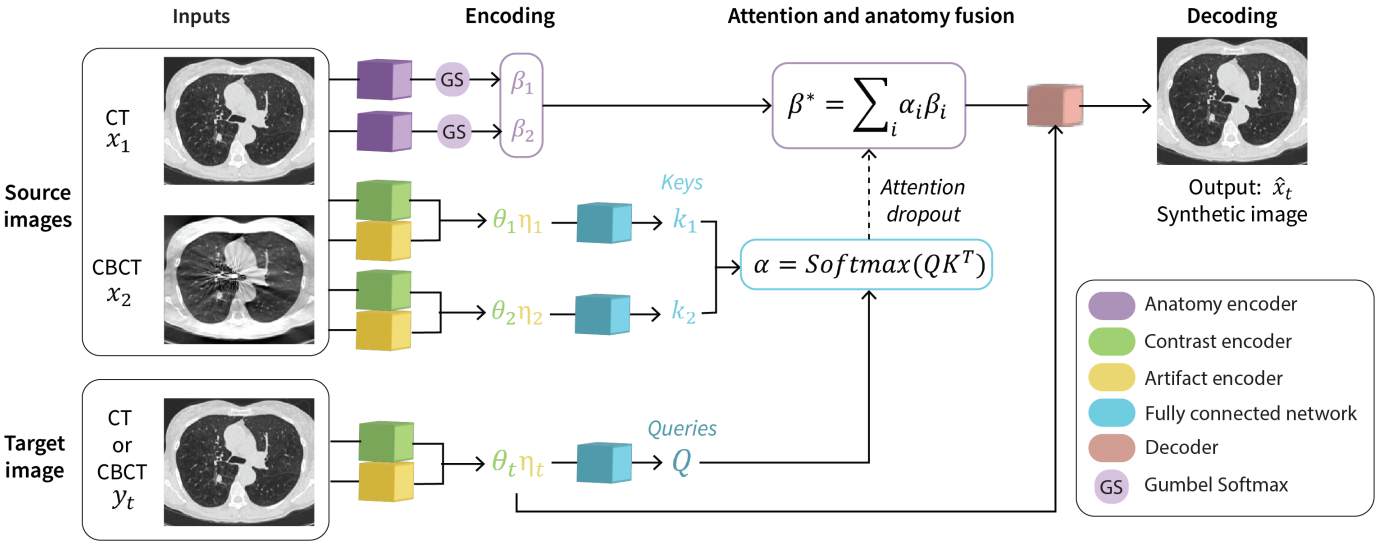

Fig. 3. A schematic framework of HACA3 for Fan-beam CT and Cone-beam CT data. The source images $x_{1,2}$ are both the FBCT and generated CBCT (genCBCT). $\beta_{1,2}$ represents the anatomical representation of the source images. θ and η represent the contrast and artifact representation respectively. α represents the attentions and β^* the optimal anatomy. The output synthetic image $\hat{x}_t$ has the contrast of the target image y_t and preserves the anatomy from the source images. The network colors indicate weight sharing.

to modify the image's intensity values, addressing potential common artifacts [29].

The model was trained for 70 epochs with a batch size of 12. Inference was then carried out on the test set using the weights from the final epoch. During inference, either a target image or the contrast and artifacts distribution has to be provided to the model. For this purpose, a randomly selected FBCT image from the test set was used as the target to obtain the desired distributions. Although the trained model can harmonize FBCT images using CBCT as a target, this study only uses FBCT as the target to achieve our goal. Employing a test set image ensures an unbiased evaluation of the model's generalization capabilities on unseen data, thereby providing a robust assessment of its performance. Given that this is a 2D approach, synthetic images were generated for all three orientations. The median value at each voxel was used as the intensity value for the final fused image [7]. Since the output synthetic images are normalized, the normalization is reversed to obtain the intensity values in the HU scale.

The implementation was based on PyTorch (v1.12.0) [30] and both training and inference were performed on a Nvidia GeForce RTX 2080 (12GB) using CUDA version 11.3.

F. Evaluation

To compare the resulting harmonized image with its respective target CT, the evaluation included the use of the structural similarity index measurement (SSIM) [31] and the peak signal-to-noise ratio (PSNR) value [32]. The SSIM is a perceptual metric for the assessment of the visual quality of images. It can be defined for two grayscale images x and y as follows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (5)$$

where μ_x, μ_y, σ_x^2, σ_y^2, σ_{xy} are the mean values, variance and covariance of x and y respectively, while c_1 and c_2 stabilize the equation. PSNR quantifies how much noise or distortion is present and can be defined between image x and reference image y as:

$$PSNR(x,y) = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right), \quad (6)$$

where the numerator is the maximum possible intensity value of the image and MSE denotes the mean squared error between the images. If we consider a 3D image with dimensions $MxNxK$, then MSE can be defined as:

$$MSE(x,y) = \frac{1}{MNK}\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}[x(m,n,k) - y(m,n,k)]^2, \quad (7)$$

The model was also evaluated on three downstream tasks: (i) lobes segmentation, (ii) nodules segmentation, and (iii) airways segmentation. These tasks provide a comprehensive assessment of the model's performance, ranging from simpler analyses, such as lobes segmentation or nodules, to more detailed, airways. All the different evaluations were performed the FBCT, the genCBCT, and the resulting harmonized CBCT. To ensure a fair comparison, all images were downsampled to a uniform size of 224x224x224, as the original images varied significantly in resolution.

1) Lobes segmentation: The lung quantification platform LungQ (version 3.0.0, Thirona, Nijmegen, The Netherlands) was used to perform the lobes segmentation. It predicts five classes corresponding to the different lobes: the left superior lobe, the left inferior lobe, the right superior lobe, the right middle lobe, and the right inferior lobe. To assess the performance of the lobes segmentation, the Dice Score

Coefficient (DSC) and the Hausdorff distance (HD) were used. The DSC quantifies the overlap between two volume segmentations and is defined as twice the intersection of the volumes divided by their union [33]. The HD calculates the distance between two point sets and it is an indicator of the largest segmentation error[34]. These metrics were computed per lobe label between the predicted LungQ and ground truth segmentations. To evaluate the differences between the FBCT, genCBCT, and the harmonized genCBCT metrics, a statistical analysis was performed. More specifically the Wilcoxon-ranked t-test [35] for non-parametric paired data with Bonferroni adjustment for multiple test comparison was used.

2) Nodules segmentation: For the nodules segmentation task, only the nodules $\geq 3mm$ in diameter in all orientations were considered since only those were present in the ground truths. To examine the performance, both precision and recall metrics were calculated between the nodules in the ground truth and their corresponding predictions. Recall is used in this context because it reflects the model's ability to identify all relevant nodules. It provides a measure of the segmentation algorithm's sensitivity, which is crucial for clinical applications where missing true positive nodules can have significant consequences. Precision complements recall by indicating the proportion of predicted nodules that are true positives, ensuring the model's predictions are accurate. It should be noted that the nodules segmentation module was previously trained on simulated CBCT data.

3) Airways segmentation: The airways segmentation task was also performed with LungQ. From the segmentation, the total airway count (TAC) can be obtained. It is directly computed by the algorithm, which selects the largest connected component and counts each airway branch and its respective children. Furthermore, the airway count per generation level was calculated to obtain a more in-depth comparison between segmentations.

III. Results

In this section, we present the results of the model performance and downstream tasks, focusing exclusively on test outcomes. For reference, HACA3 achieved a PSNR of 35.00 and an SSIM of 0.93 on magnetic resonance imaging [13], while a CycleGAN combined with a disentangled representation obtained a PSNR of 34.12 and an SSIM of 0.86 on CBCT without scope [17]. In our study, using the LIDC test set, the PSNR was 32.53 [31.09, 33.63] and the SSIM was 0.88 [0.83, 0.91] when comparing the harmonized genCBCT with the target FBCT.

An example of median harmonization performance is illustrated in Fig. 4. The results show that the model effectively reduces high-intensity values produced by the scope and minimizes artifacts, such as motion artifacts. However, some beam hardening artifacts persist, as seen in Fig. 4 A and B. Additionally, in certain areas, the airway wall is not fully reconstructed, as indicated in Fig. 4 C.

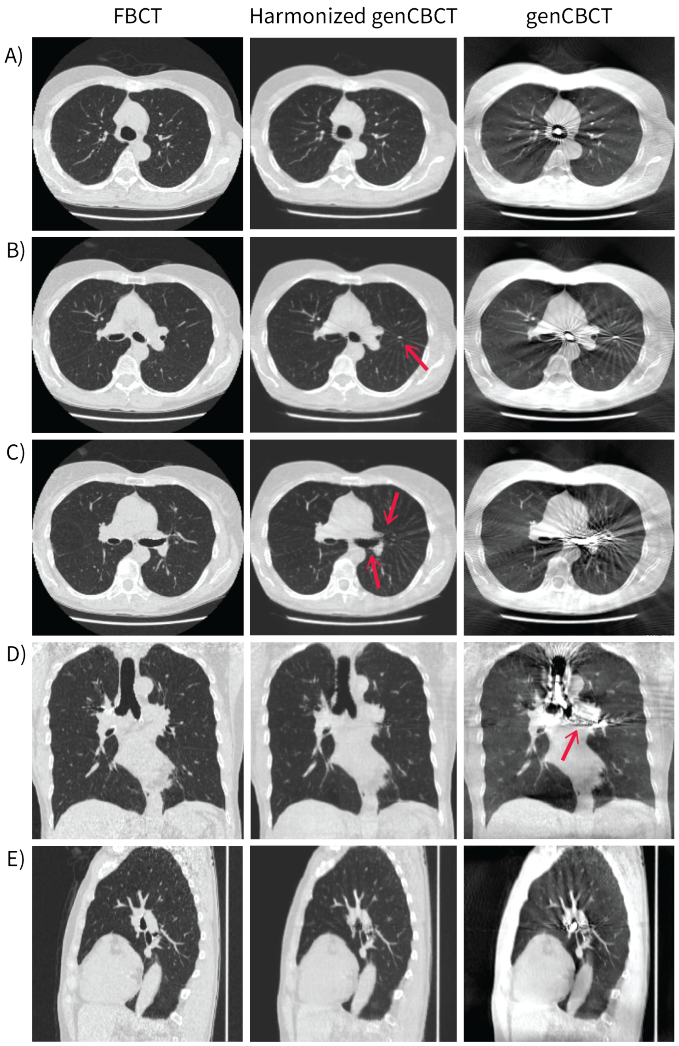

Fig. 4. Visualization of a median case based on PSNR and SSIM metrics. Slices of all three orientations are shown for the target FBCT, harmonized generated CBCT (genCBCT), and source genCBCT images. The PSNR between the target FBCT and the harmonized genCBCT is 32.71 and the SSIM is 0.86. The Red arrows are used to highlight specific regions to compare.

For visualization, Fig. 5 shows a "good" case with a PSNR of 33.99 and an SSIM of 0.91, which falls in the 75th percentile of these metrics results, indicating high-quality performance. Next to it, is the worst case with a PSNR of 29.61 and an SSIM of 0.76. The different segmentation results are described in the following subsections.

A. Lobes segmentation

The harmonized genCBCT outperformed the genCBCT images for all five lobe segmentation labels in terms of DSC, as can be seen in Fig. 6. Statistically significant differences were observed in DSC values between FBCT and both harmonized genCBCT and genCBCT for the left superior lobe ($p < 0.001$), but no HD comparisons reached significance. In label 2, all DSC comparisons were significant, and only the HD between FBCT and genCBCT was significant. Label 3 showed significant differences in DSC and HD between FBCT and genCBCT,

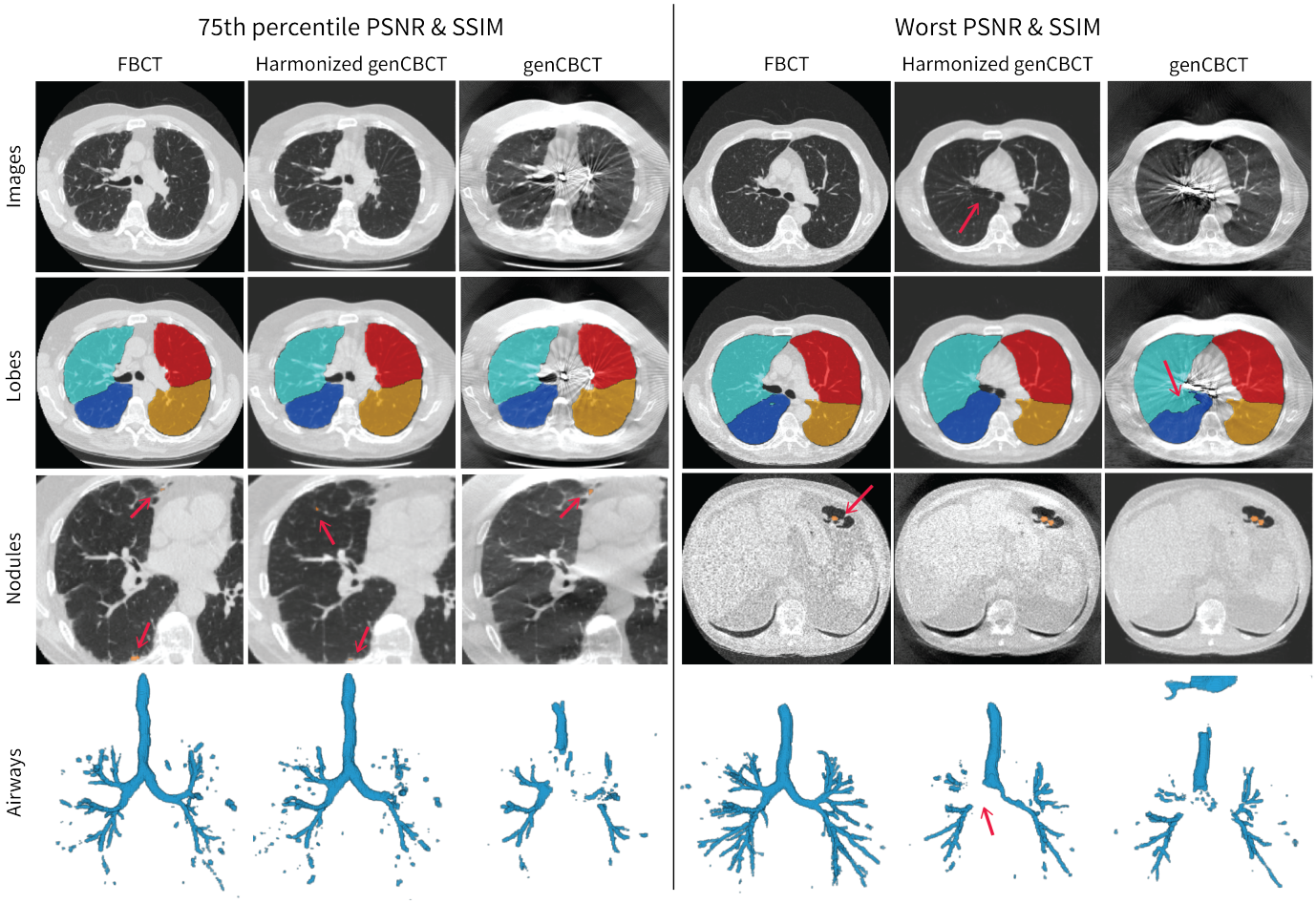

Fig. 5. Visualization of a good case (75th percentile) and the worst case based on PSNR and SSIM metrics. The first row displays the target FBCT, harmonized generated CBCT (genCBCT), and source genCBCT images. The second row shows lobe segmentations overlaid on the images, with colors representing different lobes: red for the left superior lobe, orange for the left inferior lobe, cyan for the right superior lobe, purple for the right middle lobe, and dark blue for the right inferior lobe. The third row highlights nodule segmentations in orange, overlaid on a relevant image slice. The final row presents airway segmentations with all components. Red arrows are used to highlight specific regions to compare.

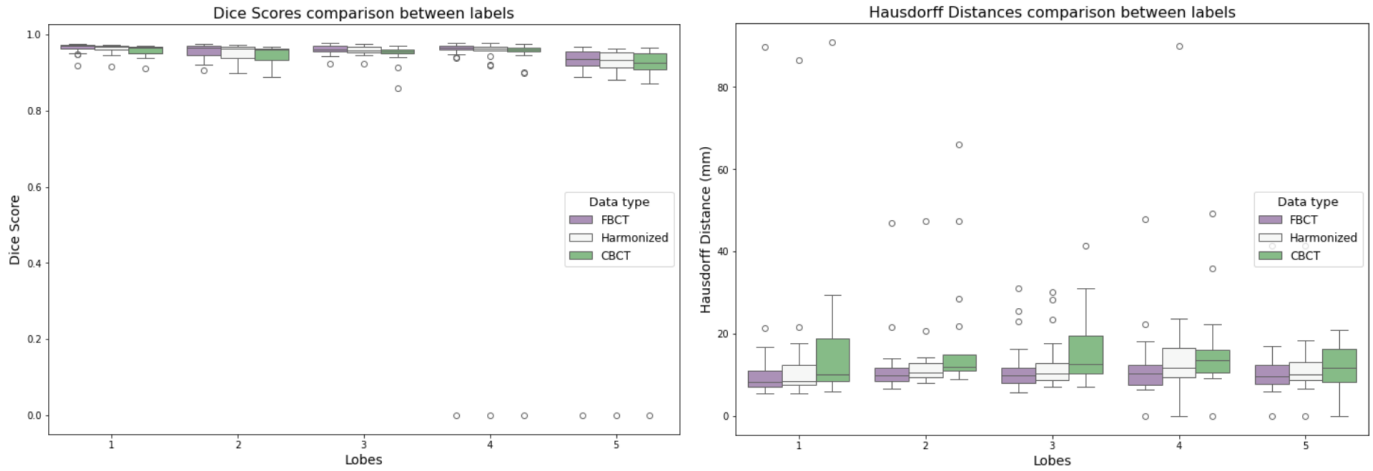

Fig. 6. Evaluation metrics results of the lobes segmentation downstream task. On the left, boxplot of the DSC values for the source FBCT (purple), harmonized genCBCT (white), and genCBCT (green). On the right, the HD values for the same sets. The lobes are as follows: (1) the left superior lobe, (2) the left inferior lobe, (3) the right superior lobe, (4) the right middle lobe, and (5) the right inferior lobe.

as well as DSC difference between harmonized genCBCT and CBCT. For Label 4, only the HD between FBCT and genCBCT was significant. Lastly, in Label 5, both DSC and HD comparisons between FBCT and genCBCT, and FBCT and harmonized genCBT were significant. Visually, the lobe segmentations appear similar across the different image types, although differences are noticeable in genCBCT in regions affected by the scope (see Fig. 5).

B. Nodules segmentation

The performance of nodule segmentation across different imaging modalities was assessed using recall and precision metrics, as detailed in Table I. For recall, both FBCT and genCBCT exhibited a median value of 0.50 with a range of [0.50, 1.00], indicating consistent detection capability across these modalities. The harmonized genCBCT had a lower median recall of 0.50 with a wider range of [0.35, 1.00], suggesting some variability in nodule detection post-harmonization. It is important to note that the model was trained on FBCT and genCBCT data, but not on harmonized genCBCT, which may affect its performance on the latter. In terms of precision, again FBCT and genCBCT achieved the same median value of 0.33, with a range of [0.25, 0.50]. The harmonized genCBCT images showed lower precision, with a median of 0.25 [0.20, 0.33]. These results indicate a decrease in recall and precision with harmonization. For example, in Fig. 5 it can be seen how in the "good" case one nodule was not segmented by either the harmonized genCBCT or CBCT. while in the worst case, where only one nodule was present, two nodules were segmented.

TABLE I
EVALUATION METRICS RESULTS FOR NODULES SEGMENTATION TASK

	FBCT	**Harmonized genCBCT**	**genCBCT**
Precision	0.33 [0.25, 0.50]	0.25 [0.20, 0.33]	0.33 [0.25, 0.50]
Recall	0.75 [0.50, 1.00]	0.50 [0.35, 1.00]	0.75 [0.50, 1.00]

C. Airways segmentation

The median of the TAC for FBCT was 85.0, for harmonized genCBCT 4.0, and for the genCBCT zero. The low or zero value for genCBCT is due to the absence of detected TAC values in 9 cases for the harmonized genCBCT and 14 cases for the genCBCT test cases. The results of airway count per generation level are shown in Fig. 7. Lower generation levels, which represent the main airway branches, are almost consistently detected across all image types. As the generation level increases, the airway count also increases. However, for both the harmonized genCBCT and genCBCT, this remains significantly lower. The airway count for harmonized genCBCT is lower than FBCT but higher than genCBCT. At the highest generation levels, harmonized genCBCT often fails to detect airways. This can be visually seen in Fig. 5, where both good and worst-case scenarios show the CBCT as disconnected, with the trachea not fully segmented. In addition, the depth of the harmonized genCBCT airway tree does not match that of the FBCT, and even in the worst case, it is disconnected in multiple regions.

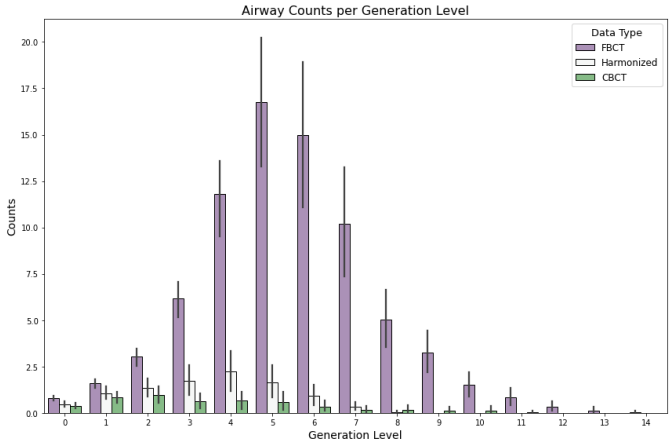

Fig. 7. Airways counts per generation level for the source FBCT (purple), harmonized genCBCT (white), and genCBCT (green).

D. External evaluation: Learn2Reg

Figure 8 shows two examples of harmonization performance using the Learn2Reg dataset, which consists of real CBCT data with inherent complexities. In Figure 8 A, there is a noticeable reduction in artifacts and smoothing of HU values in the harmonized CBCT. However, due to the presence of high levels of artifacts in the real CBCT data, such as streak artifacts, these artifacts remain visible even after harmonization. The examples depicted are from pre-treatment scans, illustrating the model's performance with real-world data that is more challenging than the simulated CBCT data used during training and evaluation.

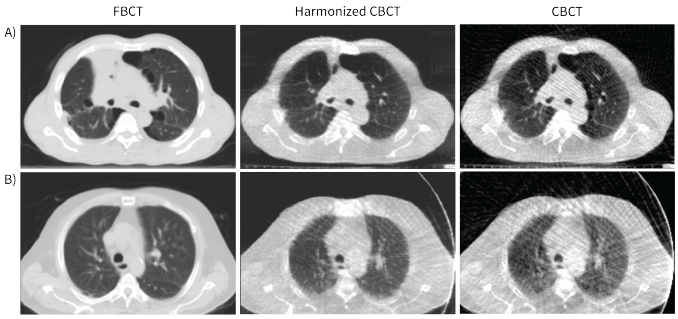

Fig. 8. Visualization of the Learn2Reg external data results are shown.

IV. DISCUSSION

In this study, we evaluate the performance of an image synthesis harmonization approach based on the disentanglement of the latent representation of anatomy, contrast, and artifacts. Our results highlight the model's ability to reduce artifacts, such as motion and scope, and to smooth the HU distribution. These enhancements translate to improved performance in

downstream tasks like lobes and airway segmentation, where the harmonized genCBCT consistently outperforms genCBCT.

Our model not only enhances CBCT images with scope artifacts but also performs exceptionally well on images without scope. Although these results are not included in the main body, the model's ability to achieve superior SSIM values on images with scope, in comparison to current approaches in images with scope [17], highlights its robustness and potential for broader clinical applications. This is significant, as harmonizing images without scope is inherently easier due to the reduced number of artifacts, thus further validating the model's capabilities.

In some cases with high artifacts, the airway wall was not fully reconstructed. This was especially true when the airway wall was thin or small (see Fig. 4 C). This was also affected by the slicing artifacts caused, by the median fusion approach to obtain the final harmonized genCBCT image. which could be solved by a fusion network [11] or 3D approach. The scope HU is reduced but not entirely. The beam hardening artifacts produce by the scope are still present in some cases.

Despite these successes, the model faced challenges in accurately reconstructing airway walls in high-artifact cases, particularly when the walls were thin or small (see Fig. 4 B and C). This limitation was partly due to slicing artifacts from the median fusion approach used to generate the final harmonized genCBCT image. Exploring alternative methods, such as a fusion network [11] or 3D approaches, could address these issues. While the model reduced the scope HU, beam-hardening artifacts persisted in some instances, indicating areas for future improvement.

In the lobes segmentation task, harmonized genCBCT consistently outperformed genCBCT, with statistically significant improvements in DSC and HD. On the contrary, nodule segmentation performance was slightly diminished. The model's tendency to overshoot on FBCT to avoid missing potential nodules resulted in lower precision and recall for harmonized genCBCT. This could also be due to the model being trained on both FBCT and CBCT. This underscores the effectiveness of the harmonization process in maintaining nodule detection capabilities but also highlights precision challenges that could impact the reliability of segmentation results. Further research is needed to address these challenges and refine the harmonization process to enhance overall segmentation performance.

For the airway segmentation, the harmonization slightly improved the performance both quantitatively with the TAC and airway count per generation level and quantitatively by correctly segmenting the trachea and in some cases connecting the left and right bronchus (see Fig. 5). However, there was a notable difference with respect to the FBCT. This indicates that the harmonized genCBCT have a lower level of detail, similar to CBCT. In addition, there can also be disconnected airways, such as the worst case in Fig. 5.

Regarding the external evaluation of Learn2Reg, the performance was variable depending on the quality and resolution of the images. The FBCT and CBCT images in this dataset were not previously aligned, and there was a notable difference in resolution between them. Additionally, the artifact distribution in Learn2Reg was not represented in our training data. Even with these challenges, it could be seen how the model was able to slightly reduce the present artifacts while preserving the anatomy. This suggests that incorporating more diverse training datasets that include a wider range of artifact types and resolutions to enhance the model's robustness to handle varied clinical scenarios.

One of the main limitations of this study is the use of a 2D approach. A 3D patch-based approach could potentially improve model performance by utilizing the full spatial context of the images rather than processing them slice-by-slice. However, this approach requires higher GPU capacity and computational resources. Moreover, we resized the images to a lower resolution to manage computational constraints, which may have affected the level of detail present. The 3D approach could mitigate this issue by allowing the use of original-size images instead of a fixed cubic dimension, potentially enhancing the model's ability to capture fine anatomical details.

In addition to the previously discussed future work, several other improvements could be explored. Firstly, enhancing data augmentation strategies by incorporating more jitter and simulating realistic CT-specific artifacts using tools such as ASTRA [20, 21] could help to better represent beam hardening effects. This includes modeling various artifacts such as strikes and dark bands. Secondly, integrating scopes with different materials and characteristics at various positions within the CBCT scan could improve the model's robustness. Furthermore, incorporating the concept of spatial attention could further enhance model accuracy and performance. By allowing attention variables to vary across spatial locations, the model could better adapt to local anatomical features, especially in complex regions, and provide more fine-grained control over the harmonization process. These advancements could address current limitations and are essential for enhancing the reliability and precision of CBCT imaging in clinical settings. For example, improved CBCT could significantly enhance tumor localization and dose planning in radiation therapy, leading to better patient outcomes and more effective treatments.

V. Conclusion

In this study, we explored the use of I2I translation to enhance simulated genCBCT images, aiming to achieve the anatomical detail of CBCT with the quality and resolution of FBCT. Our approach employed HACA3, which utilizes disentangled representations of anatomy, contrast, and artifacts for harmonization. The architecture includes an anatomy fusion module with an attention mechanism to address inherent anatomical differences between CT scans. Our results demonstrated that the model effectively reduced artifacts like motion and scope and smoothed the HU distribution. The model was evaluated on several downstream tasks, from general ones like lobe segmentation to more specific ones such as nodules

and airway segmentation. In lobe segmentation, harmonized genCBCT images consistently outperformed genCBCT. For nodule segmentation, harmonized genCBCT showed slightly lower precision and recall, as both FBCT and CBCT were used for training. For airway segmentation, a slight improvement was observed with harmonized CBCT over CBCT, though the images lacked detail. External evaluations indicated the model's potential generalizability to other distributions. Overall, the harmonization approach effectively increased the quality and diversity of the genCBCT dataset, demonstrating its potential to enhance segmentation tasks and its applicability to various imaging distributions.

REFERENCES

[1] T. M. Buzug, "Computed tomography," in *Springer handbook of medical technology*, Springer, 2011, pp. 311–342.

[2] P Whiting, N Singatullina, and J. Rosser, "Computed tomography of the chest: I. basic principles," *Bja Education*, vol. 15, no. 6, pp. 299–304, 2015.

[3] L. Lechuga and G. A. Weidlich, "Cone beam ct vs. fan beam ct: A comparison of image quality and dose delivered between two differing ct imaging modalities," *Cureus*, vol. 8, no. 9, 2016.

[4] U. V. Elstrøm, L. P. Muren, J. B. Petersen, and C. Grau, "Evaluation of image quality for different kv cone-beam ct acquisition and reconstruction methods in the head and neck region," *Acta Oncologica*, vol. 50, no. 6, pp. 908–917, 2011.

[5] P. Pelosi, L. Ball, M. G. de Abreu, and P. R. Rocco, "General anesthesia closes the lungs: Keep them resting," *Turkish journal of anaesthesiology and reanimation*, vol. 44, no. 4, p. 163, 2016.

[6] R. A. Brooks and G. Di Chiro, "Beam hardening in x-ray reconstructive tomography," *Physics in medicine & biology*, vol. 21, no. 3, p. 390, 1976.

[7] B. E. Dewey, C. Zhao, A. Carass, J. Oh, P. A. Calabresi, P. C. van Zijl, and J. L. Prince, "Deep harmonization of inconsistent mr data for consistent volume segmentation," in *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, Springer, 2018, pp. 20–30.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[9] R. K. Gebre, M. L. Senjem, S. Raghavan, C. G. Schwarz, J. L. Gunter, E. I. Hofrenning, R. I. Reid, K. Kantarci, J. Graff-Radford, D. S. Knopman, *et al.*, "Cross–scanner harmonization methods for structural mri may need further work: A comparison study," *Neuroimage*, vol. 269, p. 119 912, 2023.

[10] H. Yang, J. Sun, A. Carass, C. Zhao, J. Lee, Z. Xu, and J. Prince, "Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, Springer, 2018, pp. 174–182.

[11] L. Zuo, B. E. Dewey, Y. Liu, Y. He, S. D. Newsome, E. M. Mowry, S. M. Resnick, J. L. Prince, and A. Carass, "Unsupervised mr harmonization by learning disentangled representations using information bottleneck theory," *NeuroImage*, vol. 243, p. 118 569, 2021.

[12] B. E. Dewey, L. Zuo, A. Carass, Y. He, Y. Liu, E. M. Mowry, S. Newsome, J. Oh, P. A. Calabresi, and J. L. Prince, "A disentangled latent space for cross-site mri harmonization," in *International conference on medical image computing and computer-assisted intervention*, Springer, 2020, pp. 720–729.

[13] L. Zuo, Y. Liu, Y. Xue, B. E. Dewey, S. W. Remedios, S. P. Hays, M. Bilgel, E. M. Mowry, S. D. Newsome, and P. A. Calabresi, "Haca3: A unified approach for multi-site mr image harmonization," *Computerized Medical Imaging and Graphics*, vol. 109, p. 102 285, 2023.

[14] S. Kida, T. Nakamoto, M. Nakano, K. Nawa, A. Haga, J. Kotoku, H. Yamashita, and K. Nakagawa, "Cone beam computed tomography image quality improvement using a deep convolutional neural network," *Cureus*, vol. 10, no. 4, 2018.

[15] X. Liang, L. Chen, D. Nguyen, Z. Zhou, X. Gu, M. Yang, J. Wang, and S. Jiang, "Generating synthesized computed tomography (ct) from cone-beam computed tomography (cbct) using cyclegan for adaptive radiation therapy," *Physics in Medicine & Biology*, vol. 64, no. 12, p. 125 002, 2019.

[16] M. Selim, J. Zhang, B. Fei, G.-Q. Zhang, G. Y. Ge, and J. Chen, "Cross-vendor ct image data harmonization using cvh-ct," in *AMIA Annual Symposium Proceedings*, pmid:35308983, vol. 2021, American Medical Informatics Association, 2021, p. 1099.

[17] J. Liu, H. Yan, H. Cheng, J. Liu, P. Sun, B. Wang, R. Mao, C. Du, and S. Luo, "Cbct-based synthetic ct generation using generative adversarial networks with disentangled representation," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 12, p. 4820, 2021, pmid:34888192.

[18] Y. Hu, H. Zhou, N. Cao, C. Li, and C. Hu, "Synthetic ct generation based on cbct using improved vision transformer cyclegan," *Scientific Reports*, vol. 14, no. 1, p. 11 455, 2024.

[19] S. Armato III *et al.*, *Data from lidc-idri [data set]. the cancer imaging archive*, 2015.

[20] W. Van Aarle, W. J. Palenstijn, J. De Beenhouwer, T. Altantzis, S. Bals, K. J. Batenburg, and J. Sijbers, "The astra toolbox: A platform for advanced algorithm development in electron tomography," *Ultramicroscopy*, vol. 157, pp. 35–47, 2015.

[21] W. Van Aarle, W. J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. De Beenhouwer, K. J. Batenburg, and J. Sijbers, "Fast and flexible x-ray tomography using the astra toolbox," *Optics express*, vol. 24, no. 22, pp. 25 129–25 147, 2016.

[22] L. A. Feldkamp, L. C. Davis, and J. W. Kress, "Practical cone-beam algorithm," *Josa a*, vol. 1, no. 6, pp. 612–619, 1984.

[23] S. Hatamikia, A. Biguri, G. Kronreif, J. Kettenbach, T. Russ, H. Furtado, L. K. Shiyam Sundar, M. Buschmann, E. Unger, M. Figl, *et al.*, "Optimization for customized trajectories in cone beam computed tomography," *Medical physics*, vol. 47, no. 10, pp. 4786–4799, 2020.

[24] G. D. Hugo, E. Weiss, W. C. Sleeman, S. Balik, P. J. Keall, J. Lu, and J. F. Williamson, "Data from 4d lung imaging of nsclc patients," *The Cancer Imaging Archive*, vol. 10, K9, 2016.

[25] G. D. Hugo, E. Weiss, W. C. Sleeman, S. Balik, P. J. Keall, J. Lu, and J. F. Williamson, "A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer," *Medical physics*, vol. 44, no. 2, pp. 762–771, 2017.

[26] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, Springer, 2020, pp. 319–345.

[27] F. Pérez-García, R. Sparks, and S. Ourselin, "Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, p. 106 236, 2021, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2021.106236. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260721003102.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] E. Brion, J. Léger, A. M. Barragán-Montero, N. Meert, J. A. Lee, and B. Macq, "Domain adversarial networks and intensity-based data augmentation for male pelvic organ segmentation in cone beam ct," *Computers in Biology and Medicine*, vol. 131, p. 104 269, 2021.

[30] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[32] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" In *2012 Fourth international workshop on quality of multimedia experience*, IEEE, 2012, pp. 37–38.

[33] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool," *BMC medical imaging*, vol. 15, no. 1, pp. 1–28, 2015.

[34] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[35] R. F. Woolson, "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.