Mitigation of Motion-Induced Artifacts in Cone Beam Computed Tomography using Deep Convolutional Neural Networks

Mohammadreza Amirian, Javier A. Montoya-Zegarra, Thilo Stadelmann[1], Frank-Peter Schilling
(Zurich University of Applied Sciences ZHAW, Centre for Artificial Intelligence CAI, Winterthur, Switzerland)
Lukas Lichtensteiger, Ivo Herzig, Peter Eggenberger Hotz, Marco Morf, Alexander Züst, Rudolf Marcel Füchslin[1]
(Zurich University of Applied Sciences ZHAW, Institute for Applied Mathematics and Physics IAMP, Winterthur, Switzerland)
Pascal Paysan, Igor Peterlik, Stefan Scheib
(Varian Medical Systems Imaging Laboratory GmbH, Baden-Dättwil, Switzerland)

Version typeset December 4, 2022; Corresponding author F.-P. Schilling, scik@zhaw.ch

Abstract

Background: Cone-beam Computed Tomography (CBCT) is used for the reconstruction of images acquired by radiation therapy treatment devices (linear accelerators) in image-guided radiation therapy (IGRT). For each treatment session, it is necessary to obtain the image of the day in order to accurately position the patient, and to enable adaptive treatment capabilities including auto-segmentation and dose calculation. Reconstructed CBCT images often suffer from artifacts, in particular those induced by patient motion. Deep-learning based approaches promise ways to mitigate such artefacts.

Purpose: We propose a novel deep-learning based approach with the goal to reduce motion induced artefacts in CBCT images and improve image quality. It is based on supervised learning and includes neural network architectures employed as pre- and/or post-processing steps during CBCT reconstruction.

Methods: Our approach is based on deep convolutional neural networks which complement the standard CBCT reconstruction, which is performed either with the analytical Feldkamp-Davis-Kress (FDK) method, or with an iterative algebraic reconstruction technique (SART-TV). The neural networks, which are based on refined U-net architectures, are trained end-to-end in a supervised learning setup. Labeled training data are obtained by means of a motion simulation, which uses the two extreme phases of 4D CT scans, their deformation vector fields, as well as time-dependent amplitude signals as input. The trained networks are validated against ground truth using quantitative metrics, as well as by using real patient CBCT scans for a qualitative evaluation by clinical experts.

Results: The presented novel approach is able to generalize to unseen data and yields significant reductions in motion induced artifacts as well as improvements in image

[1] Also with European Centre for Living Technology, Venice, Italy

quality compared with existing state-of-the-art CBCT reconstruction algorithms (up to +6.3 dB and +0.19 improvements in PSNR and SSIM, respectively), as evidenced by validation with an unseen test dataset, and confirmed by a clincal evaluation on real patient scans (up to 74% preference in motion artifact reduction over standard reconstruction).

Conclusions: For the first time, it is shown that inserting deep neural networks as pre- and post-processing plugins in the existing CBCT reconstruction and trained end-to-end can yield significant improvements in image quality and reduction of motion artifacts.

Contents

I. Introduction

Computed tomography (CT) has become a versatile imaging technique in radiology and radiation therapy. In particular, cone-beam CT (CBCT) is used for the reconstruction of images acquired by radiation therapy treatment devices (linear accelerators) in image-guided radiation therapy (IGRT)[1] and by interventional radiology and intraoperative C-arm systems, providing higher spatial resolution in a cost-efficient way[2]. In IGRT, treatment is performed in up to 40 sessions. For each treatment session, it is necessary to obtain the image of the day in order to accurately position the patient. Besides, novel applications of CBCT imaging in IGRT such as online adaptive replanning[3] or daily treatment planning and dose calculation[4] have been proposed.

There are two main families of reconstruction algorithms used in modern CT scanners: (*i*) analytical techniques and (*ii*) iterative algebraic algorithms. The first group is based on filtered backprojection, and most prominently represented by the Feldkamp-Davis-Kress (FDK) method[5]. The second group consists of algorithms based on a reformulation of the reconstruction as an optimization problem. Although the development of iterative methods started in late 1960s[6], they have been employed on CT scanners only over the last 15 years[7,8] mainly because of their high computational cost. In recent years, this problem was solved due to the availability of GPUs. Iterative reconstruction algorithms such as iCBCT introduced in Ref.[9] for Varian's Halcyon® and TrueBeam® addressed the need for superior image quality compared with FDK, as demonstrated in Refs.[10,11,12,13] in terms of better noise suppression and improved contrast.

Imaging artifacts[14] are still a prevalent complication in CBCT reconstruction. The main sources of artifacts are (*i*) electrical and photon count noise, (*ii*) photons from scattered X-rays, (*iii*) extinction and beam hardening effects (e.g., due to metal implants), (*iv*) approximations in the reconstruction (due to finite beam width and detector pixel size), (*v*) aliasing (due to finite pixel size and cone beam divergence), (*vi*) ring artifacts (due to defect or miscalibrated detector elements), and (*vii*) patient motion. Motion artifacts arise since the reconstruction assumes that the scanned patient is stationary. However, periodic respiratory or cardiac (breathing and heart beat in the chest and lung region) and non-periodic (abrupt motion of the patient, gas bubbles in the abdomen and the digestive system) motion leads to acquiring projections from different states of motion. This leads to evident and undesirable, typically streak-shaped image artifacts after reconstruction. The following motion compensation strategies are used so far in IGRT clinical routine: (*i*) 4D or gated CBCT based on an external breathing signal[15], (*ii*) breath hold CBCT based on an external breathing signal and potential patient feedback, (*iii*) assisted breathing based on a ventilator system[16], (*iv*) abdominal compression devices applied to the patient[17], and (v) internal breathing signal extraction[18].

In this paper, we present a novel approach to mitigate motion artifacts in CBCT reconstruction based on deep learning. We embed the CBCT reconstruction within a deep learning pipeline, where convolutional neural networks are employed as pre- and/or post-processing steps. Those networks act on either the 2D X-ray projections (preprocessing), the reconstructed 3D volume (postprocessing), or on both. They are trained end-to-end in a supervised fashion using CBCT scans containing simulated motion, and providing a motion-free state as ground truth. We show that the presented novel approach is able to generalize to unseen data and yields significant reductions in motion induced artifacts as well as improvements in image quality compared with existing state-of-the-art CBCT reconstruction algorithms (up to +6.3 dB and +0.19 improvements in PSNR and SSIM, respectively), as evidenced by validation with an unseen test dataset, and confirmed by a qualitative clinical evaluation on real patient scans (up to 74% preference in motion artifact reduction).

I.A. Related work

Much research has been done[14,19,20] regarding the characterization and mitigation of the various kinds of artifacts which negatively impact image quality in CT and CBCT reconstruction. In recent years, deep-learning based approaches have shown promising results, including applications for IGRT and adaptive radiation therapy[21].

In Ref.[22], the components of the filtered back-projection (FBP) algorithm are mapped into a neural network by introducing a novel deep-learning enabled cone-beam back-projection layer. The backward pass of the layer is computed as a forward projection operation. The approach thus permits joint optimization of correction steps in both volume and projection domains. More formally, in Ref.[23] it is argued specifically that implementing prior knowledge (such as the back-projection operation) in the form of (differentiable) known operators into a deep learning algorithm reduces training error bounds while reducing the number of free parameters.

Limited-angle CT is employed to reduce the acquisition time and to decrease the radiation dose, which leads to a degradation of image quality and the introduction of artifacts. To overcome these issues, a recent approach presented in Ref.[24] uses an encoder-decoder architecture based on the U-net model[25] to reconstruct high-quality images. Images reconstructed using the simultaneous algebraic reconstruction (SART) method[26] are processed by a U-net to improve the image quality. Experiments on chest and abdomen CT scans demonstrated the superiority of the proposed method over existing approaches. Similarly, in Ref.[27] U-net-based networks were employed to correct limited-angle artifacts in circular tomosynthesis scans.

Having gained traction in numerous fields including CT imaging[28,29,30], deep-learning approaches have been used for *metal artifact reduction (MAR)*, e.g. in Refs.[31,32]. In Ref.[33], a dual-domain architecture (DuDoNet) was introduced to jointly compensate for metal-induced artifacts in both projection and volume domains. Experimental results on the DeepLesion CT dataset[34] showed that the proposed method outperformed both traditional and other deep-learning approaches. An improved model (DuDoNet++) was proposed[35] to compensate for over-smoothed and distorted image reconstruction and leads to improved artifact correction, especially for large metallic objects. There have also been recent efforts in MAR using unsupervised approaches, for instance the ADN model[36], which consists of a novel generative adversarial network that disentangles metal artifacts from body tissues and generates different types of artifact-affected and artifact-free CT scans in the image domain. Experimental results show that the proposed model achieves comparable results with existing supervised models. The U-DuDoNet model[37] directly models the artifact generation and compensation process in both the projection and image domains. More recently, interactive and interpretable versions of DuDoNet called InDuDoNet[38] and IDOL-Net[39] were introduced, where the former tries to improve the interpretability of the previous models and the latter aims at enhancing the interaction between projection and image domains.

Neural network based approaches have been employed to improve *sparseness artifacts* originating from low-dose CT reconstruction[40,41,42,43]. In Refs.[44,45], a new method called AirNet is presented which fuses analytical and iterative CT reconstruction integrated with deep learning to improve sparse-data 3D and 4D CBCT reconstruction. In the projection domain, deep-learning based correction of signal degradation caused by X-ray photons that are scattered within the patient body (*scatter artifacts*) has been employed[46,47]. Other examples of the application of deep learning in CT reconstruction include Refs.[48,49,50].

Finally, the compensation of *motion artifacts* using deep learning so far has received comparatively less interest. In Ref.[51], an initial study was presented consisting of a U-net based artifact reduction method in the volume domain. In Ref.[52], a U-net-based neural network is employed to compensate simulated motion artifacts in head CT scans, based on simple simulated rigid (translations, rotations, oscillations) transformations. In Ref.[53], motion artifacts in cine cardiac MRI are reduced using recurrent neural networks, while Ref.[54] addresses cardiovascular motion in short-scan CT by means of a deep partial angle-based motion compensation (Deep PAMoCo) framework.

II. Materials and Methods

II.A. CBCT Reconstruction

To reconstruct a 3D CBCT volume from 2D cone-beam projections (which we here assume to have already been corrected based on knowledge of the acquisition hardware, e.g. for beam hardening and scattering), both analytical and iterative methods are considered. *Feldkamp-Davis-Kress*[5] (FDK) is an analytical reconstruction method based on filtered back-projection (FBP). Although the *Tuy* data-sufficiency conditions[55] are not met for circular trajectories of a cone-beam source, FDK provides a fast and reliable approximation of the inverse Radon transform and has become a gold standard for 3D CBCT reconstruction[56]. In our implementation, the *Ram-Lak filter* is used to compensate for the radial non-uniformity of the sampling density and additional filtering is applied to the projections: Since FDK is applied to datasets acquired with half-fan geometry–i.e., a full 360° trajectory with detector shifted to one direction to increase the field of view–it is necessary to apply *half-fan weighing* to avoid the duplicity of data. This is followed by *cosine weighting* to decrease the longitudinal fall-off effect due to the cone-beam geometry. Finally, the projections are down-sampled so that their resolution matches the cut-off frequency requirement given by the target resolution of the reconstructed volume.

Besides FDK, we also use the *algebraic reconstruction technique* (ART) which is an iterative method originally based on the Kaczmarz algorithm[57]. It approximates the volume $\mathbf{f}$ by an iterative optimization of the data-fidelity cost function $|\mathbf{Af}-\mathbf{p}|^2$ where $\mathbf{A}$ and $\mathbf{p}$ represent the forward-projection operator and projection in attenuation space, respectively. In each iteration k, an update of the actual volume estimation is computed through the back-projection of the gradient of the cost function, i.e., $\sum_\alpha \mathbf{A}^\top([\mathbf{Af}_k]_\alpha - \mathbf{p}_\alpha)$ where $\mathbf{p}_\alpha$ and $[\mathbf{Af}_k]_\alpha$ denote the projection under angle α and corresponding forward-projection of actual volume estimation $\mathbf{f}_k$, respectively, and $\mathbf{A}^\top$ represents the back-projection operator. One of the advantages of iterative methods is that they allow for a straightforward injection of prior knowledge into the reconstruction process through a regularization term augmenting the cost function being optimized. In our implementation, we employ the edge-preserving *total variation (TV)* regularization which helps to reduce noise as well as cone-beam artifacts in the areas far from the iso-center.

In order to significantly reduce the computational cost, our GPU implementation of ART is further accelerated through the following approaches: First, the version of ART known as *simultaneous ART* (SART) is used where the volume is updated in parallel for each input projection. Further, *ordered subsets* (OS)[58] and the Nesterov *momentum method*[59] are employed. Finally, a destination-driven approach[60] is employed in forward projection (only for ART) and backward projection (both ART and FDK). Further details about the TV-regularized OS-SART with momentum (in the following referred to as SART-TV) can be found in Ref.[61] where the method is presented as a part of the iCBCT algorithm deployed clinically in Varian products.

II.B. Motion simulation

To train our models, we use a respiratory motion simulation that generates synthetic sets of CBCT volumes with motion artifacts. The simulation is originally based on Ref.[62]. It uses phase gated 4D CT scans described in Section II.C. and a set of recorded breathing curves. We use *DEEDS*[63] to perform a deformable registration between CT volumes of the end-inhale and end-exhale phases to create a patient-specific deformation vector field (DVF). We deform the CT volumes by scaling the DVFs according to the breathing amplitude at a given time to create a forward projection at each angular step in the simulated CBCT scan. This yields a full set of projections where each projection corresponds to a different respiratory

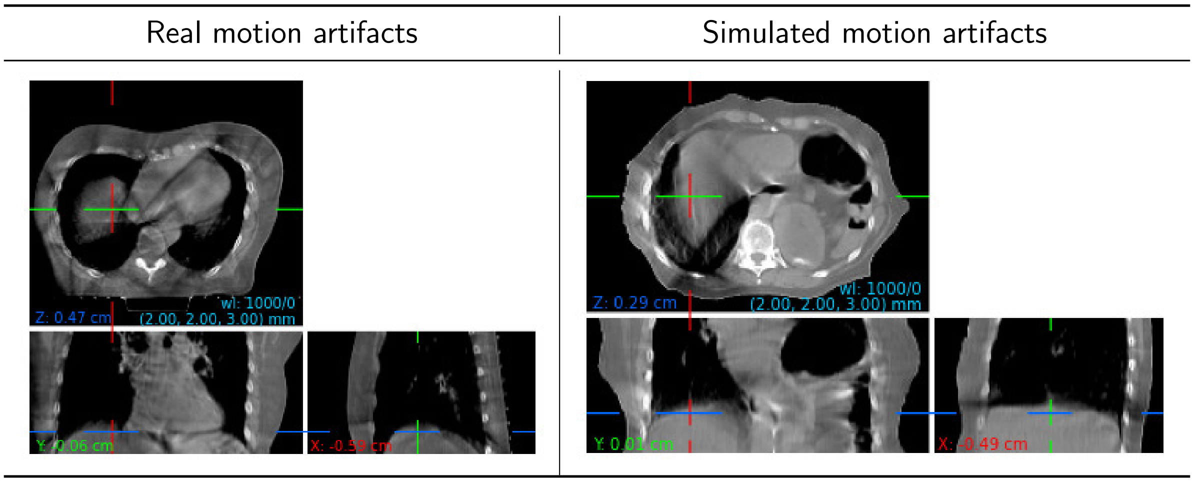


Figure 1: Motion Artifacts. Left: CBCT image with motion artifacts from the test dataset. Right: Image with artificially produced motion artifacts from the motion simulation (images are presented in HU with window and level W/L=1000/0).

state. We then reconstruct a volume using either the FDK or SART-TV reconstruction algorithms to create the CBCT volumes with motion artifacts.

In order to facilitate supervised learning we generate ground-truth volumes which correspond either to a fixed motion state, or are calculated as the average of all deformed volumes. Data augmentation is implemented by applying different breathing curves to the scan, changing the overall motion amplitudes and shifting the field-of-view in z-direction.

Figure 1 shows an example of typical motion artifacts created by patient motion in real CBCT data (test dataset, see section II.C.) side-by-side with the emulated motion artifacts from our motion simulation.

II.C. Datasets

For the training and validation of the different methods, we used a set of thoracic 4D CT scans of 80 patients, split into fractions of 60% (20%, 20%) as *training (validation, test)* datasets. They were provided as input to the motion simulation described in Section II.B.. The patient-specific anatomical correct deformation was extracted from the end in- and exhale out of the 10 breathing phases. To simulate plausible and diverse motion patterns during a virtual CBCT acquisition, we employed a set of 400 recorded breathing curves obtained using the Varian Real-time Position Management® (RPM) system.

For the testing of the developed methods on real CBCT patient scans a set of Halcyon® thoracic CBCT scans was employed (real-world *test dataset*). All pre-processed projection data and reconstructed volumes were given at the same size, resolution, and geometry to ensure consistency: The projection size is 320×80 pixels (resolution of 1.344×4.032 mm), and the volume size is $256 \times 256 \times 48$ voxels ($2 \times 2 \times 3$ mm). The source-to-imager distance is 154 cm with a detector offset of 17.5 cm.

II.D. Deep-learning enabled CBCT reconstruction

This section presents the core methodology used to correct motion artifacts in CBCT images using deep learning. Motion leads to inconsistencies in the acquired projections, which appear as artifacts in the volume domain after reconstruction. Therefore, motion corrections can be, in principle, applied before and/or after reconstruction. These correction steps are implemented as trainable neural network architectures derived from 3D encoder-decoder type architectures. The reconstruction algorithm used is either FDK or iterative CBCT (SART-TV) reconstruction, as discussed in Section II.A.. These algorithms are based on differentiable forward- and backprojection layers implemented with custom CUDA code and interfaced as PyTorch modules. In order to allow back-propagation of gradients in the case of learning in the projection domain, the CBCT reconstruction step has to be fully differentiable, which is not practical for the iterative reconstruction. Thus, projection- and dual-domain motion compensation is restricted to the FDK reconstruction.

We employ a supervised learning approach based on a simulated motion dataset (Section II.B.) for training the motion compensation networks, where the loss is calculated in the volume domain. The ground truth is either calculated as the motion-averaged volume ("average volume") or given as the volume corresponding to the fixed motion state matching the average breathing signal amplitude ("average amplitude"). The networks are validated on the held-out validation and test portions of the simulated motion dataset and on an independent real-world test dataset containing real CBCT scans (see Section II.C.). In detail, the reconstruction pipeline consists of the following components:

Projection Enhancement Network (PE-Net): To mitigate motion-induced artifacts in the projection domain, we rely on convolutional neural networks based on architectures explained in more detail in the next section. PE-Net receives as input the

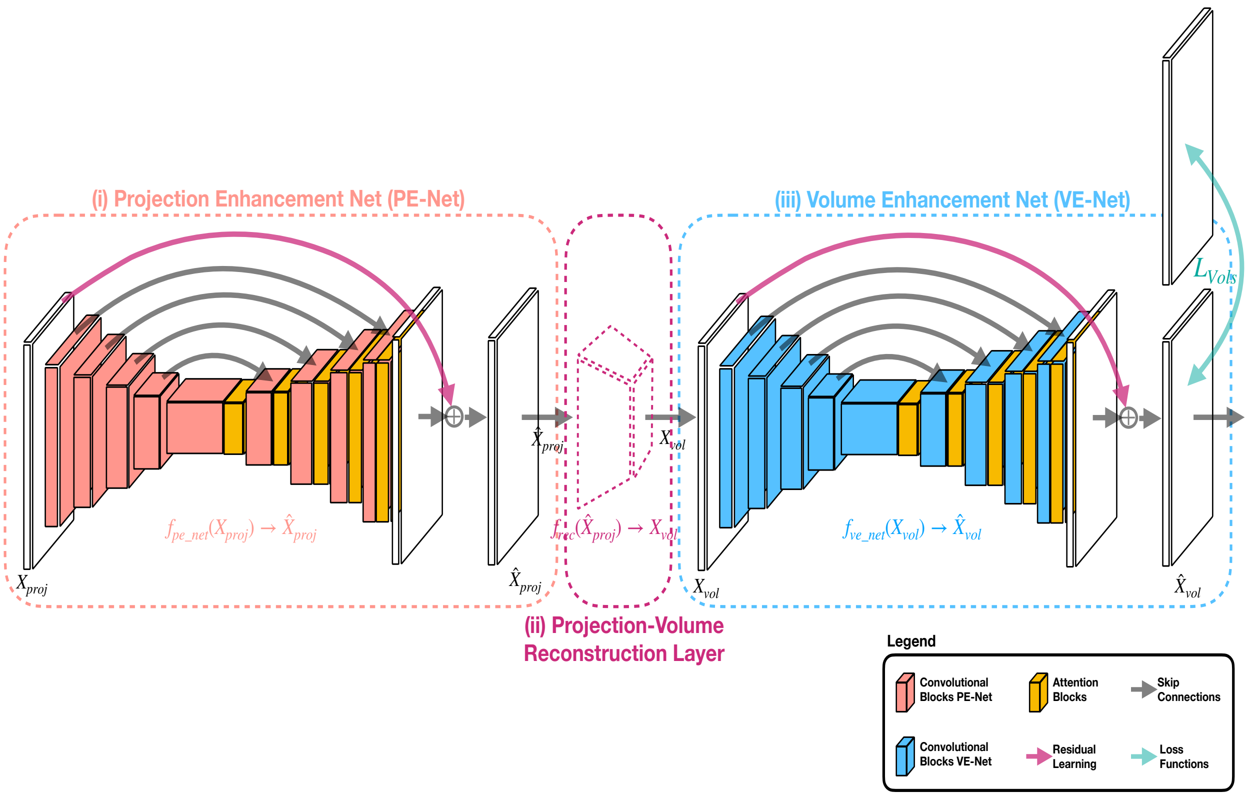


Figure 2: Architecture of the proposed end-to-end model, consisting of (i) a projection enhancement network (PE-Net), (ii) a projection-to-volume reconstruction layer, and (iii) a volume enhancement network (VE-Net).

acquired projections $\{\mathcal{X}_{proj} \in \mathcal{R}^{H_p \times W_p \times C_p}\}$, and enhances these projections $\{\hat{\mathcal{X}}_{proj}\}$, i.e. $f_{pe\_net}(\mathcal{X}_{proj}) \rightarrow \hat{\mathcal{X}}_{proj}$ to remove motion effects in the projection domain. Here, $H_p \times W_p \times C_p$ denote the projection dimensions in terms of height, width, and number of projections.

Projection-to-Volume Reconstruction Layer: The projection-to-volume reconstruction layer $f_{rec}(\cdot)$ receives as input the (enhanced) projections $\{\hat{\mathcal{X}}_{proj}\}$ and outputs a reconstructed volume $\{\mathcal{X}_{vol} \in \mathcal{R}^{H_v \times W_v \times C_v}\}$, i.e. $f_{rec}(\hat{\mathcal{X}}_{proj}) \rightarrow \mathcal{X}_{vol} : \mathcal{R}^{H_p \times W_p \times C_p} \rightarrow \mathcal{R}^{H_v \times W_v \times C_v}$, where $H_v \times W_v \times C_v$ represent the volume's height, width, and number of slices. This layer corresponds to the regular FDK or SART-TV reconstruction (Section II.A.).

Volume Enhancement Network (VE-Net): The VE-Net $f_{ve\_net}(\cdot)$ is responsible for enhancing the reconstructed volume and for compensating motion artifacts in the volume domain. As output, the VE-Net produces an enhanced volume $\{\hat{\mathcal{X}}_{vol} \in \mathcal{R}^{H_v \times W_v \times C_v}\}$, i.e. $f_{ve\_net}(\mathcal{X}_{vol}) \rightarrow \hat{\mathcal{X}}_{vol}$.

Our proposed end-to-end model, shown in Figure 2, combines the above components for motion correction in both projection and volume domain. It consists of three different modules: (*i*) a projection enhancement network (PE-Net), a (*ii*) projection-to-volume reconstruction layer, and a (*iii*) volume enhancement network (VE-Net).

We next describe the different model blocks of our proposed architecture, which is derived from the standard 3D U-net[25] architecture with refinements as discussed below. Note that these blocks are used in both PE-Net and VE-Net.

Encoder Blocks: The encoder block of the presented architecture in Figure 2 consists of four similar submodules including a 3D convolutional layer with filters of size $3 \times 3 \times 3$, followed by an instance normalization[64], the Swish activation function[65] and a 3D max-pooling layer of size $2 \times 2 \times 2$. The number of convolutional filters in the first block is doubled for every next layer; hence the latent representations of the input volume have a larger number of channels but a smaller spatial size with a higher receptive field after the first layer.

Decoder Blocks: The decoder block aims at computing the motion corrections from latent representations and has four submodules starting with a trilinear upsampling followed by a 3D convolutional layer with filters of size $3 \times 3 \times 3$, instance normalization, and Swish activation function. The number of convolutional filters is halved after each layer to make the entire model's architecture symmetric.

Attention mechanisms: To further compensate for motion artifacts, our model relies optionally on attention mechanisms. More precisely, as part of the bottleneck- and decoder-blocks of both Projection Enhancement (PE-Net) and Volume Enhancement (VE-Net) networks, we add channel-wise and spatial attention layers[66] in 3D. At each stage of the decoder, the corresponding input feature maps are multiplied with the generated attention maps to refine the original features. By using these attention layers, the model is capable of focusing on and learning more relevant features. Models including attention layers are denoted "Attn." in Table 1.

Residual Learning: Using residual learning is crucial to simplifying the learning task and improving the convergence speed. The architecture depicted in Figure 2 uses two components to enhance the gradient flow and simplify the learning task. We generally used a direct residual connection from input to output ("residual learning") to optimize the required corrections instead of reconstructing the ground truth. In addition, we optionally used internal residual connections between the input and output of the individual convolutional layers to improve the gradient flow as described in Ref.[67]. Networks including such "ResUNet" layers are labelled as such in Table 1.

II.E. Metrics

In our experiments, we report the numerical performance using several quantitative metrics[68] sensitive to the similarity of pairs of projections or volumes (x, x'). These include Root Mean Squared Error RMSE $= \sqrt{\text{MSE}}$, where $\text{MSE}(x, x') = \frac{1}{N}\sum_i ||x_i - x'_i||^2$, Peak Signal-to-Noise Ratio PSNR $= 10\log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right)$, and Structural Similarity Index (SSIM)[68]. In addition, we quote the mean and standard deviation of the difference image $(x - x')$ used for reducing the motion artifacts. All metrics are calculated in Hounsfield Units (HU) from pairs of uncorrected or corrected body-masked volumes and their corresponding ground truth counterparts.

II.F. Experiments

This section describes the experimental setup, architectural variants, optimization settings and implementation details used.

Experimental Setup: We set the volume size to $256 \times 256 \times 48$ voxels based on the neural network architectures used in this study and to optimize computational and memory costs. Based on the training dataset discussed in Section II.C., we use 720 projections per scan for training, and we add motion artifacts to the original CT volumes using the motion simulation introduced in Section II.B.. The reconstruction and forward projection geometry is selected to match the real-world test dataset as closely as possible, used in this study for clinical evaluation (Section II.C.).

Data Augmentation: We used five different patient breathing curves as input to the motion simulation for each original CT scan in the training dataset. This led to a considerable boost in the final performance of our motion correction models.

Model Architecture: The baseline model we initially considered for motion correction was a U-net with residual learning from input to output as depicted in Figure 2. A plain U-net[25] architecture without residual connections is already sufficient for correcting the artifacts in the volume domain; however, residual learning is necessary for the more complicated tasks, including projection- or dual-domain optimization. Therefore, all of our models include residual learning. We used a U-net base model with a depth of 4 and 32 filters in the first layer. After that, we double the number of filters per layer until the model's bottleneck in the middle and the architecture is reverted afterwards. The same architecture is used for both PE-Net as well as VE-Net. In the case of dual-domain learning, we use a combination of two such models. For PE-Net, the models process the projections in chunks of 192 due to memory limitations. Alternatively, we employed the same architectures, but extended with internal residual connections ("ResUNet") and/or channel-spatial attention ("Attn.").

Implementation and Optimization Settings: We implemented and trained the motion compensation models using the PyTorch[69] framework. The experiments were performed on NVIDIA V100 or A100 GPUs with 32 (40) GB of VRAM. Both projections and volumes are normalized to have zero mean and unit variance. We optimize our models by minimizing the difference between the predicted and reconstructed volume as computed by the l_1-norm$= \sum_i |x_i - x'_i|$ using the AdamW[70] optimizer with a constant learning rate of $1.4 \cdot 10^{-6}$ and weight decay of $1.9 \cdot 10^{-8}$ in the projection domain, and a learning rate of $1.1 \cdot 10^{-4}$ and weight decay of $1.4 \cdot 10^{-8}$ in the volume domain. These parameters result from a joint hyperparameter optimization together with other parameters such as number of convolutional filters, kernel size, or convolutional dilation. We used a batch size of 1 (due to GPU memory limitations) for a total number of 300 epochs. After the training, we select the model that reduces the validation loss the most.

III. Results

III.A. Quantitative Results

In order to train our neural network architectures (Figure 2) in a supervised scenario, we used the training set of the simulated motion dataset (Section II.C.). Table 1 presents the numerical performance of the architectures discussed in Section II. for the two reconstruction methods FDK and SART-TV, with two different sets of ground truth volumes ("average vol-

Model Architecture	RMSE ↓	PSNR (dB) ↑	SSIM ↑	Mean±stdev
Baseline (Average Volume GT)				
FDK	77.8875	28.3802	0.8086	-
SART-TV	76.2560	28.6741	0.8701	-
Baseline (Average Amplitude GT)				
FDK	86.9695	27.5059	0.7992	-
SART-TV	106.5914	25.6087	0.7304	-
Volume-Domain (Average Volume GT)				
3D-UNet (FDK)	**38.27(-39.62±9.06)**	**34.72(6.34±1.45)**	**0.9585(0.1499±0.0412)**	0.0154±38.2148
3D-ResUNet (FDK)	39.86(-38.03±10.53)	34.32(5.94±1.63)	0.9495(0.1410±0.0457)	-8.2486±38.8685
3D-ResUNet+Attn.(FDK)	39.65(-38.24±8.58)	34.35(5.97±1.17)	0.9559(0.1473±0.0406)	-1.9394±39.5164
3D-UNet (SART-TV)†	**44.20(-32.05±14.65)**	**33.32(4.65±1.79)**	**0.9481(0.0780±0.0400)**	-3.7927±43.9936
3D-ResUNet (SART-TV)	44.80(-31.46±14.67)	33.22(4.54±1.80)	0.9464(0.0763±0.0385)	-1.9903±44.7111
3D-ResUNet+Attn.(SART-TV)	45.75(-30.50±15.01)	33.05(4.37±1.89)	0.9377(0.0676±0.0406)	-6.0158±45.2901
Volume-Domain (Average Amplitude GT)				
3D-UNet (FDK)	51.67(-35.30±11.08)	32.10(4.59±1.10)	0.9410(0.1418±0.0431)	-3.5407±51.4552
3D-ResUNet (FDK)	**51.28(-35.69±11.87)**	**32.14(4.63±1.16)**	**0.9417(0.1425±0.0432)**	-2.9049±51.1370
3D-ResUNet+Attn.(FDK)	51.87(-35.10±11.78)	32.03(4.52±1.15)	0.9326(0.1335±0.0456)	-6.9976±51.2475
3D-UNet (SART-TV)†	**55.42(-51.17±11.50)**	**31.42(5.81±1.33)**	**0.9300(0.1996±0.0656)**	0.7139±55.2177
3D-ResUNet (SART-TV)	55.76(-50.83±12.06)	31.35(5.75±1.39)	0.9282(0.1979±0.0634)	-4.0567±55.4900
3D-ResUNet+Attn.(SART-TV)	58.78(-47.81±11.28)	30.88(5.27±1.28)	0.9131(0.1828±0.0598)	-11.9311±57.1327
Projection-Domain (Average Volume GT)				
3D-UNet (FDK)	73.88(-4.01±1.88)	28.89(0.51±0.33)	0.8654(0.0569±0.0165)	3.8085±73.5703
3D-ResUNet (FDK)	67.91(-9.98±4.86)	29.68(1.30±0.78)	0.8931(0.0845±0.0224)	-1.2820±67.7729
3D-ResUNet+Attn.(FDK)	**67.68(-10.21±7.28)**	**29.71(1.33±0.98)**	**0.8940(0.0855±0.0232)**	-1.5657±67.5189
Dual-Domain (Average Volume GT)				
3D-UNet (FDK)	49.19(-28.70±6.19)	32.43(4.05±0.62)	0.9377(0.1292±0.0349)	-0.2131±48.9999
3D-ResUNet (FDK)	**45.51(-32.38±8.13)**	**33.07(4.69±0.73)**	**0.9425(0.1339±0.0406)**	-8.9502±44.4396
3D-ResUNet+Attn.(FDK)	45.65(-32.24±9.07)	33.00(4.62±0.82)	0.9396(0.1311±0.0425)	-9.7962±44.3982

Table 1: Quantitative results of deep-learning based motion correction for CBCT data with simulated motion. The table presents the performance of our proposed motion reduction framework based on the RMSE, PSNR, and SSIM metrics, as well as the mean and standard deviation of the body-masked difference (correction) volumes. The metrics are calculated between the reconstructed and ground truth volumes (using either "average volume" or "average amplitude" ground truth (GT), see text), converted to HU with slope and intercept of 48200 and -1106, respectively. All numerical values are averaged over the test set. To make the contribution of the motion correction clearer, we report the average metric together with the average gain (or loss), as well as the standard deviation of the latter. For example, in the last row, the average PSNR is reported as 33.00 dB, corresponding to an average improvement of 4.62 dB, with a standard deviation of 0.82 dB. The models noted by † are used for clinical evaluation (Section III.B.).

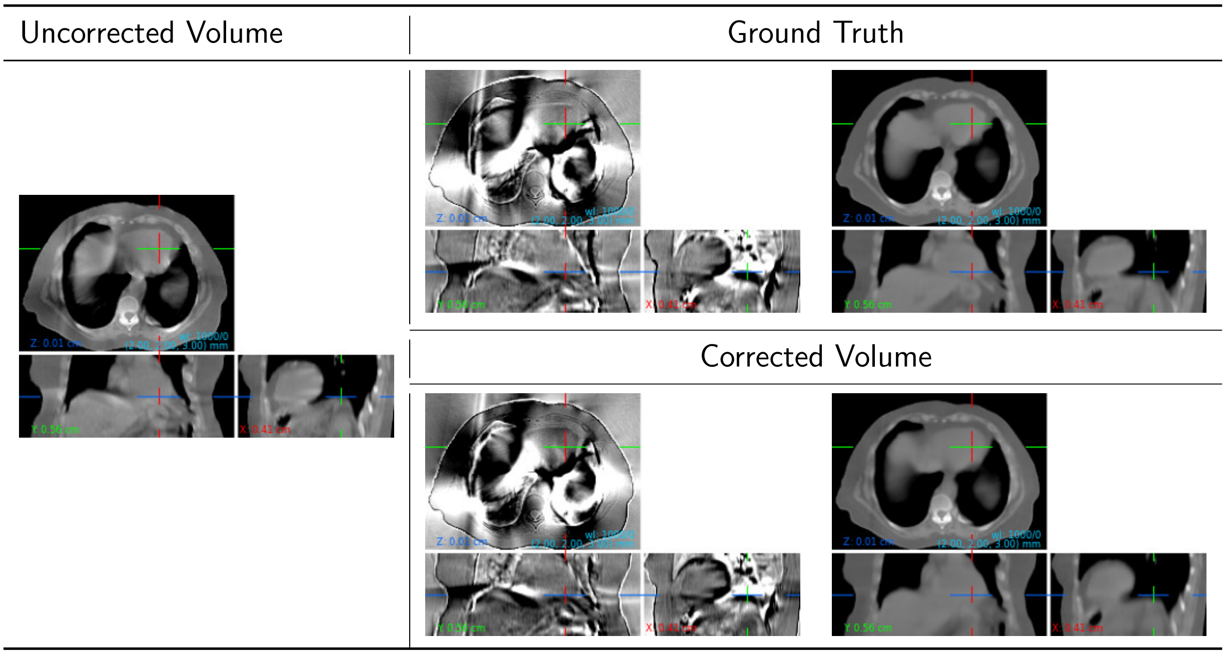


Figure 3: Example result for FDK reconstruction (volume domain optimization). Presented are the uncorrected volume using default reconstruction (left), the ground truth volume, both as difference and absolute image, ("average volume", top right), as well as the corrected volume and its difference (bottom right). Images are presented in HU with W/L=1000/0.

ume" or "average amplitude"). Three different neural network architectures are employed for experiments in projection-, volume- and dual-domain: "3D-UNet" (base architecture), "3D-ResUNet" (base enhanced with ResUNet), and "3D-ResUNet+Attn." (base enhanced with both ResUNet and attention blocks). The ground truth volumes with average amplitude differ more from their corresponding uncorrected volumes with motion artifacts than the ones with averaged volume. Therefore, the baseline RMSE is larger for average amplitude, and lower baseline performances in terms of PSNR and SSIM are reported in Table 1. Since computing the gradients in the backward pass of the reconstruction algorithm, which is required for training models in the projection-domain, is only practical for the FDK reconstruction, we do not report results based on SART-TV for optimizing in projection- and dual-domain. The numerical results are reported based on computing the metrics as introduced in Section II.E. between the body-masked ground truth and reconstructed volumes, converted to HU.

The numerical evaluation demonstrates that training 3D CNNs is consistently successful in compensating motion for deep learning in the projection, volume and dual domain,

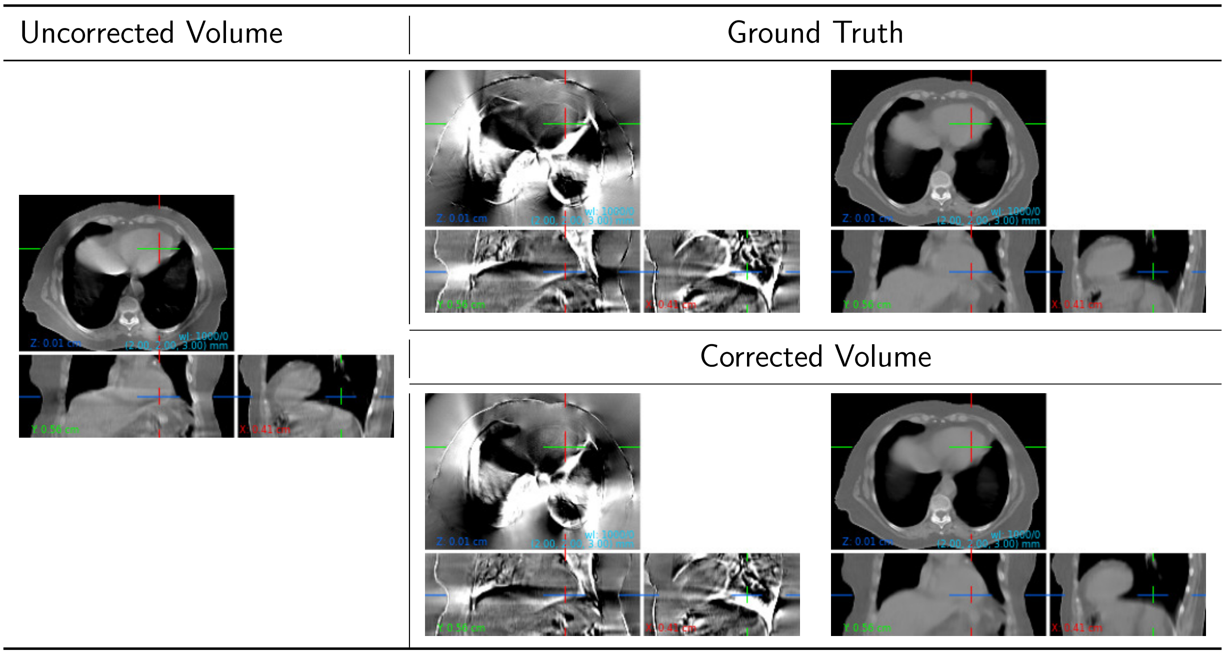


Figure 4: Example result for SART-TV reconstruction (volume domain optimization). Presented are the uncorrected volume using default reconstruction (left), the ground truth volume, both as difference and absolute image, ("average volume", top right), as well as the corrected volume and its difference (bottom right). Images are presented in HU with W/L=1000/0.

and the best performance is achieved in the volume domain. Numerically, it corresponds for FDK to an improvement of +6.34 dB in PSNR and +0.1499 for SSIM with "average volume" ground truth. The highest improvement reported for SART-TV is +5.81 dB in PSNR and +0.1996 for SSIM with "average amplitude" ground truth. We also observed a very competitive performance in dual domain optimization. However, most of the motion correction performance in the dual domain setting is based on the volume domain corrections.The maximum average gained PSNR in the case of pure projection domain optimization turned out to be +1.33 dB.

The above results represent the first successful attempt at reducing motion artifacts globally in CBCT scans using deep neural networks. The proposed method reduces motion artifacts for two reconstruction techniques (FDK and SART-TV), and with several different architectures, including variants with added internal residual connections and/or channel-spatial attention. The motion compensation performance shows a small but consistent variance with the details of the neural network architecture.

Comparing the two CBCT reconstruction algorithms, SART-TV shows more robustness against motion during acquisition time, and a slightly lower drop in baseline performance is reported. Motion artifact reduction using 3D CNNs in the volume domain for SART-TV reconstruction is successful and performs better compared with FDK reconstruction. Figures 3 and 4 present example visualizations of the observed motion artifact improvements in volume domain learning applied to the FDK and SART-TV reconstructed volumes, respectively.

III.B. Clinical Evaluation

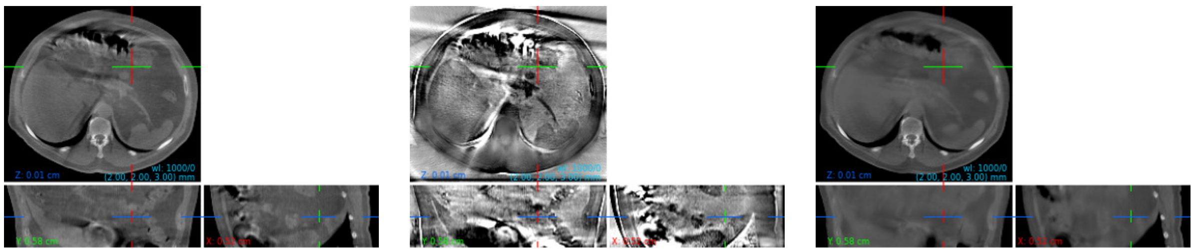
To validate the quantitative results of the previous section in a clinical setting, we applied the trained motion compensation CNN models to a real-world test dataset (see Section II.C. and Figure 5) and evaluated the performance based on the feedback obtained from clinicians in clinical routine. The real-world CBCT scans used in this study are sufficiently different from the simulated training dataset to judge the models' generalization capabilities, e.g. concerning projection count and HU calibration. To compensate for the different calibration, we rescaled the attenuation values of the real-world test dataset to a scale matching the one of the training dataset.

To collect the clinicians' feedback, we provided them with 30 pairs of SART-TV reconstructed and motion-corrected volumes, 15 each using either average-amplitude or average-volume as ground truth. We computed the motion corrections based on the developed motion compensation framework and using the best-performing CNN architectures, i.e. U-net in the volume domain without residual connections or attention, from Table 1. Subsequently, in total 20 clinicians – including radiation oncologists, medical physicists, radiation technologists and physicians – answered several questions about their preferences for using CNN models to reduce motion artifacts compared with the standard reconstruction. The clinicians identified themselves into three general categories of medical physician (26%), physicist (37%), or dosimetrist/radiation technician (37%).

Initial feedback received on the SART-TV datasets indicated the presence of severe and mild unavoidable real-world artifacts besides motion in 34% and 20% of the scans, respectively. The study participants were asked to indicate their level of agreement or preference with respect to (a) a reduction of the observed motion artifacts and (b) the usage of motion-corrected volumes for various applications including dose calculation, patient

Ground Truth: Average Volume

Uncorrected Volume	Residual Corrections	Corrected Volume



Ground Truth: Average Amplitude

Uncorrected Volume	Residual Corrections	Corrected Volume

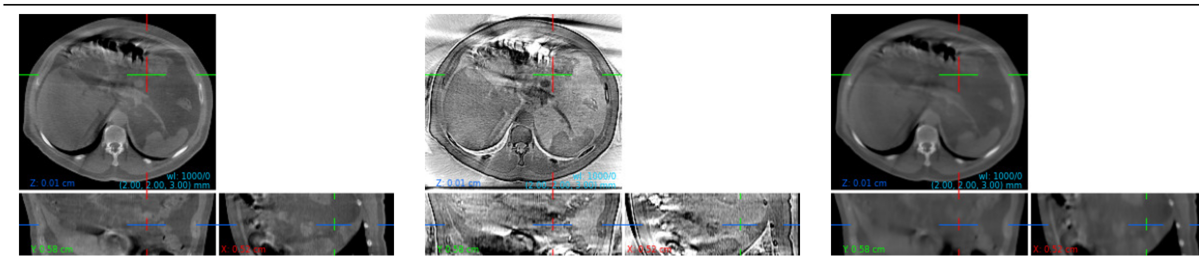


Figure 5: Example results for SART-TV reconstruction for real-world test dataset, using the two options for the choice of ground truth. Presented are the uncorrected volumes using default reconstruction (left), the residual corrections (middle), as well as the corrected volumes (right).

positioning or segmentation.

This clinical evaluation, the first of its kind to the best of our knowledge, faced the challenge of subjective assessments from experts with different clinical backgrounds. For example, physicians reported a noticeable or strong improvement in CNN-based motion artifact reduction using average volume ground truth in 80% of the scans, while for medical physicists this number is only 66%. On the other hand, medical physicists expressed preference for using CNN-corrected volumes for dose calculation in 63% of the cases, while physicians reported only 31%.

We averaged all votes and present the final results in Table 2. Despite the differences in the improvements reported by the different expert groups, there is a clear positive trend that the proposed CNN models are indeed able to reduce motion artifacts successfully. In addition, clinicians reported a weak tendency toward using CNN-corrected images (computed by models trained using average volumes as ground truth) for plan adaptation and dose

Ground Truth → ↓ Application / Preference →	Average Volume			Average Amplitude		
	CNN (%)	Equal (%)	Standard (%)	CNN(%)	Equal(%)	Standard(%)
Motion artifact reduction	**74.00**	26.00	-	58.33	41.67	-
Plan adaptation and dose calculation	**49.33**	22.00	28.67	26.33	17.33	56.33
Soft-tissue-based patient positioning	23.00	12.67	64.33	13.00	7.00	80.00
Manual and automatic tissue segmentation	24.33	14.67	61.00	13.00	10.33	76.67

Table 2: Results of the clinical evaluation. Presented are preferences for CNN-based or default SART-TV reconstruction when training CNN models using either average volume or average amplitude ground truth. The clinicians expressed on their opinion on the capability of CNN-based models for motion artifact reduction, as well as for potential applications such as plan adaptation and dose calculation, patient positioning or segmentation.

calculation. On the other hand, clinical experts expressed a preference to rather use images without CNN-based reconstruction for soft-tissue-based patient positioning as well as for manual or automatic tissue segmentation, as these images are typically sharper compared with the CNN-corrected ones.

In response to the above result, we decided to perform a quantitative evaluation to compute the level of agreement between CBCT images with and without motion artifact correction when applying an automatic segmentation algorithm to both sets of scans. We computed the average dice score over 18 organs or tissues which are visible in most of the CBCT images, including pulmonary arteries, breast, chest wall, lung, ribs and spinal canal. The high dice score of 0.89 (0.88) when using average volume (average amplitude) ground truth demonstrates a very high level of consistency between the obtained segmentation contours, despite the low preference reported by clinical experts to use the motion corrected images for segmentation.

IV. Conclusion

In this paper, we presented, for the first time to the best of our knowledge, a deep-learning based method for globally reducing motion artifacts in reconstructed 3D CBCT images, building on top of the two reconstruction algorithms FDK and SART-TV.

We implemented neural network architectures which act either on the reconstructed CBCT volumes, on the input X-ray projections, or on both for end-to-end dual-domain optimization. The proposed models were trained in a supervised way using a motion simulation framework that provides motion-free ground truth. The experimental results clearly demonstrate that motion artifacts can be corrected via deep learning. So far, the best results were obtained with the volume-domain based correction network, implementing a refined U-net architecture.

The quantitative evaluations demonstrate that the application of deep learning methods can yield significant improvements in imaging quality and reduction of motion-induced artifacts in reconstructed CBCT scans. In addition, a clinical evaluation was performed, in which clinical experts confirmed the principal quantitative results for motion artifact reduction using a real-world test dataset. While they confirmed that artifacts are reduced, and they expressed a preference for using CNN-corrected CBCT scans for dose calculation, for other applications including patient positioning or segmentation, this could not yet be demonstrated in this initial study.

There are several avenues for future research: First, the presented results show promising improvements mostly in the volume domain, independent of the acquisition parameters and reconstruction technique. However, there is room for improvement in the projection and dual-domain settings. One potential reason could be the processing of projections in batches due to GPU memory limitations, which leads to a loss of correlation between different projection batches separately processed by the neural network. In addition, great care has to be taken to ensure the backpropagation of gradients through the CBCT reconstruction layer to provide CNN models with a meaningful, precise and noiseless learning signal in the projection domain.

Second, models trained using supervised learning typically suffer from imperfect generalization to data acquired in entirely different settings[71]. Although the calibration technique we used in this study successfully reduced the performance gap of the models between simulation and real data, generalization to highly different acquisition setups and other anatomies is not granted. This encourages the investigation of unsupervised learning and/or domain adaptation techniques in future research.

Third, our motion simulation currently only simulates thoracic respiratory motion, and does not include other effects such as cardiac motion. Tackling cardiac motion in chest CBCT combined with respiratory motion is still an open problem. Furthermore, extending the presented method to abdominal CBCT requires simulating different kinds of motion artifacts.

IV. CONCLUSION

In conclusion, while the initial results are very promising, future research will aim at further improved deep learning techniques which enable improved adaptive treatment capabilities in IGRT including patient positioning and tumor targeting, auto-segmentation as well as dose calculation applications directly on the treatment device.

V. Acknowledgments

We thank the members of the radiation therapy departments of the following institutes for contributing to the clinical evaluation: University of California Los Angeles, Amsterdam University Medical Center, University Hospital Bern, Campus Bio-Medico University Rome, Assuta Hospital Israel, Alfred Health Radiation Oncology Melbourne, Australia, and Clinique de Grangettes, Genève. We thank Mário Fartaria from Varian Medical Systems Imaging Laboratory for providing the auto-segmentation results on the CBCT evaluation datasets and the comparison metrics. This work was co-financed by Innosuisse, grant no. 35244.1 IP-LS.

Declaration of Conflict of Interest: The following authors are full-time employees of Varian Medical Systems Imaging Laboratory: Pascal Paysan, Igor Peterlik, and Stefan Scheib.

References

1. D. A. Jaffray, J. H. Siewerdsen, J. W. Wong, and A. A. Martinez, Flat-panel cone-beam computed tomography for image-guided radiation therapy, International Journal of Radiation Oncology*Biology*Physics **53**, 1337–1349 (2002).

2. U. V. Elstrøm, L. P. Muren, J. B. 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 **50**, 908–917 (2011).

3. S. Yoon, H. Lin, M. Alonso-Basanta, N. Anderson, O. Apinorasethkul, K. Cooper, L. Dong, B. Kempsey, J. Marcel, J. Metz, R. Scheuermann, and T. Li, Initial Evaluation of a Novel Cone-Beam CT-Based Semi-Automated Online Adaptive Radiotherapy System for Head and Neck Cancer Treatment – A Timing and Automation Quality Study, Cureus **12** (2020).

4. T. Jarema and T. Aland, Using the Iterative kV CBCT Reconstruction on the Varian Halcyon Linear Accelerator for Radiation Therapy-Planning CT Datasets: A Feasibility Study, International Journal of Radiation Oncology*Biology*Physics **105**, E719–E720 (2019), Proceedings of the Amercian Society for Radiation Oncology 61st Annual Meeting.

5. L. A. Feldkamp, L. C. Davis, and J. W. Kress, Practical cone-beam algorithm, J. Opt. Soc. Am. A **1**, 612–619 (1984).

6. G. N. Hounsfield, Method of and apparatus for examining a body by radiation such as x or gamma radiation, (1975).

7. G. K and R. R, SAFIRE: Sinogram Affirmed Iterative Reconstruction, Siemens Healthcare, 2012.

8. T. J-B, Veo: CT Model-Based Iterative Reconstruction, GE Healthcare, 2010.

9. P. Paysan, M. Brehm, A. Wang, D. Seghers, and J. Star-Lack, Iterative image reconstruction in image-guided radiation therapy, 2018, US Patent App. 15/952,996.

[10] S. J. Gardner, W. Mao, C. Liu, I. Aref, M. Elshaikh, J. K. Lee, D. Pradhan, B. Movsas, I. J. Chetty, and F. Siddiqui, Improvements in CBCT Image Quality Using a Novel Iterative Reconstruction Algorithm: A Clinical Evaluation, Advances in Radiation Oncology **4**, 390–400 (2019).

[11] H. Kim, M. S. Huq, R. Lalonde, C. J. Houser, S. Beriwal, and D. E. Heron, Early clinical experience with varian halcyon V2 linear accelerator: Dual-isocenter IMRT planning and delivery with portal dosimetry for gynecological cancer treatments, Journal of Applied Clinical Medical Physics **20**, 111–120 (2019).

[12] W. Mao, C. Liu, S. J. Gardner, F. Siddiqui, K. C. Snyder, A. Kumarasiri, B. Zhao, J. Kim, N. W. Wen, B. Movsas, and I. J. Chetty, Evaluation and Clinical Application of a Commercially Available Iterative Reconstruction Algorithm for CBCT-Based IGRT, Technology in Cancer Research & Treatment **18**, 1533033818823054 (2019), PMID: 30803367.

[13] H. Washio, S. Ohira, Y. Funama, M. Morimoto, K. Wada, M. Yagi, H. Shimamoto, Y. Koike, Y. Ueda, T. Karino, S. Inui, Y. Nitta, M. Miyazaki, and T. Teshima, Metal artifact reduction using iterative CBCT reconstruction algorithm for head and neck radiation therapy: A phantom and clinical study, European Journal of Radiology **132**, 109293 (2020).

[14] R. K. W. Schulze, U. Heil, D. Gross, D. D. Bruellmann, E. Dranischnikow, U. Schwanecke, and E. Schoemer, Artefacts in CBCT: a review., Dento maxillo facial radiology **40 5**, 265–73 (2011).

[15] O. Dillon, P. J. Keall, C.-C. Shieh, and R. T. O'Brien, Evaluating reconstruction algorithms for respiratory motion guided acquisition, Physics in Medicine & Biology **65** (2020).

[16] S. T. H. Peeters, F. Vaassen, C. Hazelaar, A. Vaniqui, E. Rousch, D. Tissen, E. V. Enckevort, M. D. Wolf, M. C. Öllers, W. van Elmpt, K. Verhoeven, J. G. M. V. Loon, B. A. Vosse, D. K. M. D. Ruysscher, and G. Vilches-Freixas, Visually guided inspiration breath-hold facilitated with nasal high flow therapy in locally advanced lung cancer, Acta Oncologica **60**, 567–574 (2021), PMID: 33295823.

[17] M. Daly, A. McWilliam, G. Radhakrishna, A. Choudhury, and C. L. Eccles, Radiotherapy respiratory motion management in hepatobiliary and pancreatic malignancies: a systematic review of patient factors influencing effectiveness of motion reduction with abdominal compression, Acta Oncologica **61**, 833–841 (2022), PMID: 35611555.

[18] A. T. Mohd Amin, S. S. Mokri, R. Ahmad, F. Ismail, and A. A. Abd Rahni, Evaluation Methodology for Respiratory Signal Extraction from Clinical Cone-Beam CT (CBCT) using Data-Driven Methods, International Journal of Integrated Engineering **13**, 1–8 (2021).

[19] F. Boas and D. Fleischmann, CT artifacts: Causes and reduction techniques, Imaging in Medicine **4** (2012).

[20] L. Gjesteby, B. De Man, Y. Jin, H. Paganetti, J. Verburg, D. Giantsoudi, and G. Wang, Metal Artifact Reduction in CT: Where Are We After Four Decades?, IEEE Access **4**, 5826–5849 (2016).

[21] P. Paysan, I. Peterlík, T. Roggen, L. Zhu, C. Wessels, J. Schreier, M. Buchacek, and S. Scheib, Deep Learning Methods for Image Guidance in Radiation Therapy, in *Artificial Neural Networks in Pattern Recognition - 9th IAPR TC3 Workshop, ANNPR 2020, Winterthur, Switzerland, September 2-4, 2020, Proceedings*, edited by F. Schilling and T. Stadelmann, volume 12294 of *Lecture Notes in Computer Science*, pages 3–22, Springer, 2020.

[22] T. Würfl, M. Hoffmann, V. Christlein, K. Breininger, Y. Huang, M. Unberath, and A. K. Maier, Deep Learning Computed Tomography: Learning Projection-Domain Weights From Image Domain in Limited Angle Problems, IEEE Transactions on Medical Imaging **37**, 1454–1463 (2018).

[23] A. Maier et al., Learning with known operators reduces maximum error bounds, Nature Machine Intelligence **1**, 373–380 (2019).

[24] J. Wang, J. Liang, J. Cheng, Y. Guo, and L. Zeng, Deep learning based image reconstruction algorithm for limited-angle translational computed tomography, PLoS ONE **15** (2020).

[25] O. Ronneberger, P.Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351, pages 234–241, Springer, 2015, (available on arXiv:1505.04597 [cs.CV]).

[26] A. Andersen and A. Kak, Simultaneous Algebraic Reconstruction Technique (SART): A superior implementation of the ART algorithm, Ultrasonic Imaging **6**, 81–94 (1984).

[27] A. K. Schnurr, K. Chung, T. Russ, L. R. Schad, and F. G. Zöllner, Simulation-based deep artifact correction with Convolutional Neural Networks for limited angle artifacts, Zeitschrift fur Medizinische Physik **29**, 150–161 (2019).

[28] J. Schmidhuber, Deep learning in neural networks: An overview, Neural networks **61**, 85–117 (2015).

[29] T. Stadelmann, V. Tolkachev, B. Sick, J. Stampfli, and O. Dürr, Beyond ImageNet: deep learning in industrial practice, in *Applied Data Science*, pages 205–232, Springer, 2019.

[30] M. Amirian, J. A. Montoya-Zegarra, J. Gruss, Y. D. Stebler, A. S. Bozkir, M. Calandri, F. Schwenker, and T. Stadelmann, PrepNet: A Convolutional Auto-Encoder to Homogenize CT Scans for Cross-Dataset Medical Image Analysis, in *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–7, IEEE, 2021.

[31] H. S. Park, S. M. Lee, H. P. Kim, J. K. Seo, and Y. E. Chung, CT sinogram-consistency learning for metal-induced beam hardening correction, Medical Physics **45**, 5376–5384 (2018).

[32] Y. Zhang and H. Yu, Convolutional Neural Network Based Metal Artifact Reduction in X-Ray Computed Tomography, IEEE Transactions on Medical Imaging **37**, 1370–1381 (2018).

[33] W.-A. Lin, H. Liao, C. Peng, X. Sun, J. Zhang, J. Luo, R. Chellappa, and S. K. Zhou, DuDoNet: Dual Domain Network for CT Metal Artifact Reduction, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[34] K. Yan, X. Wang, L. Lu, L. Zhang, A. P. Harrison, M. Bagheri, and R. M. Summers, Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-Scale Lesion Database, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9261–9270, 2018.

[35] Y. Lyu, W.-A. Lin, H. Liao, J. Lu, and S. K. Zhou, Encoding metal mask projection for metal artifact reduction in computed tomography, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 147–157, Springer, 2020.

[36] H. Liao, W. A. Lin, S. K. Zhou, and J. Luo, ADN: Artifact Disentanglement Network for Unsupervised Metal Artifact Reduction, IEEE Transactions on Medical Imaging **39**, 634–643 (2020).

[37] Y. Lyu, J. Fu, C. Peng, and S. K. Zhou, U-DuDoNet: Unpaired Dual-Domain Network for CT Metal Artifact Reduction, in *Medical Image Computing and Computer Assisted Intervention*, edited by M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, pages 296–306, 2021.

[38] H. Wang, Y. Li, H. Zhang, J. Chen, K. Ma, D. Meng, and Y. Zheng, InDuDoNet: An Interpretable Dual Domain Network for CT Metal Artifact Reduction, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 107–118, Springer, 2021.

[39] T. Wang, Z. Lu, Z. Yang, W. Xia, M. Hou, H. Sun, Y. Liu, H. Chen, J. Zhou, and Y. Zhang, IDOL-Net: An Interactive Dual-Domain Parallel Network for CT Metal Artifact Reduction, IEEE Transactions on Radiation and Plasma Medical Sciences (2022).

[40] Y. S. Han, J. Yoo, and J. C. Ye, Deep Residual Learning for Compressed Sensing CT Reconstruction via Persistent Homology Analysis, arXiv preprint arXiv:1611.06391 (2016).

[41] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, Deep Convolutional Neural

Network for Inverse Problems in Imaging, IEEE Transactions on Image Processing **26**, 4509–4522 (2017).

[42] Z. Zhang, X. Liang, X. Dong, Y. Xie, and G. Cao, A sparse-view CT reconstruction method based on combination of DenseNet and deconvolution, IEEE transactions on medical imaging **37**, 1407–1417 (2018).

[43] A. Kofler, M. Haltmeier, C. Kolbitsch, M. Kachelrieß, and M. Dewey, A U-Nets Cascade for Sparse View Computed Tomography, in *Machine Learning for Medical Image Reconstruction: First International Workshop, MLMIR 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings*, pages 91–99, Springer, Cham, 2018.

[44] G. Chen, X. Hong, Q. Ding, Y. Zhang, H. Chen, S. Fu, Y. Zhao, X. Zhang, H. Ji, G. Wang, H. Qiu, and H. Gao, AirNet: Fused Analytical and Iterative Reconstruction with Deep Neural Network Regularization for Sparse-Data CT, Medical Physics (2020).

[45] G. Chen, Y. Zhao, Q. Huang, and H. Gao, 4D-AirNet: a temporally-resolved CBCT slice reconstruction method synergizing analytical and iterative method with deep learning, Physics in Medicine & Biology (2020).

[46] J. Maier, S. Sawall, and M. Kachelrieß, Deep scatter estimation (DSE): feasibility of using a deep convolutional neural network for real-time x-ray scatter prediction in cone-beam CT, SPIE Medical Imaging **10573** (2018).

[47] J. Erath, T. Vöth, J. Maier, and M. Kachelrieß, Forward and cross-scatter estimation in dual source CT using the deep scatter estimation (DSE), in *Medical Imaging 2019: Physics of Medical Imaging*, volume 10948, page 24, International Society for Optics and Photonics, 2019.

[48] Y. Huang, A. Preuhs, M. Manhart, G. Lauritsch, and A. Maier, Data Consistent CT Reconstruction from Insufficient Data with Learned Prior Images, arXiv preprint arXiv:2005.10034 (2020).

[49] B. Zhu, J. Z. Liu, B. R. Rosen, and M. S. Rosen, Image reconstruction by domain transform manifold learning, Nature **555**, 487–492 (2018).

[50] L. Fu and B. De Man, A hierarchical approach to deep learning and its application to tomographic reconstruction, in *15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, volume 11072, page 1107202, International Society for Optics and Photonics, 2019.

[51] P. Paysan et al., Convolutional Network Based Motion Artifact Reduction in Cone-Beam CT, in *AAPM annual meeting 2019, e-Poster*, 2019.

[52] B. Su, Y. Wen, Y. Liu, S. Liao, J. Fu, G. Quan, and Z. Li, A deep learning method for eliminating head motion artifacts in computed tomography, Medical Physics **49**, 411–419 (2022).

[53] Q. Lyu, H. Shan, Y. Xie, A. C. Kwan, Y. Otaki, K. Kuronuma, D. Li, and G. Wang, Cine cardiac MRI motion artifact reduction using a recurrent neural network, IEEE Transactions on Medical Imaging **40**, 2170–2181 (2021).

[54] J. Maier, S. Lebedev, J. Erath, E. Eulig, S. Sawall, E. Fournié, K. Stierstorfer, M. Lell, and M. Kachelrieß, Deep learning-based coronary artery motion estimation and compensation for short-scan cardiac CT, Medical Physics **48**, 3559–3571 (2021).

[55] H. K. Tuy, An inversion formula for cone-beam reconstruction, SIAM Journal on Applied Mathematics **43**, 546–552 (1983).

[56] T. M. Buzug, *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*, Springer, 2008.

[57] S. Karczmarz, Angenaherte auflosung von systemen linearer glei-chungen, Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat. , 355–357 (1937).

[58] D. Kim, S. Ramani, and J. A. Fessler, Combining Ordered Subsets and Momentum for Accelerated X-Ray CT Image Reconstruction, IEEE Transactions on Medical Imaging **34**, 167–178 (2015).

[59] Y. Nesterov, Smooth minimization of non-smooth functions, Mathematical programming **103**, 127–152 (2005).

[60] B. Keck, H. G. Hofmann, H. Scherl, M. Kowarschik, and J. Hornegger, High resolution iterative CT reconstruction using graphics hardware, in *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, pages 4035–4040, 2009.

[61] I. Peterlik, A. Strzelecki, M. Lehmann, P. Messmer, P. Munro, P. Paysan, M. Plamondon, and D. Seghers, Reducing residual-motion artifacts in iterative 3D CBCT reconstruction in image-guided radiation therapy, Medical Physics **48**, 6497–6507 (2021).

[62] P. Paysan et al., CT based simulation framework for motion artifact and ground truth generation of Cone-Beam CT, in *AAPM annual meeting 2019, e-Poster*, 2019.

[63] M. P. Heinrich, M. Jenkinson, S. M. Brady, and J. A. Schnabel, MRF-Based Deformable Registration and Ventilation Estimation of Lung CT, IEEE TRANSACTIONS ON MEDICAL IMAGING **32**, 1239–1248 (2013).

[64] D. Ulyanov, A. Vedaldi, and V. Lempitsky, Instance normalization: The missing ingredient for fast stylization, arXiv preprint arXiv:1607.08022 (2016).

[65] P. Ramachandran, B. Zoph, and Q. V. Le, Swish: a self-gated activation function, arXiv preprint arXiv:1710.05941 **7**, 5 (2017).

[66] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, Cbam: Convolutional block attention module, in *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[67] Z. Zhang, Q. Liu, and Y. Wang, Road Extraction by Deep Residual U-Net, CoRR **abs/1711.10684** (2017).

[68] 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 **13**, 600–612 (2004).

[69] A. Paszke et al., PyTorch: An Imperative Style, High-Performance Deep Learning Library, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, pages 8024–8035, Curran Associates, Inc., 2019.

70 I. Loshchilov and F. Hutter, Decoupled Weight Decay Regularization, in *International Conference on Learning Representations, ICLR 2019, New Orleans, United States, May 6-9, 2019*, pages 1–18, 2019.

71 P. Sager, S. Salzmann, F. Burn, and T. Stadelmann, Unsupervised Domain Adaptation for Vertebrae Detection and Identification in 3D CT Volumes Using a Domain Sanity Loss, Journal of Imaging **8** (2022).