Received: 4 December 2022 | Revised: 25 February 2023 | Accepted: 19 March 2023

DOI: 10.1002/mp.16405

RESEARCH ARTICLE

MEDICAL PHYSICS

Mitigation of motion-induced artifacts in cone beam computed tomography using deep convolutional neural networks

Mohammadreza Amirian[1,2] | **Javier A. Montoya-Zegarra**[1] | **Ivo Herzig**[3] | **Peter Eggenberger Hotz**[3] | **Lukas Lichtensteiger**[3] | **Marco Morf**[3] | **Alexander Züst**[3] | **Pascal Paysan**[4] | **Igor Peterlik**[4] | **Stefan Scheib**[4] | **Rudolf Marcel Füchslin**[3,5] | **Thilo Stadelmann**[1,5] | **Frank-Peter Schilling**[1]

[1]Centre for Artificial Intelligence CAI, Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

[2]Institute of Neural Information Processing, Ulm University, Ulm, Germany

[3]Institute for Applied Mathematics and Physics IAMP, Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

[4]Varian Medical Systems Imaging Laboratory GmbH, Baden, Switzerland

[5]European Centre for Living Technology, Venice, Italy

Correspondence
Frank-Peter Schilling, Centre for Artificial Intelligence CAI, Zurich University of Applied Sciences ZHAW, Technikumstrasse 71, 8400 Winterthur, Switzerland.
Email: scik@zhaw.ch

Funding information
Innosuisse - Schweizerische Agentur für Innovationsförderung, Grant/Award Number: 35244.1 IP-LS

Abstract

Background: Cone beam computed tomography (CBCT) is often employed on radiation therapy treatment devices (linear accelerators) used 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 artifacts.

Purpose: We propose a novel deep-learning based approach with the goal to reduce motion induced artifacts 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 quality compared with existing state-of-the-art CBCT reconstruction algorithms (up to +6.3 dB and +0.19 improvements in peak signal-to-noise ratio, PSNR, and structural similarity index measure, SSIM, respectively), as evidenced by validation with an unseen test dataset, and confirmed by a clinical evaluation on real patient scans (up to 74% preference for motion artifact reduction over standard reconstruction).

Conclusions: For the first time, it is demonstrated, also by means of clinical evaluation, that inserting deep neural networks as pre- and post-processing plugins in the existing 3D CBCT reconstruction and trained end-to-end yield significant improvements in image quality and reduction of motion artifacts.

KEYWORDS
cone beam computed tomography, deep learning, motion artifacts

1 | INTRODUCTION

Cone beam computed tomography (CBCT) is a technique often used to acquire volumetric X-ray images on board of radiation therapy treatment devices (linear accelerators) in image-guided radiation therapy (IGRT),[1] as well as 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 CBCT scanners: (*i*) analytical techniques and (*ii*) iterative algebraic algorithms. The first group is inspired by 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 CBCT 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[9] for Varian's Halcyon and TrueBeam addressed the need for superior image quality compared with FDK, as demonstrated[10–13] 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 peak signal-to-noise ratio (PSNR) and structural similarity (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).

1.1 | 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] As the existing literature on deep-learning based CBCT motion compensation is scarce, and the developed methods generally are often applicable to artifact types other than motion, as well as for both CT and CBCT, we extend the discussion beyond the field of CBCT motion artifacts, which is the main focus of our study.

The components of the filtered back-projection (FBP) algorithm were mapped into a neural network by introducing a novel deep-learning enabled cone beam back-projection layer.[22] 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, it has been argued specifically[23] 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.

In *Limited-angle CT*, a recent approach[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. Similarly, U-net-based networks were employed[27] to correct limited-angle artifacts in circular tomosynthesis scans.

Having gained traction in numerous fields including CT imaging,[28–30] deep-learning approaches have been used for *metal artifact reduction* (MAR).[31,32] A dual-domain network (DuDoNet)[33] 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. There have also been recent efforts in MAR using unsupervised approaches, for instance the artifact disentanglement network (ADN) model.[36] 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.

Neural network based approaches have been employed to improve *sparseness artifacts* originating from low-dose CT reconstruction.[40–43] A new method called AirNet[44,45] 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]

Finally, the compensation of *motion artifacts* using deep learning so far has received comparatively less attention. An initial study[48] demonstrated a U-net-based artifact reduction method in the volume domain. A U-net-based neural network was employed[49] to compensate simulated motion artifacts in head CT scans, based on simple simulated rigid (translations, rotations, oscillations) transformations. Motion artifacts in cine cardiac MRI were reduced[50] using recurrent neural networks, and cardiovascular motion in short-scan CT was addressed by means of a deep partial angle-based motion compensation (Deep PAMoCo) framework.[51] Specifically for CBCT and including the 4D case, detecting and avoiding slices with considerable motion artifacts has proven to be a promising strategy to reduce the negative effect of motion artifacts.[52] CNNs were used to reduce streak artifacts caused by fewer projections for each breathing phase to reconstruct motion-resolved 4D CBCT scans.[53] This was extended to using deep learning through prior-guided CNNs to alleviate sparseness streaks using the information of the same volume in different breathing phases.[54] The quality of motion resolved 4D CBCT scans can also be improved using dual-encoder convolutional neural networks (DeCNN) to realize an average-image-constrained 4D CBCT reconstruction.[55]

2 | MATERIALS AND METHODS

2.1 | 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, for example, 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[56] 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.[57] 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 – that is, 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 downsampled 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.[58] 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, that is, $\sum_\alpha \mathbf{A}^\top([\mathbf{A}\mathbf{f}_k]_\alpha - \mathbf{p}_\alpha)$ where $\mathbf{p}_\alpha$ and $[\mathbf{A}\mathbf{f}_k]_\alpha$ 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. Furthermore, *ordered subsets* (OS)[59] and the Nesterov *momentum method*[60] are employed. Finally, a destination-driven approach[61] is employed in forward projection (only for ART) and backward projection (both ART and FDK). The method has been discussed[62] as *TV-regularized OS-SART with momentum* as part of the iCBCT algorithm deployed clinically in Varian products.

2.2 | Motion simulation

To train our models, we use a respiratory motion simulation[63] that generates synthetic sets of CBCT volumes with motion artifacts. It uses phase gated 4D CT scans described in Section 2.3 and an independently recorded set of breathing curves. We use *DEEDS*[64] 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 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 using two different methods: In the first method, called "average volume", we take the average of all deformed volumes which result from application of the DVF scaled with the breathing signal amplitude matching the acquisition time offset at each angular step. The second method, called "average amplitude", corresponds to a single deformed volume at a fixed time offset matching the average amplitude of the breathing signal. While the motion-averaged ground truth is used more often in previous research[62] on motion compensation, images based on this method tend to smoothen sharp edges. On the other hand, images based on the average-amplitude method are able to retain more sharp details. The two methods will be compared in the quantitative as well as clinical evaluation of our approach.

Data augmentation is implemented by: (*i*) applying different breathing curves to the scan, (*ii*) changing the overall motion amplitudes and (*iii*) 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 2.3) side-by-side with the emulated motion artifacts from our motion simulation.

2.3 | 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. The 4D CT scans were provided as input to the motion simulation described in Section 2.2. We simulate a CBCT scan with 720 projections (size 320×76 pixels, resolution 1.344×4.032 mm) using the Halcyon geometry, and a scanning time of 15 s ($24°/s$), representing a simplified Halcyon scan.

To simulate plausible and diverse motion patterns during a virtual CBCT acquisition, we employed a set of 150 recorded free-breathing amplitude signal traces obtained using the Varian Real-time Position Management (RPM) system. The breathing traces used for augmentation are randomly selected from the breathing curves dataset. They contain minor irregularities and baseline drift within the short scanning time.

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 preprocessed projection data and reconstructed volumes were given at the same size, resolution, and geometry to ensure consistency: There are typically between 489 and 697 projections, with one scan having 858 projections, of size 320×76 pixels (resolution 1.344×4.032 mm) for each patient, 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. The acquisition time of the scans varies typically between 16.5 and 24.7 s, with one scan each at 30.8 and 40.5 s, and the test dataset includes both free-breathing and breath-hold scans. The projection count, acquisition time, and variable scan velocity are determined by the Halcyon machine standard acquisition protocols.

2.4 | 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

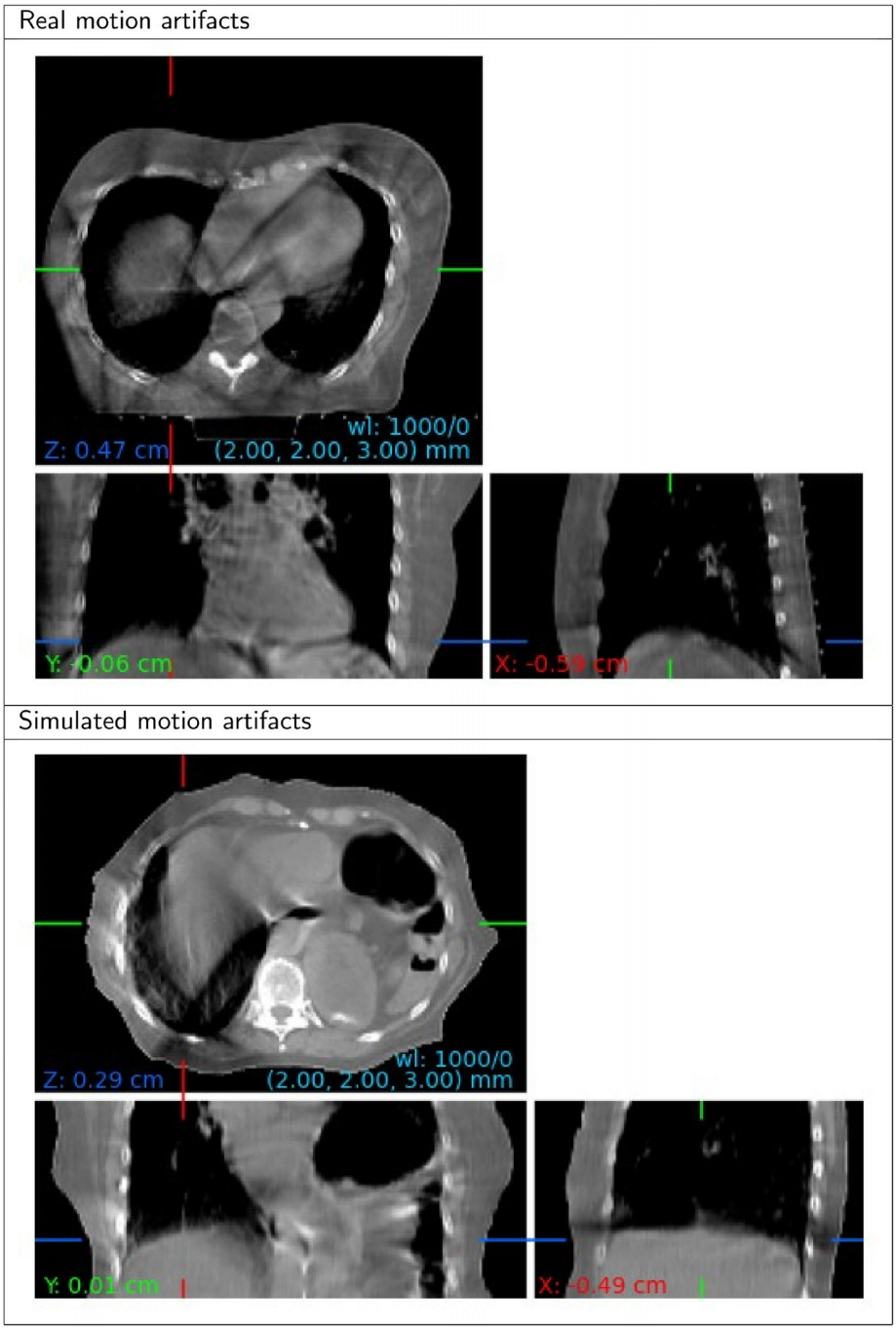

FIGURE 1 Motion artifacts. Top: CBCT image with motion artifacts from the test dataset. Bottom: Image with artificially produced motion artifacts from the motion simulation (images are presented in HU with window and level W/L=1000/0).

be, in principle, applied before and/or after reconstruction. These correction steps are implemented as trainable neural networks derived from 3D encoder–decoder type architectures. The reconstruction algorithm used is either FDK or iterative CBCT (SART-TV) reconstruction, as discussed in Section 2.1. 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 2.2) 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 2.3). 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 acquired projections $\{\mathcal{X}_{proj} \in \mathbb{R}^{H_p \times W_p \times C_p}\}$, and enhances these projections $\{\hat{\mathcal{X}}_{proj}\}$, that is, $f_{pe\_net}(\mathcal{X}_{proj}) \rightarrow \hat{\mathcal{X}}_{proj}$ 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 $\{\mathcal{X}_{vol} \in \mathbb{R}^{H_v \times W_v \times C_v}\}$, that is, $f_{rec}(\hat{\mathcal{X}}_{proj}) \rightarrow \mathcal{X}_{vol} : \mathbb{R}^{H_p \times W_p \times C_p} \rightarrow \mathbb{R}^{H_v \times W_v \times C_v}$, 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 2.1).

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 $\{\hat{\mathcal{X}}_{vol} \in \mathbb{R}^{H_v \times W_v \times C_v}\}$, that is, $f_{ve\_net}(\mathcal{X}_{vol}) \rightarrow \hat{\mathcal{X}}_{vol}$.

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,[65] the Swish activation function[66] 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. 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[67] in 3D. More precisely, given a 3D intermediate feature map $\mathbf{F} \in \mathbb{R}^{C \times D \times H \times W}$, the attention mechanism infers first a channel-wise attention map $\mathbf{M}_c \in \mathbb{R}^{C \times 1 \times 1 \times 1}$ that helps the network to focus on useful channels. Its output is element-wise multiplied with the intermediate feature map $\mathbf{F}$ to re-weight the importance of each channel and generates a new feature map $\mathbf{F}_c \in \mathbb{R}^{C \times D \times H \times W}$. The spatial attention mechanism follows the channel attention mechanism and aids the network to enhance informative spatial local regions by inferring a spatial attention map $\mathbf{M}_s \in \mathbb{R}^{1 \times D \times H \times W}$. Its output is then multiplied with the channel-attention feature map $\mathbf{F}_c$ to re-weight the importance along the spatial domain and generates a new feature map $\mathbf{F}_s \in \mathbb{R}^{C \times D \times H \times W}$. The complete channel-wise and spatial-wise attention mechanism is given by:

$$\mathbf{F}_c = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F}$$

$$\mathbf{F}_s = \mathbf{M}_s(\mathbf{F}_c) \otimes \mathbf{F}_c$$

By using these attention layers, the model is capable of focusing on and learning more relevant features. More precisely, the attention mechanism aids the network on *what* to focus and on *where* to focus. 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.[68] Networks including such residual connections within layers are labeled as "ResUNet" in Table 1.

2.5 | Metrics

In our experiments, we report the numerical performance using several quantitative metrics[69] sensitive to the similarity of pairs of projections or volumes

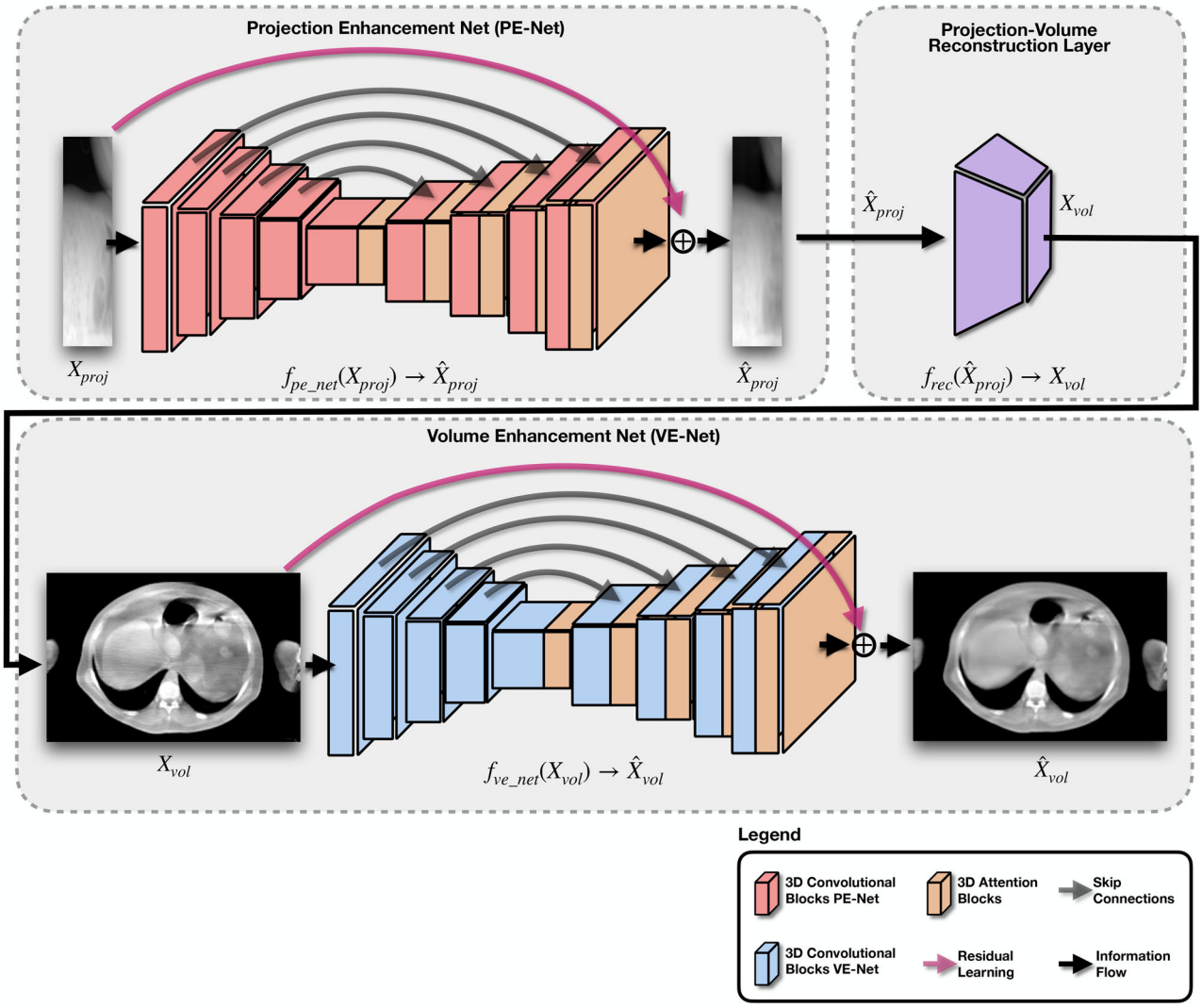

FIGURE 2 Architecture of the proposed end-to-end model, consisting of a projection enhancement network (PE-Net), a projection-to-volume reconstruction layer, and a volume enhancement network (VE-Net).

(x, x'). These include root mean squared error $\text{RMSE} = \sqrt{\text{MSE}}$, where $\text{MSE}(x, x') = \frac{1}{N}\sum_i ||x_i - x'_i||^2$, peak signal-to-noise ratio $\text{PSNR} = 10\log_{10}(\frac{\text{MAX}^2}{\text{MSE}})$, and SSIM).[69] 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.

2.6 | 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 2.3, we use 720 projections of size 320×76 for training, and we add motion artifacts to the original CT volumes using the motion simulation introduced in Section 2.2. 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 2.3).

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

TABLE 1 Quantitative results of deep-learning based motion correction for CBCT data with simulated motion.

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

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 48 200 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 3.2).

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-based 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 and 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[70] 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[71] 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), and trained the models for a total number of 300 epochs. After the training, we select the model that reduces the validation loss the most.

3 | RESULTS

3.1 | 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 2.3). Table 1 presents the numerical performance of the architectures discussed in Section 2 for the two reconstruction methods FDK and SART-TV, with two different sets of ground truth volumes ("average volume" 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 2.5 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, 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 3D 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.

3.2 | 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 2.3 and Figure 5) and evaluated the performance based on the feedback obtained from clinicians. The real-world CBCT scans used in this study are sufficiently different from the simulated training dataset to judge the models' generalization capabilities, for example, 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, that is, 3D-UNet 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 clinical experts determined the level of severity of artifacts besides motion through an additional question asked for each scan. The study

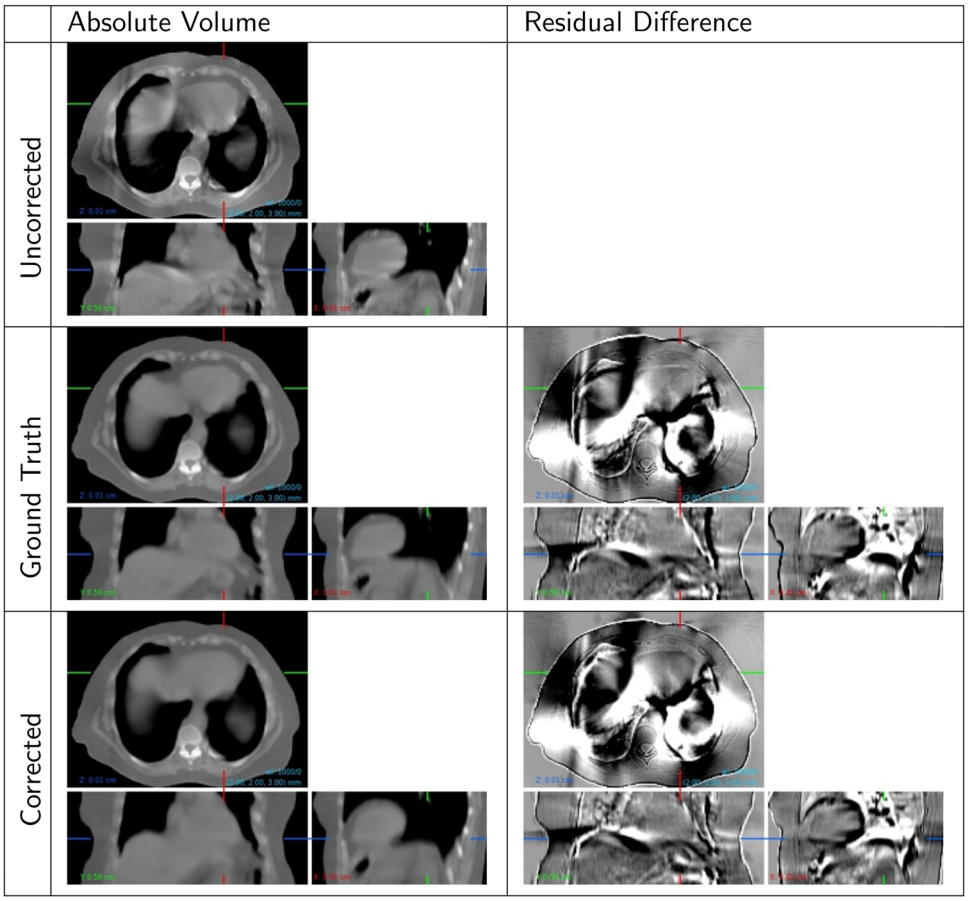

FIGURE 3 Example result for FDK reconstruction (volume domain optimization). Presented are the uncorrected volume using default reconstruction (top), the ground truth volume, both as absolute image and its difference with the uncorrected volume, ("average volume" ground truth, middle), as well as the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

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 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 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. Regarding the choice of ground truth when training the models, the results suggest that "average volume" ground truth based images are preferable for dose calculation due to their time-averaged representation of the mass, while "average amplitude" ground truth based images are preferable for segmentation due to the contrast at organ boundaries.

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

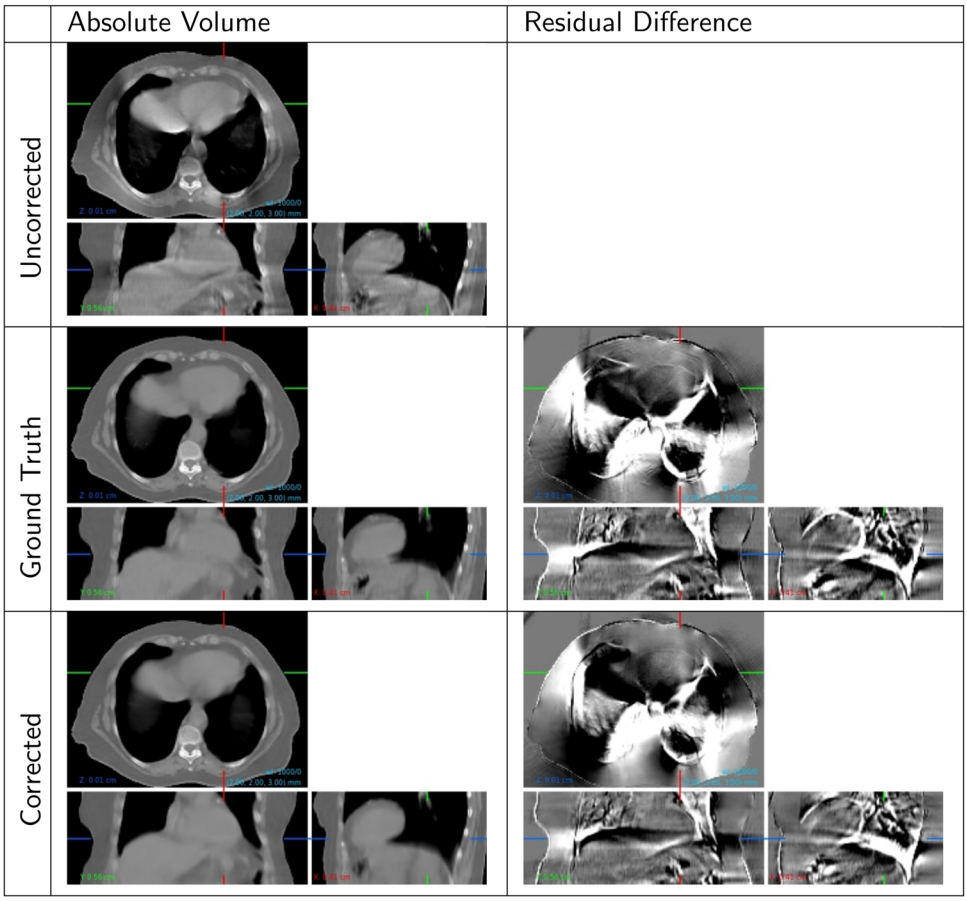

FIGURE 4 Example result for SART-TV reconstruction (volume domain optimization). Presented are the uncorrected volume using default reconstruction (top), the ground truth volume, both as absolute image and its difference with uncorrected volume, ("average volume" ground truth, middle), as well as the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

TABLE 2 Results of the clinical evaluation.

Ground Truth →	Average volume			Average amplitude		
↓ Application/Preference →	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

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 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.

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.

4 | 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,

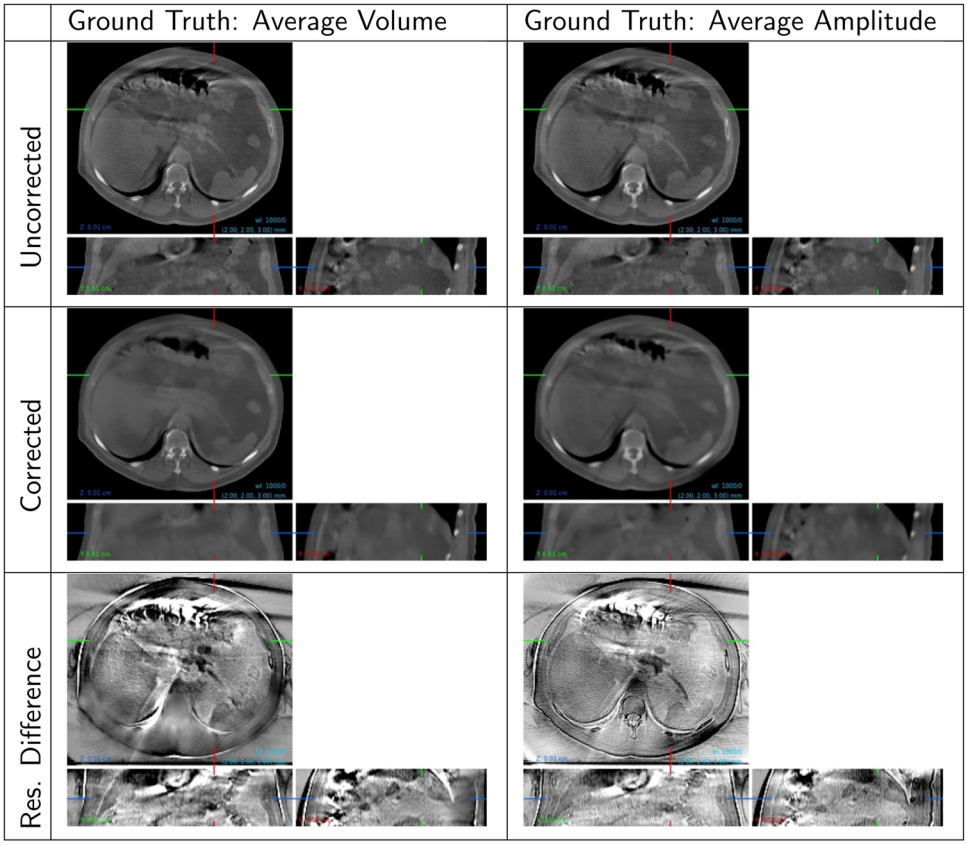

FIGURE 5 Example results for SART-TV reconstruction for real-world test dataset, using the two options for the choice of ground truth, "average volume" (left) and "average amplitude" (right). Presented are the uncorrected volumes using default reconstruction (top), the corrected volumes (middle) as well as the residual corrections (bottom).

the best results were obtained with the volume-domain based correction network, implementing a refined U-net-based 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.

In contrast to time-resolved 4D CBCT acquisition, our proposed solution requires lower compute since it requires processing only a single 3D volume, it does not require additional information, for example, a breathing trace, and finally it can be applied as a pure software upgrade to existing machines.

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.[72] Although we could demonstrate that the trained models are able to generalize to the unseen test set despite different acquisition times, projection counts and breathing patterns compared with the training data set (see Section 2.3), and also helped by the calibration technique we used, 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.

Fourth, when employing 4D reconstruction, guided, for example, by an external breathing signal, the problem of resolving motion can be addressed through explicit prediction of deformations and their application during reconstruction, which is the subject of a follow-up study.

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.

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, Genéve. We thank Giorgia Nicolini and Eugenio Vanetti from RadiQA Services for contributing to the clinical evaluation, for their help organizing this evaluation, and for feedback. 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.

Open access funding provided by Zurcher Hochschule fur Angewandte Wissenschaften.

CONFLICT OF INTEREST STATEMENT

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

REFERENCES

1. Jaffray DA, Siewerdsen JH, Wong JW, Martinez AA. Flat-panel cone-beam computed tomography for image-guided radiation therapy. *Int J Radiat Oncol Biol Phys*. 2002;53:1337-1349.
2. Elstrøm UV, Muren LP, Petersen JBB, Grau C. Evaluation of image quality for different kV cone-beam CT acquisition and reconstruction methods in the head and neck region. *Acta Oncologica*. 2011;50:908-917.
3. Yoon S, Lin H, Alonso-Basanta M, et al. 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*. 2020;12(8):e9660.
4. Jarema T, Aland T. Using the iterative kV CBCT reconstruction on the Varian Halcyon linear accelerator for radiation therapy-planning CT datasets: a feasibility study. *Int J Radiat Oncol Biol Phys*. 2019;68:112-116. Proceedings of the American Society for Radiation Oncology 61st Annual Meeting.
5. Feldkamp LA, Davis LC, Kress JW. Practical cone-beam algorithm. *J Opt Soc Am A*. 1984;1:612-619.
6. Hounsfield GN. Method of and apparatus for examining a body by radiation such as x or gamma radiation. 1975.
7. Grant K, Raupach R. SAFIRE: sinogram affirmed iterative reconstruction. Siemens Healthcare; 2012.
8. Thibault JB. Veo: CT model-based iterative reconstruction. GE Healthcare; 2010.
9. Paysan P, Brehm M, Wang A, Seghers D, Star-Lack J. Iterative image reconstruction in image-guided radiation therapy. 2018. US Patent App. 15/952,996.
10. Gardner SJ, Mao W, Liu C, et al. Improvements in CBCT image quality using a novel iterative reconstruction algorithm: a clinical evaluation. *Adv Radiat Oncol*. 2019;4:390-400.
11. Kim H, Huq MS, Lalonde R, Houser CJ, Beriwal S, Heron DE. Early clinical experience with Varian Halcyon V2 linear accelerator: dual-isocenter IMRT planning and delivery with portal dosimetry for gynecological cancer treatments. *J Appl Clin Med Phys*. 2019;20(11):111-120.
12. Mao W, Liu C, Gardner SJ, et al. Evaluation and clinical application of a commercially available iterative reconstruction algorithm for CBCT-based IGRT. *Technol Cancer Res Treat*. 2019;18.
13. Washio H, Ohira S, Funama Y, et al. Metal artifact reduction using iterative CBCT reconstruction algorithm for head and neck radiation therapy: a phantom and clinical study. *Eur J Radiol*. 2020;132:109293.
14. Schulze RKW, Heil U, Gross D, et al. Artefacts in CBCT: a review. *Dentomaxillofac Radiol*. 2011;40(5):265-273.
15. Dillon O, Keall PJ, Shieh CC, O'Brien RT. Evaluating reconstruction algorithms for respiratory motion guided acquisition. *Phys Med Biol*. 2020;65.
16. Peeters STH, Vaassen F, Hazelaar C, et al. Visually guided inspiration breath-hold facilitated with nasal high flow therapy in locally advanced lung cancer. *Acta Oncol*. 2021;60:567-574.
17. Daly M, McWilliam A, Radhakrishna G, Choudhury A, Eccles CL. Radiotherapy respiratory motion management in hepatobiliary and pancreatic malignancies: a systematic review of patient factors influencing effectiveness of motion reduction with abdominal compression. *Acta Oncol*. 2022;61:833-841.
18. Mohd Amin AT, Mokri SS, Ahmad R, Ismail F, Abd Rahni AA. Evaluation methodology for respiratory signal extraction from clinical cone-beam CT (CBCT) using data-driven methods. *Int J Integr Eng*. 2021;13:1-8.
19. Boas F, Fleischmann D. CT artifacts: causes and reduction techniques. *Imaging Med*. 2012;4.
20. Gjesteby L, De Man B, Jin Y, et al. Metal artifact reduction in CT: where are we after four decades? *IEEE Access*. 2016;4:5826-5849.
21. Paysan P, Roggen T, Zhu L, et al. Deep learning methods for image guidance in radiation therapy. In: Schilling FP, Stadelmann T, eds. Artificial Neural Networks in Pattern Recognition - 9th IAPR TC3 Workshop, ANNPR 2020, Winterthur, Switzerland, September 2–4, 2020, Proceedings, *Lecture Notes in Computer Science*. Springer; 2020;12294:3-22.
22. Hoffmann M, Christlein V, Breininger K. Deep learning computed tomography: learning projection-domain weights from image domain in limited angle problems. *IEEE Trans Med Imaging*. 2018;37:1454-1463.
23. Maier A, Syben C, Stimpel B. Learning with known operators reduces maximum error bounds. *Nat Mach Intell*. 2019;1:373-380.
24. Wang J, Liang J, Cheng J, Guo Y, Zeng L. Deep learning based image reconstruction algorithm for limited-angle translational computed tomography. *PLoS ONE*. 2020;15.
25. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. *Medical Image Computing*

and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing; 2015;234-241.
26. Andersen A, Kak A. Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. *Ultrason Imaging*. 1984;6:81-94.
27. Schnurr AK, Chung K, Russ T, Schad LR, Zöllner FG. Simulation-based deep artifact correction with convolutional neural networks for limited angle artifacts. *Zeitschrift fur Medizinische Physik*. 2019;29:150-161.
28. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw*. 2015;61;85-117.
29. Stadelmann T, Tolkachev V, Sick B, Stampfli J, Dürr O. Beyond ImageNet: deep learning in industrial practice. In: *Applied Data Science*. Springer; 2019:205-232.
30. Amirian M, Montoya-Zegarra JA, Gruss J, et al. 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)*. IEEE;2021:1-7.
31. Park HS, Lee SM, Kim HP, Seo JK, Chung YE. CT sinogram-consistency learning for metal-induced beam hardening correction. *Med Phys*. 2018;45:5376-5384.
32. Zhang Y, Yu H. Convolutional neural network based metal artifact reduction in X-ray computed tomography. *IEEE Trans Med Imaging*. 2018;37:1370-1381.
33. Lin WA, Liao H, Peng C, et al. 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. Yan K, Wang X, Lu L, et al. 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)*. 2018:9261-9270.
35. Lyu Y, Lin WA, Liao H, Lu J, Zhou SK. Encoding metal mask projection for metal artifact reduction in computed tomography. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2020:147-157.
36. Liao H, Lin WA, Zhou SK, Luo J. ADN: artifact disentanglement network for unsupervised metal artifact reduction. *IEEE Trans Med Imaging*. 2020;39:634-643.
37. Lyu Y, Fu J, Peng C, Zhou SK. U-DuDoNet: unpaired dual-domain network for CT metal artifact reduction. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng Y, Essert C, eds. *Medical Image Computing and Computer Assisted Intervention*. 2021:296-306.
38. Wang H, Li Y, Zhang H, et al. InDuDoNet: an interpretable dual domain network for CT metal artifact reduction. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2021:107-118.
39. Wang T, Lu Z, Yang Z. IDOL-Net: an interactive dual-domain parallel network for CT metal artifact reduction. *IEEE Trans Radiat Plasma Med Sci*. 2022;6:874-885.
40. Han Y, Yoo JJ, Ye JC. Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis. *CoRR*. 2016;abs/1611.06391.
41. Jin KH, McCann MT, Froustey E, Unser M. Deep convolutional neural network for inverse problems in imaging. *IEEE Trans Image Process*. 2017;26:4509-4522.
42. Zhang Z, Liang X, Dong X, Xie Y, Cao G. A sparse-view CT reconstruction method based on combination of DenseNet and deconvolution. *IEEE Trans Med Imaging*. 2018;37:1407-1417.
43. Kofler A, Haltmeier M, Kolbitsch C, Kachelrieß M, Dewey M. 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*, Springer; 2018:91-99.
44. Chen G, Hong X, Ding Q, et al. AirNet: fused analytical and iterative reconstruction with deep neural network regularization for sparse-data CT. *Med Phys*. 2020;47:2916-2930.
45. Chen G, Zhao Y, Huang Q, Gao H. 4D-AirNet: a temporally-resolved CBCT slice reconstruction method synergizing analytical and iterative method with deep learning. *Phys Med Biol*. 2020;65:175020.
46. Maier J, Sawall S, Kachelrieß M. 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*. 2018;10573.
47. Erath J, Vöth T, Maier J, Kachelrieß M. Forward and cross-scatter estimation in dual source CT using the deep scatter estimation (DSE). In: *Medical Imaging 2019: Physics of Medical Imaging*. vol 10948. International Society for Optics and Photonics, 2019: 24.
48. Paysan P, Strzelecki A, Arrate F. Convolutional network based motion artifact reduction in cone-beam CT. In: *AAPM Annual Meeting 2019, e-Poster*. 2019.
49. Su B, Wen Y, Liu Y, Liao S, Fu J, Quan G, Li Z. A deep learning method for eliminating head motion artifacts in computed tomography. *Med Phys*. 2022;49:411-419.
50. Lyu Q, Shan H, Xie Y, et al. Cine cardiac MRI motion artifact reduction using a recurrent neural network. *IEEE Trans Med Imaging*. 2021;40:2170-2181.
51. Maier J, Lebedev S, Erath J, et al. Deep learning-based coronary artery motion estimation and compensation for short-scan cardiac CT. *Med Phys*. 2021;48:3559-3571.
52. Hansch A, Dicken V, Klein J, Morgasb T, Haas B, Hahn H. Artifact-driven sampling schemes for robust female pelvis CBCT segmentation using deep learning. In: Mori K, Hahn H, eds. *Medical Imaging 2019: Computer-Aided Diagnosis*, vol 10950; 2019.
53. Zhang Z, Liu J, Yang D, Kamilov US, Hugo GD. Deep learning-based motion compensation for four-dimensional cone-beam computed tomography (4D-CBCT) reconstruction. *Med Phys*. 2022.
54. Zhi S, Duan J, Cai J, Mou X. Artifacts reduction method for phase-resolved Cone-Beam CT (CBCT) images via a prior-guided CNN. In: Schmidt TG, Chen GH, Bosmans H, eds. *Medical Imaging 2019: Physics of Medical Imaging*. vol 10948. International Society for Optics and Photonics, SPIE; 2019:1094828.
55. Jiang Z, Zhang Z, Chang Y, Ge Y, Yin FF, Ren L. Enhancement of 4-D cone-beam computed tomography (4D-CBCT) using a dual-encoder convolutional neural network (DeCNN). *IEEE Trans Radiat Plasma Med Sci*. 2022;6:222-230.
56. Tuy HK. An inversion formula for cone-beam reconstruction. *SIAM J Appl Math*. 1983;43:546-552.
57. Buzug TM. *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. Springer; 2008.
58. Karczmarz S. Angenäherte Auflösung von Systemen linearer Gleichungen. *Bull Int Acad Pol Sic Let, Cl Sci Math Nat*. 1937;355-357.
59. Kim D, Ramani S, Fessler JA. Combining ordered subsets and momentum for accelerated X-ray CT image reconstruction. *IEEE Trans Med Imaging*. 2015;34:167-178.
60. Nesterov Y. Smooth minimization of non-smooth functions. *Math Program*. 2005;103;127-152.
61. Keck B, Hofmann HG, Scherl H, Kowarschik M, Hornegger J. High resolution iterative CT reconstruction using graphics hardware. In: *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*. 2009:4035-4040.
62. Peterlik I, Strzelecki A, Lehmann M, et al. Reducing residual-motion artifacts in iterative 3D CBCT reconstruction in image-guided radiation therapy. *Med Phys*. 2021;48:6497-6507.
63. Paysan P, Munro P, Scheib S. CT based simulation framework for motion artifact and ground truth generation of cone-beam CT. In: *AAPM Annual Meeting 2019, e-Poster*. 2019.

64. Heinrich MP, Jenkinson M, Brady SM, Schnabel JA. MRF-Based Deformable Registration and Ventilation Estimation of Lung CT. *IEEE Trans Med Imaging*. 2013;32:1239-1248.
65. Ulyanov D, Vedaldi A, Lempitsky VS. Instance normalization: the missing ingredient for fast stylization. *CoRR*. 2016;abs/1607.08022.
66. Ramachandran P, Zoph B, Le QV. Searching for activation functions. *CoRR*. 2017;abs/1710.05941.
67. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018:3-19.
68. Zhang Z, Liu Q, Wang Y. Road extraction by deep residual U-Net. *CoRR*. 2017;abs/1711.10684.
69. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13:600-612.
70. Paszke A, Gross S, Massa F. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc; 2019:8024-8035.
71. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International Conference on Learning Representations, ICLR 2019, New Orleans, United States, May 6-9, 2019*. 2019:1-18.
72. Sager P, Salzmann S, Burn F, Stadelmann T. Unsupervised domain adaptation for vertebrae detection and identification in 3D CT volumes using a domain sanity loss. *J Imaging*. 2022;8:222.

How to cite this article: Amirian M, Montoya-Zegarra JA, Herzig I, et al. Mitigation of motion-induced artifacts in cone beam computed tomography using deep convolutional neural networks. *Med Phys*. 2023;50:6228–6242. https://doi.org/10.1002/mp.16405