

# Dual-domain U-Nets with embedded back projection operators for motion-resolved 4D CBCT reconstruction

Ivo Herzig<sup>1</sup>, Pascal Paysan<sup>3</sup>, Daniel Barco<sup>2</sup>, Marc André Stadelmann<sup>2</sup>, Frank-Peter Schilling<sup>2</sup>, Igor Peterlik<sup>3</sup>, Michal Walczak<sup>3</sup>, Lijin Aryananda<sup>1</sup>, Woo Sang Ahn<sup>4</sup>, Rudolf Marcel Fuchsli<sup>1</sup>, Lukas Lichtensteiger<sup>1</sup>

<sup>1</sup> Zurich University of Applied Sciences ZHAW, Institute for Applied Mathematics and Physics IAMP, Winterthur, Switzerland

<sup>2</sup> Zurich University of Applied Sciences ZHAW, Centre for Artificial Intelligence CAI, Winterthur, Switzerland

<sup>3</sup> Varian Medical Systems Imaging Laboratory GmbH, Baden-Dättwil, Switzerland

<sup>4</sup> Radiation Oncology, Gangneung Asan Hospital, University of Ulsan College of Medicine, Republic of Korea

## Abstract

**Background:** Four-dimensional cone beam computed tomography (4D CBCT) plays an important role in image-guided radiation therapy (IGRT) of thoracic cancers. However, its clinical utility is hindered by long scan times, resulting in high patient dose and compromised image quality such as streaking artifacts caused by respiratory motion and sparse sampling of projections.

**Purpose:** We propose a novel deep learning-based method to enable motion-resolved 4D CBCT reconstruction from conventional clinical free-breathing scans without the need for a respiratory signal or explicit projection binning. The proposed deep convolutional neural network (CNN) takes conventional free breathing 3D CBCT projection data as input and predicts a static volume representing the maximum inhalation phase and a set of ten displacement vector fields (DVF) representing patient motion during a complete breathing cycle.

**Methods:** The network extends the well-known U-Net architecture. The downsampling encoder path acts directly on filtered CBCT projection stacks, whereas the upsampling-decoder part acts in the volume domain. The skip connections found in conventional U-Nets are replaced with non-trainable back projection functions at multiple levels of resolution to effectively transfer features from the projection to the volume domain. The model is trained in a supervised fashion on simulated CBCT scans. The method is evaluated on simulated scans of 11 held-out unseen patients from the same simulation dataset, and on 13 clinical free-breathing CBCT scans. Two additional models with different acquisition settings (scan durations of 60s and 6s) were trained and evaluated using three and two scans, respectively, in a clinical expert evaluation. The experts were presented with side-by-side views of a single phase from our 4D reconstruction and a reference 3D SART-TV image and indicated their preference with regard to tumor and esophagus visibility.

**Results:** Clinical experts attested improved tumor visibility (59%, vs. 36% no preference, 5% preferred a reference 3D reconstruction), and esophagus (47%, vs. 42% and 11%). Using the simulated test dataset, we observed stable image quality metrics compared to a traditional SART-TV 3D reconstruction (mean RMSE: -1.19 HU, PSNR: +0.09 dB, SSIM: -0.009) while enabling motion-resolved 4D reconstruction. For clinical CBCT scans, where ground-truth is unavailable, our method demonstrates improved resolution of dynamic structures (e.g., diaphragm) while greatly mitigating motion streak artifacts when compared to traditional reconstructions.

**Conclusions:** Our proposed novel, non-patient-specific CNN model enables the prediction of static volumes and complete 4D respiratory motion models from a conventional clinical free-breathing scan without the need for a recorded respiratory surrogate signal or explicit projection binning. Our study serves as a proof-of-concept for this method, showcasing reduced motion artifacts in images while adding motion modeling capability.

## I. Introduction

Four-dimensional cone beam computed tomography (4D CBCT) has great potential in Image-Guided Radiation Therapy (IGRT) for thoracic cancers, enabling motion-resolved tumor localization and adaptive treatment planning.

However, the required long acquisition times can result in higher patient radiation dose, and 4D CBCT image acquisition protocols typically require additional hardware and setup time to allow recording of respiratory surrogate signals, further reducing clinical utility<sup>1</sup>.

Furthermore, reconstruction algorithms, such as the Feldkamp-Davis-Kress (FDK) method<sup>2</sup> or algebraic reconstruction techniques (ART)<sup>3</sup> are sensitive to motion corruption and incomplete data, leading to degraded image quality.

We introduce a novel deep learning (DL) approach for motion-resolved 4D CBCT reconstruction from standard clinical free-breathing scans, eliminating the need for a recorded respiratory signal or explicit projection binning<sup>4</sup>. A 4D CBCT method, which does not require respiratory signals or explicit binning, can improve robustness, efficiency, and image quality by simplifying workflow and reducing errors from surrogate signals.

### I.A. Related Work

Numerous approaches have been proposed to address motion artifacts and enable time-resolved (4D) CBCT reconstruction, including post-processing of reconstructed images, deep learning-based deformable image registration, projection-binned reconstructions, dual-domain methods, and reconstruction strategies that avoid explicit projection binning<sup>5,6</sup>.

Post-processing methods operate on reconstructed 3D or 4D volumes and have been used to suppress various types of image defects. We focus on artifacts introduced by motion or sparse sampling. Typically, these methods work by learning mappings from artifact-corrupted to artifact-free images<sup>7</sup>. Types of architecture include CNNs<sup>8</sup> and generative methods such as generative adversarial networks (GANs) and diffusion models<sup>9,10,11</sup>. While many 3D artifact-reduction methods can also be ap-

plied to 4D CBCT, some studies specifically focus on the special properties of 4D scans<sup>12,13</sup>.

Motion compensated (MoCo) approaches start with an initial (usually phase correlated) 4D reconstruction and then use motion modeling to combine the 4D volumes to a new set of volumes with reduced sparse-sampling artifacts<sup>14,15,16,17</sup>. Deformable image registration (DIR) is often used for motion modeling. One widely adopted framework is VoxelMorph, which employs a U-Net architecture to perform DIR<sup>18</sup>. Extensions to groupwise registration and settings with sparse or artifact-laden data have also been studied<sup>19,20</sup>.

Most 4D techniques, including MoCo, require the binning of projections according to motion state (most often respiratory phases). This requires a respiratory surrogate signal which is recorded during the scan using, e.g., an optical marker block (such as the Varian Real-Time Position Management® (RPM) system), belts, or other optical methods. However, the need for such external equipment increases scan setup complexity and duration. Alternatively, surrogate signals can also be extracted after the scan from projection data, for example using the well-known Amsterdam Shroud<sup>21</sup> or intensity analysis method<sup>22</sup>. For these methods, however, the accuracy of the signal depends on the quality and properties of the projection data, and usually only works well for a limited range of scan angles.

Several methods have been proposed to combine operations in both the projection and volume domains<sup>23</sup>. One such technique incorporates differentiable non-trainable projection operators directly into neural networks<sup>24</sup>. Other dual-domain strategies have jointly optimized projection interpolation and image-domain enhancement in a unified network<sup>25</sup> or integrate deep learning components into iterative reconstruction pipelines<sup>26,27</sup>.

Methods that do not require a surrogate signal are attractive in clinical settings, where such signals are often unavailable. Recent work has explored reconstruction of dynamic 4D CBCT techniques that aim to reconstruct volumes at each projection, bypassing the need for explicit projection binning<sup>28,29</sup>. Patient-specific learning-based models have been developed to reconstruct images from single projections by estimating deformation models that map a refer-

ence image to the observed motion state using explicit motion fields<sup>30</sup> or an implicit neural fields based approach<sup>31</sup>.

### I.B. Contributions

We propose a deep learning-based method to enable motion-resolved 4D CBCT reconstruction from conventional clinical free-breathing scans without the need for a recorded respiratory surrogate signal and without any explicit projection binning. The neural network is a novel style of dual-domain U-Net with non-trainable back projection operations within its skip connections to allow end-to-end joint motion modeling and 4D image reconstruction directly from projections. The model is not patient-specific and can be trained and evaluated on a mid-tier GPU.

## II. Materials and Methods

### II.A. Synthetic data generation

Our models are trained on synthetic free-breathing CBCT scans, generated using respiratory motion simulation based on a previously published method<sup>8</sup>.

DIR is performed across the ten phases of four-dimensional computed tomography (4D CT) using a registration neural network<sup>18,20</sup>. The network registers each phase to the next in a cyclic manner, wrapping from the final phase back to the initial phase, resulting in a series of patient-specific deformation vector fields (DVF).

Each synthetic scan is generated from a set of DVFs and the corresponding maximum-inhale reference volume combined with a randomly selected respiratory amplitude signal. For each respiratory cycle within this signal, the amplitude is converted into a phase signal and a per-cycle amplitude ratio. The amplitude ratio measures how much a given cycle's inhale-to-exhale amplitude compares to the largest such amplitude in the scan.

During simulated X-ray acquisition, the DVF corresponding to the current phase of the respiratory signal is selected and scaled by the per-cycle amplitude ratio. The reference volume is deformed using this scaled DVF and then forward-projected to generate digitally reconstructed radiographs (DRRs). The

result is a set of synthetic CBCT projections that contain realistic respiratory motion patterns with inter-phase variation in amplitude.

Additionally, motion artifact-free ground-truth volumes are generated by deforming the reference volume using the DVFs at ten discrete phases (ranging from 0% to 90%). These ground-truth volumes correspond to the breathing cycle with the highest inhale-to-exhale amplitude in the scan.

### II.B. U-Net with embedded back projection operator

The proposed neural network takes a set of free-breathing CBCT projections as input and predicts a static volume (max-inhale) and a set of circular DVFs that model a complete respiratory cycle (Figure 1). The DVFs are applied to the static volume to create ten breathing phases, resulting in a 4D reconstruction consistent in format with traditional phase-correlated 4D CBCT methods. The network architecture is based on the original U-Net design, adapted to process projection data and output volumetric motion representations. We use six hierarchical levels.

**Projection pre-processing:** Following the traditional FDK reconstruction approach<sup>2</sup>, the scatter-corrected input projections are converted to attenuation space and filtered using a Shepp-Logan filter. The resulting projection stack is standardized to zero mean and unit variance before being fed into the network.

**Projection domain encoder:** The projections are arranged into a 3D tensor (sinogram) of dimensions  $N \times H \times W$ , where  $N$  is the number of projections (temporal dimension), and  $H, W$  are the spatial projection height and width, respectively. The U-Net encoder uses strided 3D convolutions for hierarchical feature extraction. A stride of 2 is applied along the spatial dimensions at each level to progressively reduce resolution. Along the temporal axis, the first two encoding levels use a stride of 2 to expand the temporal receptive field, while the remaining four levels use a stride of 1 to mitigate sparse-sampling artifacts in the embedded back projection operators. The number of filters at each level is 32, 64, 64, 64, 64, and 64, respectively.

**Projection-to-volume skip connections:** At each

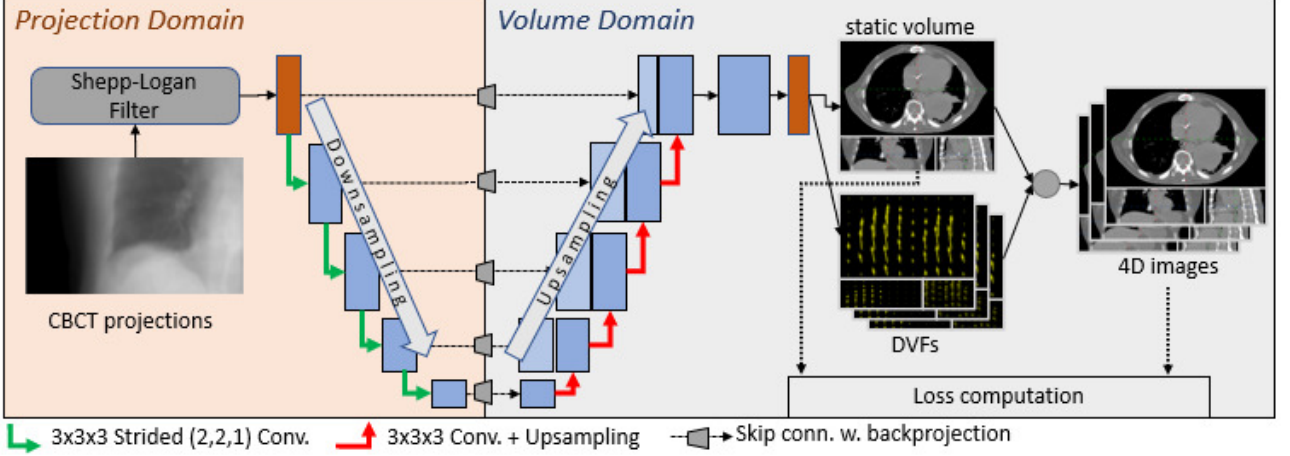


Figure 1: Overview of the CNN-based reconstruction method. Starting with CBCT projections from a scan and finally predicting a static volume and a set of cyclic DVFs representing patient motion during a breathing period.

level of the U-Net, the projection domain feature maps are translated to the volume domain via embedded non-trainable back projection functions<sup>24</sup>. Instance normalization is applied to the backprojected feature maps and they are then concatenated with the corresponding features in the volume-domain decoder path. The back projection functions are implemented as custom CUDA kernels. When combined with the Shepp-Logan filter applied to the input projections, this resembles the FDK reconstruction method. To allow end-to-end training of the network, the backward pass of these layers is implemented as a forward projection operation.

**Volume domain decoder:** Each level applies a block consisting of  $3 \times 3 \times 3$  convolutions followed by Swish activation<sup>32</sup> and 2x trilinear upsampling. In addition to the upsampling layers, we add three additional convolution layers without upsampling on the top level of the network to refine the resulting volumetric outputs at the final output resolution. All convolutional layers in the decoder use 64 filters, making the down- and upsampling paths of the U-Net symmetric except for the top level.

**Network outputs and volume warping:** The last layer of the network is a convolutional layer without an activation function; it maps the feature maps to the output volume with dimensions  $31 \times D \times H \times W$ . This output volume is split into a static image

( $1 \times D \times H \times W$ ) and a set of 10 DVFs ( $10 \times D \times H \times W \times 3$ ), where  $D, H, W$  are the volume sizes in  $z, y, x$  directions, respectively. Each DVF maps the volume from one breathing phase to the next, thus representing a periodic breathing cycle. These DVFs are used to warp the static volume and generate a 10-phase 4D image sequence.

### II.B.1. Training

The model consists of just 2.2 million parameters. All models were trained using a batch size of 1, utilizing approximately 30 GB of the 48 GB available on a single NVIDIA RTX A6000. Training for 300 epochs takes approximately 100 hours. We track validation performance on a held-out portion of the dataset and select the model with the best performance. We use the Adam optimizer<sup>33</sup> with a learning rate of  $10^{-5}$ . The training uses MAE loss on the predicted static image and the moving 4D images with equal weights, and bending energy<sup>34</sup> regularization (weight: 0.001) on the predicted DVFs. We performed basic manual hyperparameter tuning on the 30 s synthetic dataset and used the same settings for training the 60 s and 6 s scan duration models.

## II.C. Evaluation

Model performance is assessed on a synthetic test dataset and real clinical CBCT scans.

For the synthetic test dataset, where ground-truth images are available, we report Root Mean Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) for the predicted volumes and compare them to traditional 3D (SART-TV) reconstructions. These metrics are computed on the body-only part of the image; background air and couches are removed prior to computing all metrics. Background removal is performed by thresholding the image based on HU and selecting the second-largest connected component in the image as the body (the largest component being the background).

To assess suitability for downstream clinical tasks and motion modeling performance, we use TotalSegmentator<sup>35</sup> to perform segmentation on both the ground-truth and reconstructed volumes. TotalSegmentator produces segmentation masks for over 100 structures; we focus on the relevant structures right lung, left lung, liver, heart, esophagus, and spinal cord, and report Dice similarity coefficients (referred to as Dice scores) for these segments.

Robustness and adaptability of our proposed method were evaluated by training two additional models on a 60s and 6s free-breathing setup. Although a short 6s CBCT naturally exhibits fewer motion artifacts, it presents a significant challenge for motion-resolved 4D reconstruction due to the very limited number of projections per respiratory phase.

Predictions from these models were used to conduct a clinical expert evaluation. The experts were presented a side-by-side view of a SART-TV 3D reconstruction and the exhale phase image from our 4D reconstruction. They were asked to assess tumor and esophagus visibility and indicated their preferred reconstruction.

For the clinical scan test datasets, there is no ground-truth available. We provide images for qualitative assessment, comparing our method to 3D iterative algebraic (SART-TV)<sup>36</sup> and 4D McKinnon-Bates (MKB)<sup>37,38</sup> reconstructions.

### II.C.1. Input data consistency

There is no ground-truth available for the real CBCT scans. To assess projection consistency of the predicted motion pattern with the actual motion in the scan, we assess the diaphragm motion in projection space by forward projecting our 4D reconstruction and comparing it to the original input projections. The predicted 10-phase volumes are each forward projected with acquisition settings that match the original input data, resulting in a set of 10 sinograms which are used to assess the input data consistency.

**Amsterdam-Shroud method to extract diaphragm motion:** A diaphragm displacement signal is extracted from the original input projection using the Amsterdam-Shroud (AS)<sup>21</sup> method for the right lung. A region of interest (ROI,  $13 \times 13 \times 32$  voxels) containing the diaphragm dome is manually selected for each scan. To extract the actual signal from the AS images, we use a path-finding algorithm<sup>39</sup> to find the minimum-cost path through all columns of the AS image. Due to the presence of minor residual sorting artifacts (ghost edges) from the original 4D-CT scans for the synthetic datasets (Figure 3), and the slightly blurred results for our reconstruction method, we have found this to be a more robust way to extract curves from AS images than the better-known maximum-intensity or optimization-based methods. In particular, the path-finding approach also alleviates issues for scan angles where the left lung diaphragm is also visible in the AS image, which can lead to discontinuities when using other methods. Furthermore, all AS images are manually reviewed and we mark invalid regions due to the ROI leaving the field-of-view, occlusion of the diaphragm, or other artifacts.

**Comparison of extracted diaphragm displacements:** We extract AS images and diaphragm positions for our predicted max-inhale and max-exhale phase volumes. The two extracted curves represent the maximum predicted diaphragm displacement in the superior-inferior direction over all projections. To compare this to the input projections, we apply peak detection to the original AS curves and calculate the average displacement between all lower and all upper peaks. We compute the ratio of the overall displacement and a position offset for the

prediction over the original input projections. The displacement ratio quantifies the predicted motion amplitude compared to the input; a value of  $> 1$  means that the predicted motion is larger than the input data, while values  $< 1$  mean that the motion is underestimated. The results of this method can be seen in Figures 3 and 6.

## II.D. Datasets

### II.D.1. Clinical testing datasets

We evaluated our model on three datasets acquired under different scan protocols.

**30 s Acquisition:** Thirteen thoracic scans from seven radiotherapy patients from Gangneung Asan Hospital, South Korea, were acquired on Varian Halcyon™ systems in half-fan mode (860 projections). Each scan spans 4–8 respiratory cycles. Scans were centered on the planning target volume (PTV), occasionally resulting in truncation artifacts.

*Geometry:* source-imager distance (SID): 154 cm, source-to-axis distance (SAD): 100 cm, detector lateral offset: 17.5 cm.

**60 s Acquisition:** Four 360° full-fan scans (879 projections) from individual patients, each covering 18–22 respiratory cycles. *Geometry:* SID: 154 cm, SAD: 100 cm.

**6 s Acquisition:** Five 360° full-fan scans (407 projections) covering 1–2 respiratory cycles. *Geometry:* SID: 154 cm, SAD: 100 cm.

The 60 s and 6 s scans were acquired using the Varian HyperSight system on a Halcyon machine with an imager size of  $86 \times 43$  cm (pixel size:  $0.28 \times 1.12$  mm).

Due to GPU memory limits, projection images were downsampled from their original size. *30 s:* image width 366 and height between 85 and 147 pixels (pixel size:  $1.76 \times 2.688$  mm); *60 s & 6 s:*  $768 \times 192$  pixels (pixel size:  $1.12 \times 2.24$  mm). The reconstructed volume sizes are: *30 s:*  $256 \times 256 \times 64$  voxels ( $2 \times 2 \times 3$  mm); *60 s & 6 s:*  $256 \times 256 \times 122$  voxels ( $2.1 \times 2.1 \times 2.0$  mm).

### II.D.2. Synthetic Datasets

The model is trained using synthetic data generated using the method from Section II.A.. Syn-

thetic datasets matching each clinical setting (30 s, 60 s, 6 s) were generated using a 4D-CT dataset of 78 patients (split 70%/15%/15% for training/validation/testing). For each patient, 10 synthetic scans were simulated using randomly selected amplitude signals from a set of 150 traces acquired using the Varian RPM system. This resulted in 560 training, 110 validation, and 110 test samples per acquisition type.

## III. Results

### III.A. Simulated dataset

#### III.A.1. Image quality

Table 1 presents the mean RMSE, PSNR, and SSIM performance metrics for our 4D reconstruction method compared to the SART-TV 3D reconstruction on the simulated test datasets. To ensure a conservative comparison, SART-TV 3D metrics were computed against all ten ground-truth phases, and only the best-matching (maximum) value was used. Our 4D results are reported as the average value over all ten phases.

The image quality metrics depend on intrinsic scan properties, such as motion amplitude and tissue distribution. Figure 2 shows the distribution of per-sample *gained* image quality of our method compared to SART-TV. The 60 s and 6 s retrained models exhibit a slight quantitative decrease, likely because initial model development and hyperparameter tuning were focused on the 30 s protocol (see Section II.B.1.). Furthermore, the 6 s scan presents an extra challenge, providing only 1–2 full respiratory periods for motion inference. Overall, these results show that our method can achieve the benefits of 4D motion resolution with quantitative image quality close to a full-projection-count 3D reconstruction.

As illustrated in Figure 3, motion streak artifacts present in the SART-TV reconstruction are reduced, and tissue interfaces of moving structures (diaphragm, tumor) are better defined. However, perceived sharpness and visibility of fine details and small structures are reduced, giving the predicted image a blurred appearance. This blurring effect is a known limitation of networks trained with pixel-wise loss functions such as MAE<sup>40</sup>.

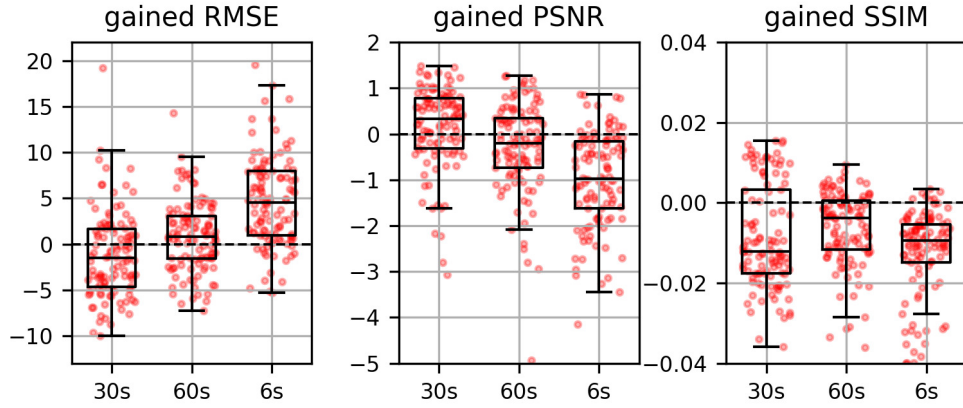


Figure 2: Distribution of per-sample gained image quality for the 30 s, 60 s and 6 s model versions relative to SART-TV. Image quality is quantified in terms of gained RMSE (lower is better), gained PSNR (higher is better), and gained SSIM (lower is better). The boxes span the interquartile range (IQR), and the whiskers extend to the most extreme data points within  $1.5 \cdot \text{IQR}$ . The box plots are overlaid with a scatter plot of all individual samples in the test dataset.

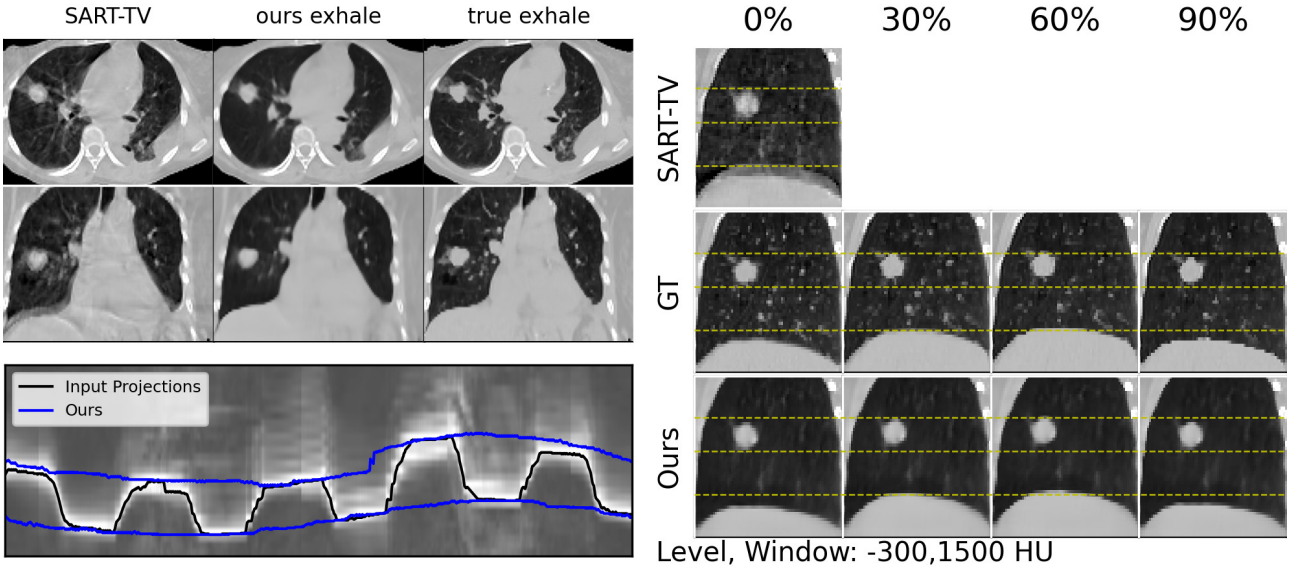


Figure 3: Sample from synthetic test dataset with visible lung tumor. *Top left*: Image quality comparison of our method with SART-TV and ground-truth image, showing reduced motion artifacts. *Right*: 4D reconstruction of our method compared to the ground-truth, and a 3D SART-TV reconstruction. Shown are four out of 10 respiratory phases. *Bottom left*: Extracted diaphragm motion (AS) from the original input projections and in-/exhale phases from our 4D reconstruction. The reconstructed phases are consistent with the in-/exhale positions.

The predicted tumor and diaphragm motion over the breathing cycle matches the positions in the ground-truth images. In the AS image, the diaphragm position of the predicted in-/exhale phases matches the peak positions of the original input projections. However, compared to AS images of real CBCT scans, the diaphragm trace curve is flatter, and there are some smeared-out or ghost edges visible in the image. These effects stem from the synthetic CBCT and motion simulation framework used in this study.

|             | SART-TV     | Ours(Static) | Ours(4D)    |
|-------------|-------------|--------------|-------------|
| <b>30 s</b> |             |              |             |
| RMSE        | 47.64±9.98  | 49.67±8.39   | 46.45±7.13  |
| PSNR        | 32.70±2.16  | 32.22±1.46   | 32.79±1.37  |
| SSIM        | 0.956±0.017 | 0.943±0.015  | 0.947±0.014 |
| <b>60 s</b> |             |              |             |
| RMSE        | 38.29±7.28  | 61.95±7.13   | 39.16±5.41  |
| PSNR        | 34.53±1.77  | 30.24±1.00   | 34.25±1.21  |
| SSIM        | 0.961±0.015 | 0.935±0.016  | 0.954±0.014 |
| <b>6 s</b>  |             |              |             |
| RMSE        | 39.79±7.93  | 84.59±6.88   | 44.39±7.86  |
| PSNR        | 34.20±1.77  | 27.50±0.72   | 33.21±1.57  |
| SSIM        | 0.967±0.012 | 0.911±0.018  | 0.955±0.015 |

Table 1: Image quality metrics for the simulated dataset (RMSE in HU, PSNR in dB).

### III.A.2. Segmentation Results

Mean segmentation Dice scores for segments most affected by respiratory motion are consistently higher compared to SART-TV (right lung: +0.0089, left lung: +0.0077, liver: +0.10, heart: +0.028, esophagus: +0.032) and stable for the stationary spinal cord (+0.0032). Figure 4 shows the distribution of Dice scores for the samples in the test dataset. The Dice scores for SART-TV are the maximum scores against all 10 ground-truth phases, while for our 4D method Dice scores are averaged over all 10 phases.

### III.B. Clinical scan dataset

Ground-truth is not available for the real clinical scans. Figure 6 provides a visual comparison of reconstructed image quality and motion fidelity across respiratory phases. Our method shows improved overall image quality compared to SART-TV and MKB reconstructions. Motion streak artifacts are greatly reduced and moving structures are better defined than in SART-TV reconstructions. In addition, there are no visible sparse sampling artifacts compared to MKB.

Compared to MKB, our method preserves respiratory motion while reducing noise and sparse sampling artifacts. Diaphragm tracking results show that predicted motion is consistent with the input projections (Figure 6). Quantitatively, in terms of predicted diaphragm displacement compared to the input projections, our method predicts diaphragm motion better than MKB, which often underestimates the motion (Figure 5). Relative diaphragm displacement: MKB mean 0.58 (std 0.1), ours mean 0.98 (std 0.1).

### III.C. Retrained models for different acquisition geometries

Our trained models are inherently sensitive to CBCT acquisition parameters such as scan duration and projection count, which affect the spatial and temporal resolution of the acquired data. To evaluate robustness across different CBCT protocols, we retrained our model using data matched to the geometry of HyperSight 60 s and 6 s acquisitions, maintaining the same network architecture and training configuration as used for the 30 s model.

The 30 s model achieved the best image quality in terms of RMSE and PSNR. The 60 s and 6 s models performed slightly worse (mean PSNR differences of  $-0.28$  dB and  $-0.99$  dB, respectively, compared to SART-TV). This may reflect that model development and hyperparameter tuning were performed for the 30 s setting only. No protocol-specific adjustments were made for the 60 s and 6 s models, indicating potential for improvement through targeted optimization.

For the 60 s and 6 s scans, we observe a larger difference in RMSE and PSNR between the predicted static image and the final (warped) 4D reconstruction

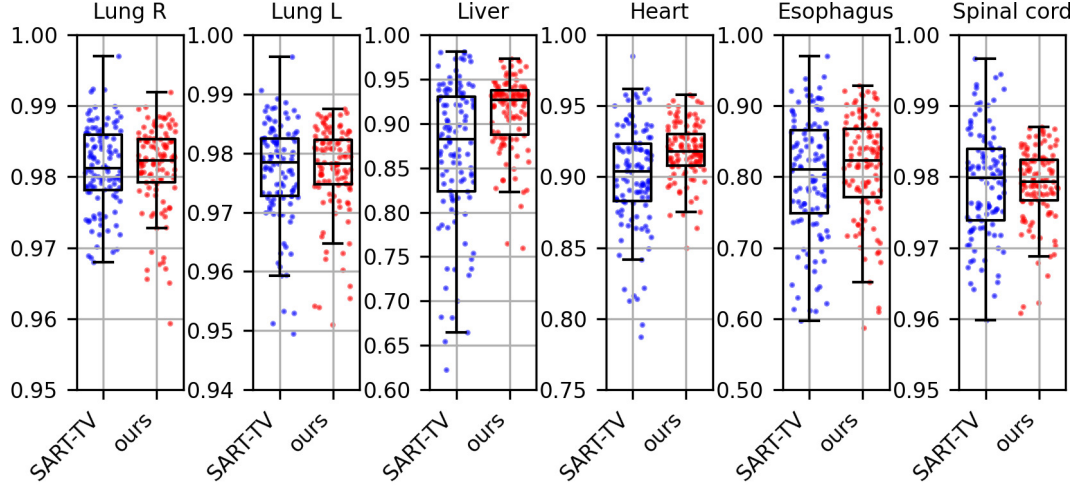


Figure 4: Segmentation Dice scores for SART-TV and our method, showing improved segmentation performance for segments affected by respiratory motion (lungs, liver, heart, esophagus) and consistent segmentation scores for the stationary spinal cord.

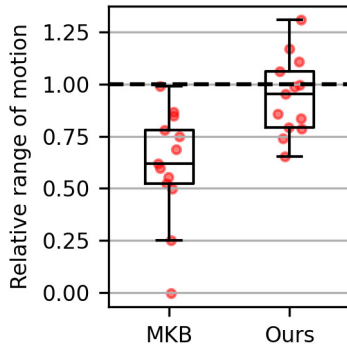


Figure 5: Relative diaphragm displacement for the 30s clinical test dataset for the MKB reconstruction and our proposed CNN method. A value of 1.0 means that the predicted diaphragm displacement exactly matches the displacement in the original X-ray projections; values  $> 1$  or  $< 1$  mean the displacement is over- or underestimated, respectively.

than for the 30 s scan (60 s: +4.01 dB; 6 s: +5.71 dB; 30 s: +0.57 dB), suggesting that the networks not only learn to use the predicted DVFs to represent actual patient motion but also use the warping step to increase image quality. This further suggests that there is room for improvement through hyperparameter tuning and DVF regularization.

Despite the slight decrease in image quality met-

rics, the model still produced 4D volumes with reduced motion artifacts and improved anatomical boundary definition. The extracted diaphragm motion traces showed good agreement with the input projection data, indicating robust motion estimation (Figure 7).

### III.D. Clinical Expert Evaluation

A total of eleven clinical experts (seven radiation oncologists and four medical physicists) participated in the evaluation. Each expert assessed two images from the HyperSight 6 s dataset and three images from the HyperSight 60 s dataset. The experts were shown side-by-side a SART-TV 3D image and a single static image (exhale phase) of our 4D reconstruction. They were asked to select their preferred image for (a) tumor visibility and (b) esophagus visibility.

For tumor visibility, averaged over all five samples and experts, 59% of votes indicated a preference for our reconstruction, 36% indicated no preference, and 5% favored the iterative 3D reconstruction. Similarly, for esophagus visibility, 47% preferred our reconstruction, while 42% had no preference and 11% favored the iterative 3D approach (Figure 8 and Table 2). The high percentage of "no preference" votes suggests that our 4D reconstruction provides comparable structural visibility to the 3D images, while also

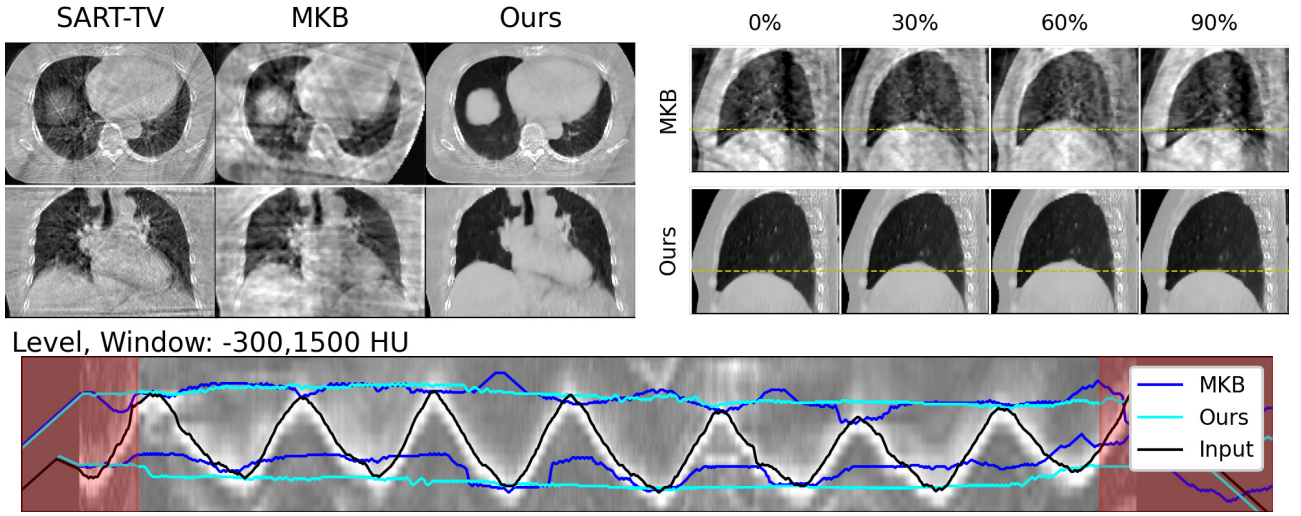


Figure 6: Sample from clinical test dataset. *Top left*: Image quality comparison of our method with 3D SART-TV and 4D MKB reconstructions. Motion streaking and sparse sampling artifacts are reduced by our method. *Top right*: 4D reconstruction of our method compared to MKB. Shown are four out of ten respiratory phases. *Bottom*: Extracted diaphragm motion of the input projections and from the in-/exhale phases from our 4D and a 4D MKB reconstruction. The reconstructed phases match the in-/exhale position of the original input projections for both MKB and our methods.

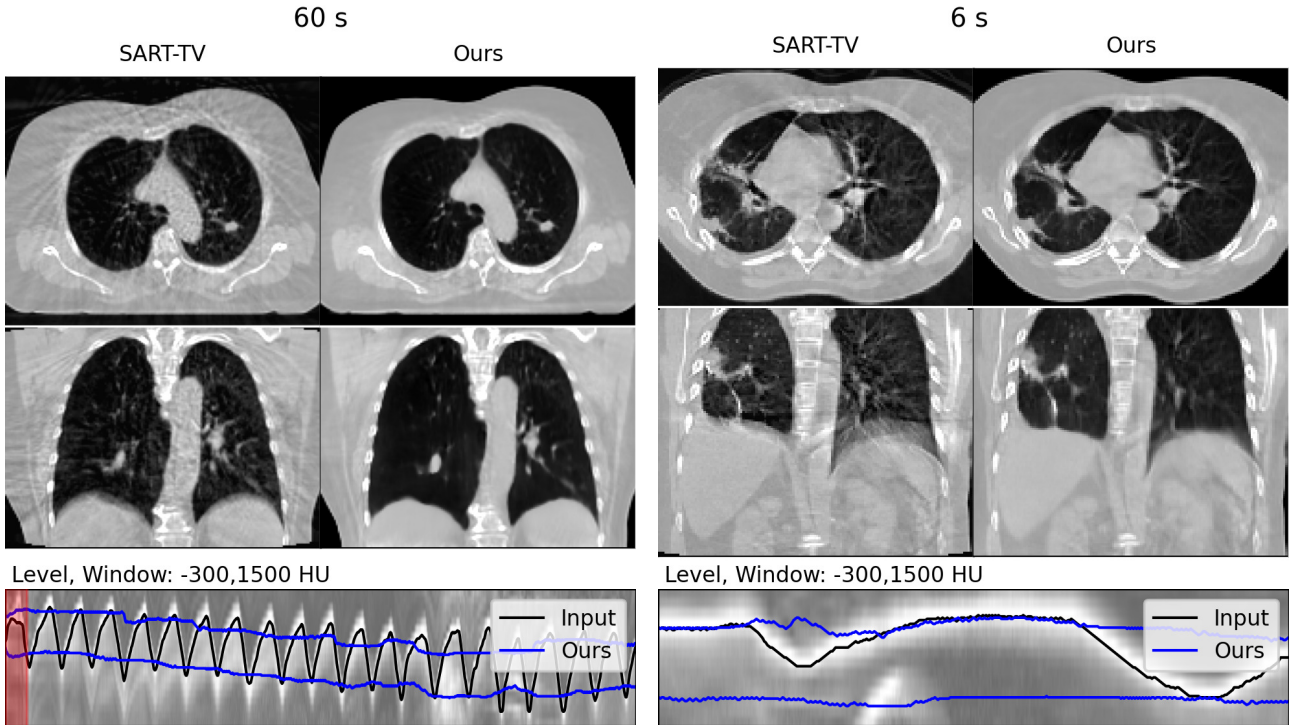


Figure 7: Comparison of SART-TV 3D to our 4D reconstruction for two HyperSight scans with AS images and diaphragm positions (left: 60 s, right: 6 s).

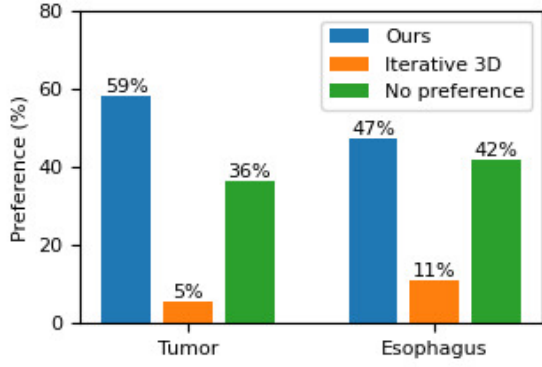


Figure 8: Clinical expert evaluation for the two tasks: tumor visibility and esophagus visibility. The bars show the indicated preference of the clinical experts for our 4D or a SART-TV 3D reconstruction method, averaged over all five scans.

|      | Tumor visibility |          |          | Esophagus visibility |          |          |
|------|------------------|----------|----------|----------------------|----------|----------|
|      | Ours             | Iter. 3D | No pref. | Ours                 | Iter. 3D | No pref. |
| 6 s  | 10               | 1        | 0        | 4                    | 1        | 6        |
| 6 s  | 4                | 0        | 7        | 5                    | 1        | 5        |
| 60 s | 7                | 0        | 4        | 8                    | 1        | 2        |
| 60 s | 5                | 2        | 4        | 4                    | 1        | 6        |
| 60 s | 6                | 0        | 5        | 5                    | 2        | 4        |

Table 2: Results for all five cases from the clinical evaluation. The table lists two 6 s and three 60 s scans and the number of experts (out of a total of 11) who preferred our 4D reconstruction, an iterative 3D reconstruction, or indicated no preference.

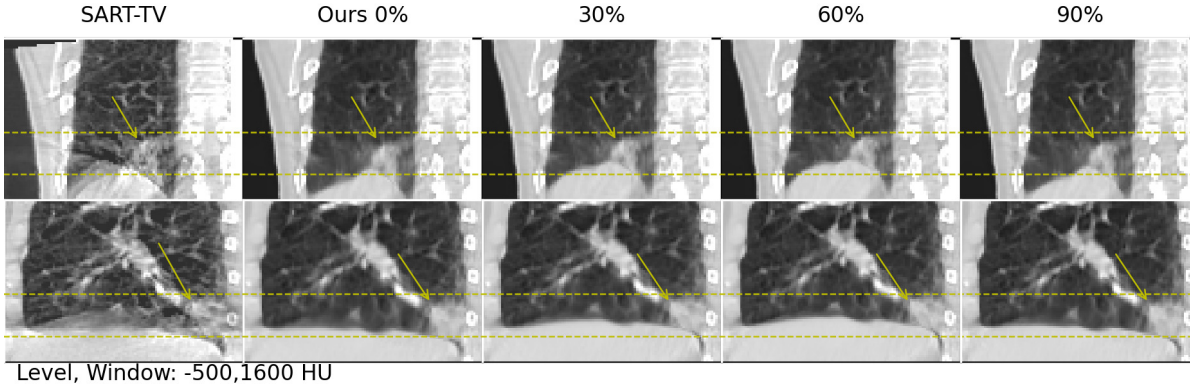


Figure 9: Comparison of SART-TV 3D to our 4D reconstruction of a HyperSight 6 s scan of a case with a tumor in the lower right lung. Improved tumor visibility and delineation of our 4D vs. 3D SART-TV reconstruction. This case was presented in the expert evaluation (first row of Table 2), with 10 out of 11 experts agreeing on improved tumor visibility for our 4D reconstruction.

adding temporal information. Figure 9 shows a scan from a difficult case with a lung tumor in the right lower lung. In the SART-TV 3D reconstruction, tumor visibility is hindered by strong streaking artifacts. In comparison, our proposed 4D method reduces the streaking artifacts and provides a motion-resolved reconstruction, with 10 out of 11 experts agreeing on better tumor visibility.

## IV. Discussion

We proposed a CNN-based 4D CBCT reconstruction method and evaluated its performance across simulated and clinical datasets. On simulated data, our method achieved image quality metrics comparable to conventional 3D reconstruction techniques, while additionally providing motion resolution across respiratory phases. However, our reconstruction results exhibit a blurring effect that can obscure high-frequency details. This is a known limitation of training with pixel-wise loss functions such as Mean Ab-

solute Error (MAE). To improve sharpness, future investigations could explore alternative or combined loss formulations, such as SSIM-based or perceptual losses, to improve image sharpness while preserving robustness to motion artifacts.

In clinical scenarios, where ground-truth is unavailable, we assessed the method through qualitative visual comparison and diaphragm motion analysis using AS images. Our approach substantially reduced sparse sampling artifacts and increased motion fidelity compared to MKB reconstructions. A key advantage of our method is that it does not require external respiratory signals, signal extraction, or any explicit binning of projections. Instead, the network infers temporal phase information directly from the projection data.

We demonstrated adaptability to different acquisition protocols by retraining the model to match HyperSight 6s and 60s scans, suggesting potential for integration into different clinical workflows. Even under high-speed, sparse-view conditions, the method produced motion-resolved volumes that aligned well with expected motion patterns, as indicated by diaphragm trajectory estimates. The results for HyperSight 6s scans are especially relevant, since current motion modeling or 4D reconstruction approaches generally require longer scans covering multiple respiratory periods, in contrast to just 1–2 full periods in the 6s scans. Traditional 4D CBCT requires scan times of 60 or more seconds. Our study shows 4D reconstructions from 60s, 30s, and even 6s scans (a reduction of 50% to 90%). Our method does not require a recorded surrogate signal, which can further reduce the required setup time. Furthermore, standard 4D acquisition protocols often utilize higher mAs settings to compensate for the image quality degradation caused by projection binning. For instance, a Varian Halcyon standard 4D thorax protocol delivers 11.1 mGy compared to 4.45 mGy for a 3D scan ( $CTDI_{vol}$ , 32 cm), a dose reduction of roughly 60%.

A clinical expert evaluation reinforced our quantitative findings. Across five patient cases, the majority of experts preferred our reconstructions over SART-TV in terms of both tumor and esophagus visibility. This consensus was particularly pronounced in the challenging case of a lower-lung tumor, where strong streaking artifacts in the SART-TV reconstruction

obscured the target, while our method yielded much-improved tumor visibility. We recognize that our clinical evaluation is limited by the number of experts and by using only five patient cases; it should therefore be seen as a preliminary validation of our method. Nevertheless, the results of the study suggest that our method successfully enables time-resolved 4D reconstructions without impacting perceived image quality compared to traditional 3D reconstructions.

Direct validation of the predicted motion remains challenging due to the lack of ground-truth. Our approach of comparing the motion present in the original input projections to the predicted diaphragm motion states in the AS images provides a useful proxy to judge motion fidelity. However, additional studies may be necessary for more rigorous motion validation. Our findings reveal that quantitative image quality metrics are consistently lower for the single predicted static volume but improve substantially when DVFs are applied. This observation suggests that the model not only leverages the DVFs for motion modeling but also implicitly enhances image quality through their application.

Image quality assessment is relatively straightforward through established quantitative metrics, whereas motion modeling lacks equally accessible and reliable evaluation criteria. This potential trade-off complicates efforts to fine-tune model hyperparameters.

## V. Conclusion

We introduced a novel CNN-based U-Net architecture for 4D CBCT reconstruction directly from projection data, eliminating the need for recorded respiratory signals or explicit projection binning.

The model is trained on a population dataset, requiring no patient-specific adaptation. We implemented this as a compact network with just over 2 million parameters, enabling 10-phase 4D reconstruction on a single mid-tier GPU in under 4 seconds. This efficient approach offers high-quality, motion-resolved imaging and demonstrates strong potential for integration into clinical workflows by reducing complexity, setup time, and reliance on respiratory surrogates.

## Acknowledgments

Clinical test dataset (30 s scans) supplied by Yonsei University and Gangneung Asan Hospital, University of Ulsan College of Medicine, IRB approval number GNAH IRB 2022-07-008.

HyperSight 60 s and 6 s datasets supplied by Maastricht Clinic and Nova Scotia Health.

We also thank Yonsei University for facilitating the clinical expert evaluation.

This work was supported by Innosuisse under grant 120.480 IP-LS.

## Conflict of Interest Statement

Pascal Paysan, Igor Peterlik, and Michal Walczak are employees of Varian Medical Systems Imaging Laboratory GmbH.

## References

- <sup>1</sup> C. Nardi, R. Molteni, C. Lorini, G. G. Taliani, B. Matteuzzi, E. Mazzoni, and S. Colagrande, Motion artefacts in cone beam CT: an in vitro study about the effects on the images, *The British journal of radiology* **89**, 20150687 (2016).
- <sup>2</sup> L. A. Feldkamp, L. C. Davis, and J. W. Kress, Practical Cone-Beam Algorithm, *Journal of the Optical Society of America A* **1**, 612–619 (1984), DOI:10.1364/JOSAA.1.000612.
- <sup>3</sup> R. Gordon, R. Bender, and G. T. Herman, Algebraic Reconstruction Techniques (ART) for Three-Dimensional Electron Microscopy and X-ray Photography, *Journal of Theoretical Biology* **29**, 471–481 (1970), DOI:10.1016/0022-5193(70)90109-8.
- <sup>4</sup> J.-J. Sonke, L. Zijp, P. Remeijer, and M. van Herk, Respiratory Correlated Cone Beam CT, *Medical Physics* **32**, 1176–1186 (2005), DOI:10.1118/1.1869074.
- <sup>5</sup> Y. Zhang, Z. Jiang, Y. Zhang, and L. Ren, A Review on 4D Cone-Beam CT (4D-CBCT) in Radiation Therapy: Technical Advances and Clinical Applications, **51**, 5164–5180 (2024), DOI:10.1002/mp.17269.
- <sup>6</sup> M. Amirian, D. Barco, I. Herzig, and F.-P. Schilling, Artifact Reduction in 3D and 4D Cone-Beam Computed Tomography Images With Deep Learning: A Review, *IEEE Access* **12**, 10281–10295 (2024), DOI:10.1109/access.2024.3353195.
- <sup>7</sup> S. Kida, T. Nakamoto, M. Nakano, K. Nawa, A. Haga, J. Kotoku, H. Yamashita, and K. Nakagawa, Cone Beam Computed Tomography Image Quality Improvement Using a Deep Convolutional Neural Network, *Cureus* **10**, e2548 (2018), DOI:10.7759/cureus.2548.
- <sup>8</sup> M. Amirian, J. A. Montoya-Zegarra, I. Herzig, P. Eggenberger Hotz, L. Lichtensteiger, M. Morf, A. Züst, P. Paysan, I. Peterlik, S. Scheib, R. M. Fuchsli, T. Stadelmann, and F.-P. Schilling, Mitigation of Motion-Induced Artifacts in Cone Beam Computed Tomography Using Deep Convolutional Neural Networks, *Medical Physics* **50**, 6228–6242 (2023), DOI:10.1002/mp.16405.
- <sup>9</sup> R. L. J. Qiu, Y. Lei, J. Shelton, K. Higgins, J. D. Bradley, W. J. Curran, T. Liu, A. H. Kesarwala, and X. Yang, Deep Learning-Based Thoracic CBCT Correction with Histogram Matching, *Biomedical Physics & Engineering Express* **7** (2021), DOI:10.1088/2057-1976/ac3055.
- <sup>10</sup> G. Dong, Y. He, X. Liu, J. Dai, Y. Xie, and X. Liang, Better Cone-Beam CT Artifact Correction via Spatial and Channel Reconstruction Convolution Based on Unsupervised Adversarial Diffusion Models, *Bioengineering* **12**, 132 (2025), DOI:10.3390/bioengineering12020132.
- <sup>11</sup> C. Hu, N. Cao, X. Li, Y. He, and H. Zhou, CBCT-to-CT Synthesis Using a Hybrid U-Net Diffusion Model Based on Transformers and Information Bottleneck Theory, *Scientific Reports* **15**, 10816 (2025), DOI:10.1038/s41598-025-92094-6.
- <sup>12</sup> F. Madesta, T. Sentker, T. Gauer, and R. Werner, Self-Contained Deep Learning-Based Boosting of 4D Cone-Beam CT Reconstruction, *Medical Physics* **47**, 5619–5631 (2020), DOI:10.1002/mp.14441.
- <sup>13</sup> Z. Deng, H. Chen, Y. Zhou, H. Hu, Z. Xu, T. Lyu, Y. Xi, Y. Chen, J. Sun, and J. Zhao, RSTAR4D: Rotational Streak Artifact Reduction in 4D CBCT Using Separable 4D Convolutions, *IEEE Transactions on Radiation and Plasma Medical Sciences* (2025), DOI:10.1109/TRPMS.2025.3553866.
- <sup>14</sup> S. Rit, J. W. H. Wolthaus, M. van Herk, and J.-J. Sonke, On-the-Fly Motion-Compensated Cone-Beam CT Using an a Priori Model of the Respiratory Motion, *Medical Physics* **36**, 2283–2296 (2009), DOI:10.1118/1.3115691.

- <sup>15</sup> S. Rit, J. Nijkamp, M. van Herk, and J.-J. Sonke, Comparative Study of Respiratory Motion Correction Techniques in Cone-Beam Computed Tomography, *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* **100**, 356–359 (2011), DOI:10.1016/j.radonc.2011.08.018.
- <sup>16</sup> Z. Zhang, J. Liu, D. Yang, U. S. Kamilov, and G. D. Hugo, Deep Learning-Based Motion Compensation for Four-Dimensional Cone-Beam Computed Tomography (4D-CBCT) Reconstruction, *Medical physics* **50**, 808–820 (2023), DOI:10.1002/mp.16103.
- <sup>17</sup> Z. Zhang, Y. Hao, X. Jin, D. Yang, U. S. Kamilov, and G. D. Hugo, Fast Motion-Compensated Reconstruction for 4D-CBCT Using Deep Learning-Based Groupwise Registration, *Biomedical Physics & Engineering Express* **11**, 015030 (2024), DOI:10.1088/2057-1976/ad97c1.
- <sup>18</sup> G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, VoxelMorph: A Learning Framework for Deformable Medical Image Registration, *IEEE Transactions on Medical Imaging* **38**, 1788–1800 (2019), DOI:10.1109/TMI.2019.2897538.
- <sup>19</sup> Y. Zhang, X. Wu, H. M. Gach, H. Li, and D. Yang, GroupRegNet: A Groupwise One-Shot Deep Learning-Based 4D Image Registration Method, *Physics in Medicine & Biology* **66**, 045030 (2021), DOI:10.1088/1361-6560/abd956.
- <sup>20</sup> I. Herzig, P. Paysan, S. Scheib, A. Züst, F.-P. Schilling, J. Montoya, M. Amirian, T. Stadelmann, P. Eggenberger Hotz, R. M. Füchslin, and L. Lichtensteiger, Deep Learning-Based Simultaneous Multi-Phase Deformable Image Registration of Sparse 4D-CBCT, in *Medical Physics*, volume 49, pages e325–e326, American Association of Physicists in Medicine, 2022.
- <sup>21</sup> L. Zijp, J.-J. Sonke, and M. van Herk, Extraction of the Respiratory Signal from Sequential Thorax Cone-Beam X-ray Images, in *International Conference on the Use of Computers in Radiation Therapy*, pages 507–509, 2004.
- <sup>22</sup> A. Kavanagh, P. M. Evans, V. N. Hansen, and S. Webb, Obtaining Breathing Patterns from Any Sequential Thoracic X-Ray Image Set, *Physics in Medicine and Biology* **54**, 4879–4888 (2009), DOI:10.1088/0031-9155/54/16/003.
- <sup>23</sup> L. Gao, K. Xie, J. Sun, T. Lin, J. Sui, G. Yang, and X. Ni, A Transformer-Based Dual-Domain Network for Reconstructing FOV Extended Cone-Beam CT Images from Truncated Sinograms in Radiation Therapy, *Computer Methods and Programs in Biomedicine* **241**, 107767 (2023), DOI:10.1016/j.cmpb.2023.107767.
- <sup>24</sup> 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), DOI:10.1109/TMI.2018.2833499.
- <sup>25</sup> X. Zhao, Y. Du, and Y. Peng, DLPVI: Deep Learning Framework Integrating Projection, View-by-View Backprojection, and Image Domains for High- and Ultra-Sparse-View CBCT Reconstruction, *Computerized Medical Imaging and Graphics* **121**, 102508 (2025), DOI:10.1016/j.compmedimag.2025.102508.
- <sup>26</sup> G. Chen, Y. Lin, E. Elder, B. Ghavidel, M. W. McDonald, K. M. Langen, J. D. Bradley, T. Liu, and H. Gao, AirNet: Fused Analytical and Iterative Image Reconstruction Method With Deep Learning Regularization for High-Quality Sparse-Data On-Board CBCT, *International Journal of Radiation Oncology, Biology, Physics* **108**, e305 (2020), DOI:10.1016/j.ijrobp.2020.07.729.
- <sup>27</sup> 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 and Biology* **65**, 175020 (2020), DOI:10.1088/1361-6560/ab9f60.
- <sup>28</sup> C. Jailin, S. Roux, D. Sarrut, and S. Rit, Projection-Based Dynamic Tomography, *Physics in Medicine & Biology* **66**, 215018 (2021), DOI:10.1088/1361-6560/ac309e.
- <sup>29</sup> Y. Huang, K. Thielemans, G. Price, and J. R. McClelland, Surrogate-Driven Respiratory Motion Model for Projection-Resolved Motion Estimation and Motion Compensated Cone-Beam CT Reconstruction from Unsorted Projection Data, *Physics in Medicine & Biology* **69**, 025020 (2024), DOI:10.1088/1361-6560/ad1546.
- <sup>30</sup> A. Zakeri, A. Hokmabadi, M. G. Nix, A. Gooya, I. Wijesinghe, and Z. A. Taylor, 4D-Precise: Learning-based 3D Motion Estimation and High Temporal Resolution 4DCT Reconstruction from Treatment 2D+t X-ray Projections, *Computer Methods*

- and Programs in Biomedicine **250**, 108158 (2024), DOI:10.1016/j.cmpb.2024.108158.
- 31 H.-C. Shao, T. Mengke, T. Pan, and Y. Zhang, Dynamic CBCT Imaging Using Prior Model-Free Spatiotemporal Implicit Neural Representation (PMF-STINR), *Physics in Medicine & Biology* **69**, 115030 (2024), DOI:10.1088/1361-6560/ad46dc.
  - 32 P. Ramachandran, B. Zoph, and Q. V. Le, Searching for Activation Functions, *arXiv preprint arXiv:1710.05941* (2017), DOI:10.48550/arXiv.1710.05941.
  - 33 D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, *arXiv preprint arXiv:1412.6980* (2017), DOI:10.48550/arXiv.1412.6980.
  - 34 D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes, Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images, *IEEE Transactions on Medical Imaging* **18**, 712–721 (Aug./1999), DOI:10.1109/42.796284.
  - 35 J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, M. Bach, and M. Segeroth, TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images, **5**, e230024 (2023), DOI:10.1148/ryai.230024.
  - 36 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), DOI:10.1002/mp.15236.
  - 37 G. C. Mc Kinnon and R. H. T. Bates, Towards Imaging the Beating Heart Usefully with a Conventional CT Scanner, *IEEE Transactions on Biomedical Engineering* **BME-28**, 123–127 (1981), DOI:10.1109/TBME.1981.324785.
  - 38 S. Leng, J. Zambelli, R. Tolakanahalli, B. Nett, P. Munro, J. Star-Lack, B. Paliwal, and G.-H. Chen, Streaking Artifacts Reduction in Four-Dimensional Cone-Beam Computed Tomography, *Medical Physics* **35**, 4649–4659 (2008), DOI:10.1118/1.2977736.
  - 39 S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, Scikit-Image: Image Processing in Python, *PeerJ* **2**, e453 (2014), DOI:10.7717/peerj.453.
  - 40 V. Ghodrati, J. Shao, M. Bydder, Z. Zhou, W. Yin, K.-L. Nguyen, Y. Yang, and P. Hu, MR image reconstruction using deep learning: evaluation of network structure and loss functions, *Quantitative Imaging in Medicine and Surgery* **9** (2019).