Received: 27 September 2022 | Revised: 23 April 2023 | Accepted: 12 May 2023

DOI: 10.1002/mp.16705

RESEARCH ARTICLE

MEDICAL PHYSICS

Simultaneous object detection and segmentation for patient-specific markerless lung tumor tracking in simulated radiographs with deep learning

Lili Huang[1,2] | **Christopher Kurz**[1] | **Philipp Freislederer**[1] | **Farkhad Manapov**[1] | **Stefanie Corradini**[1] | **Maximilian Niyazi**[1] | **Claus Belka**[1,3,4] | **Guillaume Landry**[1] | **Marco Riboldi**[2]

[1]Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany

[2]Department of Medical Physics, Faculty of Physics, Ludwig-Maximilians-Universität München, München, Germany

[3]German Cancer Consortium (DKTK), partner site Munich, a partnership between DKFZ and LMU University Hospital Munich, Germany

[4]Bavarian Cancer Research Center (BZKF), Munich, Germany

Correspondence
Guillaume Landry, Department of Radiation Oncology, LMU University Hospital, LMU Munich, 81377 Munich, Germany.
Email: Guillaume.Landry@med.uni-muenchen.de

Present address
Philipp Freislederer, Brainlab AG, Munich, Germany

Guillaume Landry and Marco Riboldi shared senior authorship equally.

Funding information
FöFoLe program of the Faculty of Medicine of the LMU Munich, Grant/Award Number: 1113

Abstract

Background: Real-time tumor tracking is one motion management method to address motion-induced uncertainty. To date, fiducial markers are often required to reliably track lung tumors with X-ray imaging, which carries risks of complications and leads to prolonged treatment time. A markerless tracking approach is thus desirable. Deep learning-based approaches have shown promise for markerless tracking, but systematic evaluation and procedures to investigate applicability in individual cases are missing. Moreover, few efforts have been made to provide bounding box prediction and mask segmentation simultaneously, which could allow either rigid or deformable multi-leaf collimator tracking.

Purpose: The purpose of this study was to implement a deep learning-based markerless lung tumor tracking model exploiting patient-specific training which outputs both a bounding box and a mask segmentation simultaneously. We also aimed to compare the two kinds of predictions and to implement a specific procedure to understand the feasibility of markerless tracking on individual cases.

Methods: We first trained a Retina U-Net baseline model on digitally reconstructed radiographs (DRRs) generated from a public dataset containing 875 CT scans and corresponding lung nodule annotations. Afterwards, we used an independent cohort of 97 lung patients to develop a patient-specific refinement procedure. In order to determine the optimal hyperparameters for automatic patient-specific training, we selected 13 patients for validation where the baseline model predicted a bounding box on planning CT (PCT)-DRR with intersection over union (IoU) with the ground-truth higher than 0.7. The final test set contained the remaining 84 patients with varying PCT-DRR IoU. For each testing patient, the baseline model was refined on the PCT-DRR to generate a patient-specific model, which was then tested on a separate 10-phase 4DCT-DRR to mimic the intrafraction motion during treatment. A template matching algorithm served as benchmark model. The testing results were evaluated by four metrics: the center of mass (COM) error and the Dice similarity coefficient (DSC) for segmentation masks, and the center of box (COB) error and the DSC for

bounding box detections. Performance was compared to the benchmark model including statistical testing for significance.

Results: A PCT-DRR IoU value of 0.2 was shown to be the threshold dividing inconsistent (68%) and consistent (100%) success (defined as mean bounding box DSC > 0.6) of PS models on 4DCT-DRRs. Thirty-seven out of the eighty-four testing cases had a PCT-DRR IoU above 0.2. For these 37 cases, the mean COM error was 2.6 mm, the mean segmentation DSC was 0.78, the mean COB error was 2.7 mm, and the mean box DSC was 0.83. Including the validation cases, the model was applicable to 50 out of 97 patients when using the PCT-DRR IoU threshold of 0.2. The inference time per frame was 170 ms. The model outperformed the benchmark model on all metrics, and the comparison was significant ($p < 0.001$) over the 37 PCT-DRR IoU > 0.2 cases, but not over the undifferentiated 84 testing cases.

Conclusions: The implemented patient-specific refinement approach based on a pre-trained baseline model was shown to be applicable to markerless tumor tracking in simulated radiographs for lung cases.

KEYWORDS
DRR, deep learning, lung tumor tracking, markerless tracking

1 | INTRODUCTION

Respiratory motion of tumors for lung cancer patients poses challenges for accurate dose delivery in radiotherapy. Many motion management methods have been proposed to address this problem, including motion encompassment,[1,2] respiratory gating,[3,4] breath holding,[5,6] forced shallow breathing,[7,8] and real-time tumor tracking.[9–11] Real-time tumor tracking requires accurate knowledge of the tumor position while the dose is delivered. One of the most used tumor tracking techniques is kV X-ray imaging.[12–15] The low soft-tissue contrast of X-ray images, however, often limits the visibility of the tumor in projection images. A common solution is to implant fiducial markers at or near the target site, and then establish a correspondence between the highly visible fiducials and the target.[16,17] This tracking-with-markers approach is not ideal, because marker implantation prolongs the treatment period and is an invasive procedure incurring risks of complications.[18] Furthermore, the potential migration of markers suggested by previous studies adds to the uncertainty of the prebuilt correspondence between the marker and the target location.[18] Research efforts have been put into markerless tumor tracking using traditional methods such as template matching[19] and using a correlation model.[20] A study investigating the applicability of markerless tracking for lung cases showed that only 66% of the already pre-selected patients had passed the initial tumor visualization test and proceeded to be treated with this approach.[21]

In recent years, multiple investigations have started exploring markerless tumor tracking with kV or MV X-ray imaging using machine learning-based methods.[22–26] Considering the difficulty of collecting a large amount of data to train generic models, most studies[22–25] directly employed a patient-specific approach as the training strategy. For each patient they trained the model from scratch on digitally reconstructed radiographs (DRRs) generated from pre-treatment four-dimensional computed tomography (4DCT) images of the specific patient, employing various forms of data augmentation. Zhao et al. investigated the localization of pancreatic[22] and prostate[23] tumors on 2D DRRs by using Faster R-CNN and predicting coordinates of the top-left corner of target bounding boxes. They interpolated 4DCT data between phases and randomly split the interpolated 4DCT for training and testing. Zhou et al.[26] fine-tuned a Mask R-CNN model, which was pre-trained on a non-medical dataset (COCO dataset), on augmented 4DCT-DRRs. They employed a deep leaning-based model to interpolate directly 4DCT images, and did a similar split of interpolated data as Zhao et al. Although their Mask R-CNN model could output both bounding boxes and segmentation masks, they only evaluated the performance of the segmentation prediction. Takahashi et al.[24] used a fully convolutional neural network (FCN) and conducted both a digital simulation study and an epoxy chest phantom study for lung tumor tracking with the prediction output being pixel-wise segmentation masks. For their digital simulation study, they trained on augmented 4DCT-DRRs and tested on original unaugmented 4DCT-DRRs of the same digital phantom. Therefore, the generalization capabilities of the developed model were not specifically tested. The train/test division of datasets is crucial for the proper and accurate evaluation of machine learning-based models, thus it is highly important to ensure the independence between training and testing data.[27] On the other hand, the phantom study carried out by

Takahashi et al. used all 4DCT-DRRs of an epoxy respiratory motion phantom for training and tested on X-ray images of the same phantom. This type of train/test scheme would fit well into the current treatment procedure, nevertheless, this study was limited by the oversimplified phantom anatomy. Without the need of performing any interpolation, Sakata et al.[25] conducted a lung tumor tracking study on eight patients exploiting the extremely randomized trees (ERT) method to predict segmentation masks. They used all 4DCT-DRRs for training and tested on incoming fluoroscopic images of the same patient, which represented treatment scenarios the best. However, for markerless tumor tracking with kV X-ray imaging, current studies lack prediction and evaluation of both the target bounding boxes and the segmentation masks. Attempts to investigate the applicability of the proposed models on individual cases are also missing.

In this paper, we aimed to emulate real-time tumor tracking on in-room kV X-ray images by tracking tumors in anterior-posterior DRRs as a first step to examine the feasibility of simultaneously detecting and segmenting the target. To this purpose, we trained and evaluated a Retina U-Net model for markerless lung tumor tracking, which is able to output simultaneously both the target bounding boxes and the segmentation masks to provide redundant information for tracking. A bounding box prediction provides the two-dimensional coordinates of a rectangular box enclosing the detected target along with a predicted class label for the target, fulfilling the task of object detection. A segmentation mask, on the other hand, gives a pixel-wise classification on the input image separating the foreground object from the background. Having both types of predictions allows for better flexibility for dynamic multi-leaf collimator (MLC) tracking. For rigid target motion, prediction of the target bounding box is sufficient for adapting the MLC aperture; whereas for cases where the target deforms while moving, prediction of the target mask conforming to the target's changing shape is required. We chose the Retina U-Net architecture for this, which combines object detection and segmentation tasks by adding additional semantic segmentation supervision to its feature extractor.

While other works used the same motion pattern contained in 4DCT data for training and testing, we separated the two by using DRRs from 4DCT data exclusively for testing while using DRRs from 3D planning CT (PCT) data for training. In general, two methodologies are possible for developing models for tumor motion tracking: one is to develop a general prediction model that is applicable to all patients, the other is to develop a systematic approach to produce models customized and applicable to each patient individually. A hybrid method combining both has also been previously examined.[28] Having a general model ready to work at any time requires less efforts and time compared to the necessity of adapting the patient-specific approach to each patient before treatment. Also, a general model likely has been tested on many data before application, as opposed to a newly created patient-specific model. On the other hand, given that in radiotherapy patients' pre-treatment data such as the PCT are usually available, taking advantage of such prior knowledge to create a patient-specific model might yield a performance superior to what a generic model could achieve. In the present study, we developed patient-specific models by individually fine tuning a baseline model, that was trained on a large public dataset, on DRRs from a given patient's 3D PCT image which would be available prior to treatment in a clinical workflow.

2 | MATERIALS AND METHODS

2.1 | Datasets

Two separate datasets were used for training the baseline and patient-specific models. The first dataset was a public dataset, the Lung Imaging Database Consortium (LIDC),[29] that contained 875 CT scans and corresponding lung nodule annotations. The CT scans have varying voxel size and CT physical extent but were resampled to the same voxel size of 0.7 mm× 0.7 mm × 1.25 mm before being used (while preserving their physical extent). Originally, the LIDC dataset has multiple versions of annotation for the same nodule, which were drawn by multiple radiologists separately. Since inter-observer variability is out of scope of this study, we randomly selected one annotation for each nodule as the ground-truth segmentation label.

The second dataset was a cohort of 97 lung-cancer patients treated at the University Hospital of LMU Munich. For each of the 97 patients, the dataset contained a PCT with its corresponding gross tumor volume (GTV) segmentation, and a 10-phase 4DCT. All CT volumes had a voxel size of 1.074 mm × 1.074 mm × 3 mm in left-right (LR), anterior-posterior (AP), and superior-inferior (SI) directions. Their voxel arrays had a size of 512 × 512 in LR and AP directions, and varying sizes in SI direction. The 97 patients were selected so that the volumes of their GTVs were between 725 and 26485 mm^3, which were respectively the minimal and maximal volumes of LIDC nodules used during baseline model training (see section 2.4). We performed B-spline deformable image registration between PCT and 4DCT to warp the GTV segmentations of PCT to those of 4DCT (see details in Table A-1 in the supplementary materials).

2.2 | Radiograph simulation

To prepare data for training and testing, all CT volumes were forward projected into simulated X-ray images,

DRRs, using the open-source software Reconstruction Toolkit (RTK).[30] Before simulation, the Hounsfield unit (HU) values of the CT volumes were converted to their corresponding relative attenuation coefficients according to the formula:

$$\frac{\mu_{\text{material}}}{\mu_{\text{water}}} = 0.001 \times \text{HU} + 1 \tag{1}$$

where μ_{material} and μ_{water} represent the attenuation coefficients for the material and water respectively. Geometry settings for the simulation, except the projection angle, were taken from the ExacTrac Dynamic (Brainlab, Munich, Germany) X-ray imaging system in our clinic, which could be used in the future for lung-cancer markerless tracking. The source-to-isocenter distance was set to 2190 mm, and the source-to-detector distance was 3509 mm. In this proof of principle study, we chose 0° as the projection angle rather than the oblique angles of the actual imaging system. The detector panel has a size of 768 × 768 pixels with a pixel size of 0.388 mm× 0.388 mm after a 2 × 2 binning. To get a larger field of view of the lungs and more patches for our patch-based training, we chose to simulate DRRs with a double size of 1536 × 1536 in pixels and the same pixel size of 0.388 mm× 0.388 mm. The chosen model architecture of this study, explained in the next section, can work with DRRs of any size larger than the patch size (see section 2.4.3 for more details on patch size). At the time of testing the model, a single patch of size similar to the clinical detector size was used.

2.3 | Model architecture and training strategy

We used Retina U-Net[31] as the model architecture for the target localization and segmentation task. Retina U-Net is a one-stage detector, which is able to fuse an object detection task with a semantic segmentation task by adding additional segmentation supervision signals to the feature pyramid network. Its loss function is composed of three parts: cross-entropy loss for the classification task, smooth L1 loss for the bounding box regression task, and cross-entropy and soft Dice loss with equal weights for the segmentation task. Compared to a regular Retina Net, Retina U-Net has an additional segmentation loss, which is shown in Equation (2).

$$\begin{aligned} L_{\text{seg}} &= L_{\text{CE}} - \frac{2}{|K|} \sum_{k \in K} \frac{\sum_{i \in I} u_{i,k} v_{i,k}}{\sum_{i \in I} u_{i,k} + \sum_{i \in I} v_{i,k}}, \\ L_{\text{CE}} &= \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise} \end{cases} \end{aligned} \tag{2}$$

where L_{seg} is the total segmentation loss and L_{CE} represents the cross-entropy loss with y being a binary indicator (which is assigned as 1, representing the condition "True", when the class label is correctly predicted) and p being the output probability for the class. The second term of the total segmentation loss represents the soft Dice similarity coefficient (DSC) loss (calculated as 1 - DSC), where u is the network's softmax output and v is the ground-truth segmentation map in one-hot encoding. u and v have the same shape of $I \times K$ with I being the total number of pixels and K being the number of classes. The notations, $u_{i,k}$ and $v_{i,k}$, respectively represent the i-th pixel for class k in the softmax output and in the ground-truth one-hot segmentation map.

The input to the model was an image and the output was a target bounding box, a class label along with confidence score, and a segmentation mask. Figure 1 illustrates the network architecture that can be divided into three parts. The first part is the feature extractor on the left, which is built on a symmetric Feature Pyramid Network (FPN). In Retina U-Net, the FPN is extended to a full U-Net by adding high resolution pyramid levels enabling the output of a segmentation mask. As the backbone of the entire network, the FPN extracts convolutional features from the input image at different resolution levels (from highest at the bottom to lowest at the top). Two parallel sub-networks, one for classification and the other for box regression, are attached to coarser pyramid levels of the FPN. At inference time the final bounding box prediction is obtained by merging and thresholding predictions made on all levels via the weighted box clustering algorithm.[31]

In this study only one class, i.e. tumor, is of interest and was labeled as 1 (background as 0). We used the 2D version of the model since DRRs are two-dimensional projection images. We used the code shared by Jaeger on github https://github.com/MIC-DKFZ/medicaldetectiontoolkit and performed minor adaptations to enable access to the segmentation object. We also added post-processing steps for segmentation, which included first performing a connected component analysis, then filtering out pixels that were not connected to any pixels inside the predicted bounding boxes, and finally binary hole filling, closing and opening.

The training was done in two stages: (1) We first trained a baseline model using patient data in the LIDC dataset. (2) Then we refined the baseline model for each hospital patient by continuing training using patient-specific 3D PCT-DRRs only. 4DCT-DRRs were never used for training the model weights.

The code for training, validating and testing was implemented within an NVIDIA CUDA docker container (Docker 20.10.12, CUDA 9.0, cuDNN 7.6) that has PyTorch 0.4.1 and Python 3.6 installed. The experiments were ran on a server equipped with 376 GB of RAM, an

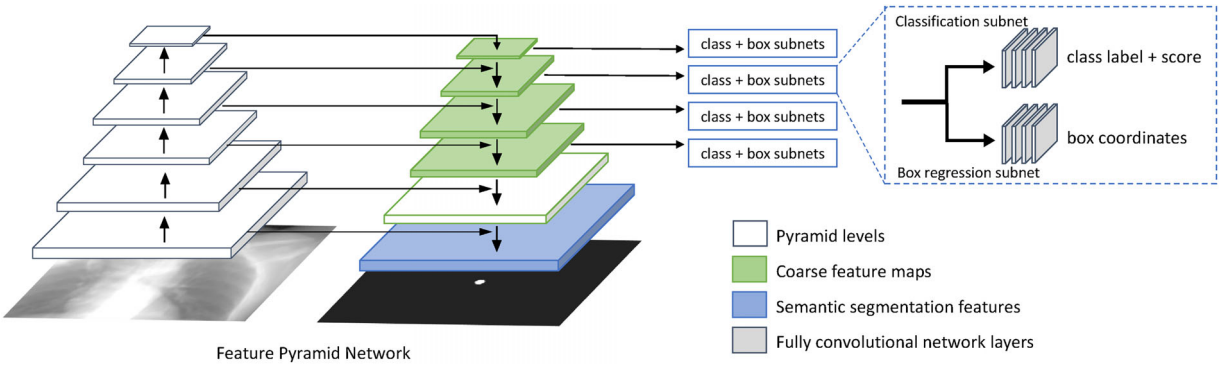

FIGURE 1 Network architecture. The network input on the left is a 2D DRR. The outputs consist of bounding boxes with labels and scores on the rightmost part of the figure, as well as a 2D segmentation mask at the end of the decoding arm. Only one of the four classification and box subnets has been expanded for better visibility. DRR, digitally reconstructed radiograph.

Intel Xeon Gold 6254 CPU (3.10 GHz, 36 cores), and an NVIDIA Quadro RTX 8000 GPU (48 GB).

2.4 | Training of baseline model

2.4.1 | Training data selection

To explore the feasibility of detecting tumors on DRRs with Retina U-Net, we first targeted high-visibility cases from the LIDC data defined by two criteria: (1) The tumor volume was greater than 700 mm^3. (2) The tumor had a good contrast against its background in DRRs, which was subjectively judged by visually checking all the DRRs. Of 2625 distinct LIDC nodules, approximately 14% satisfied the volume criterion, and of these volume-filtered nodules, around 22% passed the visual inspection. Finally, we selected 82 patients with high-visibility tumors among 863 LIDC patients (12 patients had cone beam computed tomography scans and were excluded).

2.4.2 | Data augmentation

Given the small remaining dataset size and inspired by Schultheiss et al.,[32] we synthesized a larger amount of CT volumes with nodules that matched both the volume and contrast standards by inserting nodules of those high-visibility patients into CT volumes of other low-visibility patients. The procedure is illustrated in Figure 2. Before insertion, nodules from the 82 high-visibility cases were extracted from their original CT volumes and split into two pools: 80% training and 20% validation. In parallel, CT volumes of the rest of the 781 low-visibility cases first went through nodule removal, in which voxel values of nodules were replaced with normal lung tissue's HU value of −850. Then their nodule-free CTs were divided into training and validation pools according to the same 80:20 ratio. Nodules in the training pool were only inserted into CT volumes of the training pool, and the same for the validation pool. All CT volumes and nodules were resampled to have the same voxel size of 0.7 mm× 0.7 mm × 1.25 mm.

Upon insertion, nodules were randomly scaled by a ratio within 0.9 and 1.2, randomly rotated within 0° to 360° in all three axes, and inserted at a randomly sampled location within the lung mask of the selected CT volume. These lung masks were eroded with a kernel radius of 20 × 20 × 20 pixels to ensure that nodules stayed properly within the lungs and were not too close to the mediastinal region after insertion. To mimic the natural nodule number distribution in the original LIDC dataset, 80%, 15%, and 5% of the clean CT volumes were inserted with 1, 2, or 3 nodules, respectively. The nodule-only masks for the synthetic CT volumes were obtained by padding the much smaller binary nodule masks extracted before insertion to the same size of their corresponding inserted CTs and setting their new origin coordinates as those of the inserted CTs.

After insertion, all synthetic CT volumes and their corresponding nodule masks were forward projected to generate DRRs and their nodule segmentations, which were then screened by their nodules' foreground-to-background contrast in the projection image to ensure high visibility. The foreground-to-background contrast was calculated as the difference between the median pixel values of the nodule foreground and the nodule's surrounding background region (defined by a bounding box centered at the nodule center with its length and width being 1.2 times the nodule's length and width). Only DRRs with contrasts higher than 5.0 were used for baseline model training, which amounts to 61% of the total generated DRRs.

In total, we generated 15 452 synthetic DRRs with corresponding nodule masks for training and 1429 for validation. At train-time the input further went through an on-the-fly augmentation pipeline consisting of rotation, scaling, and elastic deformation, using the python package Batchgenerators.[33] The rotation angle was between 0° and 360° and the scaling range was between 0.8 and 1.1. For elastic deformation, we used the default parameters for alpha and sigma as (0, 1500) and (30, 50).[34] Values were randomly sampled from the given

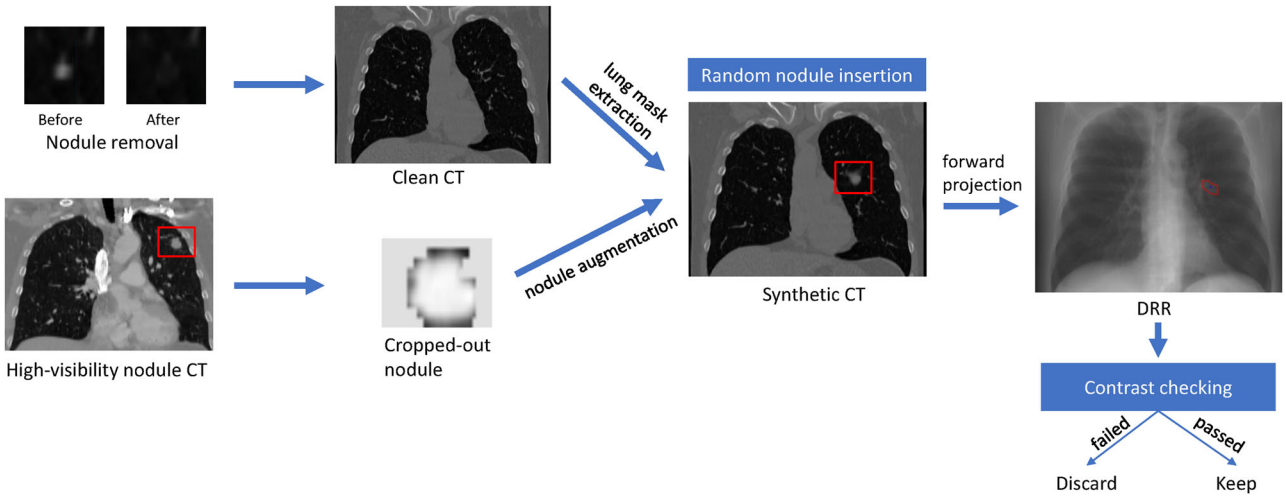

FIGURE 2 An example of the nodule insertion procedure.

intervals during augmentation. The generated DRRs had a pixel intensity range of 0 to 300, and no intensity normalization was involved.

2.4.3 | Training and validation

To account for the potential discrepancy between original LIDC data and synthesized LIDC data, we introduced another validation set composed of 16 unmodified LIDC patients, who were from the 20% validation pool of the initial 82 high-visibility patients. This unmodified validation set served as an indicator of how well the baseline model will work on DRRs of unmodified CT scans.

The input DRRs had a size of 1536 × 1536 in pixels. Since the model's prediction is patched-based, cropping of the input images would become necessary when a patch size different from the input image size is selected. For the baseline model, we tuned the following hyperparameters: the image patch size, the learning rate, the matching IoU threshold, image intensity clipping, and augmentation probability. After tuning the baseline model hyperparameters, we set the pre-crop size to 700 × 700 and the patch size to 672 × 672. At runtime, the data loading pipeline first cropped every 1536 × 1536 input image into nine sub-images of size 700 × 700 with overlap between each other, then sampled these sub-images to make a training batch with a 50:50 balance between target-containing and non-target-containing patches. Afterwards, during the on-the-fly data augmentation, the pre-cropped patches were centrally cropped to slightly smaller patches of size 672 × 672. The second cropping avoids potential border artifacts induced by the spatial augmentation. The Adam optimizer was used and the learning rate was set to 0.0001. The matching intersection over union (IoU), defined as the overlapping area of the predicted and the ground-truth target bounding boxes over the combined area of the two boxes, was set to 0.1. This hyperparameter is used during training, where the evaluation of the Retina U-Net classification loss function assumes true positive detection for IoU > 0.1.

The baseline model was evaluated in terms of average precision (AP) and detection accuracy. AP is a commonly used metric for object detection tasks and captures the precision-recall curve. It is defined as the weighted sum of precisions at each confidence score threshold with the weights being the increase in recall from the previous score threshold, as described in Equation (3):

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{3}$$

where P_n denotes the precision determined at confidence score threshold n, and R_n and R_{n-1} are respectively the recalls at score threshold n and n-1. The other metric, detection accuracy, is defined as the percentage of tumors detected with IoU higher than the matching IoU per patient. Among positive detections, the mean absolute error of centers of predicted boxes, the mean distance between centers of mass (COM) and the DSC between ground-truth and predicted segmentations were further evaluated.

2.5 | Training and testing of patient-specific models

2.5.1 | Evaluation metrics

For patient-specific models, we focused on metrics that are more relevant to the tracking purpose of this study: center of mass (COM) error for segmentation prediction, center of box (COB) error for box prediction,

and DSC for both box and segmentation predictions. The COM error (the COB error) was calculated as the Euclidean distance at isocenter plane between predicted and ground-truth (GT) segmentations (bounding boxes).

2.5.2 | Data partition

The second LMU dataset containing 97 patients was used for investigating a scheme of patient-specific training. The baseline model was applied to each of the 97 patients' PCT-DRR. For each patient, we recorded the IoU value with which the baseline model detected the tumor in the PCT-DRR. We labelled this as the PCT-DRR IoU. In the scenario of tracking tumors in real time on kV X-ray images, no X-ray images are acquired at the treatment planning stage. Because of this, the patient-specific training procedure cannot be validated during training and needs to be fully automated in order to be meaningful for clinical scenarios. This necessitates finding optimal hyperparameters for patient-specific training prior to final testing, similar to the idea explored by Teo et al.[28] Hence, we selected 13 patients with higher PCT-DRR IoU (all greater than 0.7) among the 97 patients for hyperparameter tuning and validation of the patient-specific training approach. We chose this relatively high IoU threshold to ensure that hyperparameter selection was not biased by cases for which tracking is not possible. The 84 remaining patients, with PCT-DRR IoU ranging from 0.0 to 0.9, were reserved for final testing. The four evaluation metrics were calculated respectively per PCT-DRR IoU intervals of 0.1 width.

2.5.3 | Training and hyperparameter tuning

Each incoming patient's PCT-DRR and its tumor segmentation mask were taken as input by the already pre-trained baseline model for patient-specific fine-tuning. The single image went through the same augmentation pipeline as that of the baseline model training (see section 2.4.2) on the fly, but with slightly different parameters for rotation and elastic deformation. The maximal rotation angle was limited to 10° for all patient-specific training. A refined model specific to a given patient was obtained at the end of the patient-specific training. Afterwards, to simulate the tracking scenario, the patient-specific model was applied on the 10-phase 4DCT-DRRs, which were obtained from the 10-phase 4DCT through forward projection. Figure 3 describes the complete procedure.

As explained in section 2.5.2, hyperparameters should be shared across all patients. To search for the optimal set of hyperparameters, we experimented on 13 patients to determine the best learning rate, magnitude of elastic deformation during augmentation and number of epochs for patient-specific training. Essentially, we evaluated the average performance on the 4DCT-DRRs of the 13 patients for different combinations of hyperparameters and chose the best configuration for final testing using the metrics mentioned above.

2.5.4 | Testing

Following the procedure shown in Figure 3, we refined the baseline model on each of the 84 remaining test-set patients' PCT-DRR using the pre-determined hyperparameters, and finally tested on their 4DCT-DRRs. To speed up the inference step, 4DCT-DRRs were directly cropped to the patch size of 672×672 beforehand while keeping the image center unchanged. This choice of patch size was optimized at the baseline model stage and it does not differ much from the regular image size 768×768 of our X-ray imaging system after binning.

2.5.5 | Benchmarking

As a benchmark, we implemented the template matching method with fast normalized cross-correlation (using the Python package scikit-image[35]). For each of the 84 test-set patients, the PCT-DRR and its ground-truth tumor segmentation mask were used to extract the bounding box template for the patient. At test time, the location in the testing image that had the highest cross-correlation value with the pre-extracted template was computed and considered as the location of the bounding box. The method was tested on the same 84 test-set patients' 4DCT-DRRs and evaluated in terms of COB errors and box DSC for the resulting box predictions. The COB error per phase and the box DSC per phase were calculated and compared via boxplots for patient-specific models and the template matching method, and for two groups of test-set patients: the full group of 84 patients and the subgroup made up of 37 patients with PCT-DRR IoU > 0.2. The Wilcoxon signed-rank test was performed on the paired per-phase results for the two methods in order to test the statistical significance.

3 | RESULTS

As presented in Figure 4, over the course of training, the baseline model showed slight overfitting in terms of total loss and no overfitting in terms of AP on the unmodified LIDC validation set. The baseline model reached an AP of 69% and a detection accuracy of 71% on the unmodified validation set that included 16 LIDC patients and 21 nodules in total. Figure 5 showcases two success/miss examples of the model's predictions on two patients from the unmodified validation set. The top row shows a successful case where the baseline model correctly

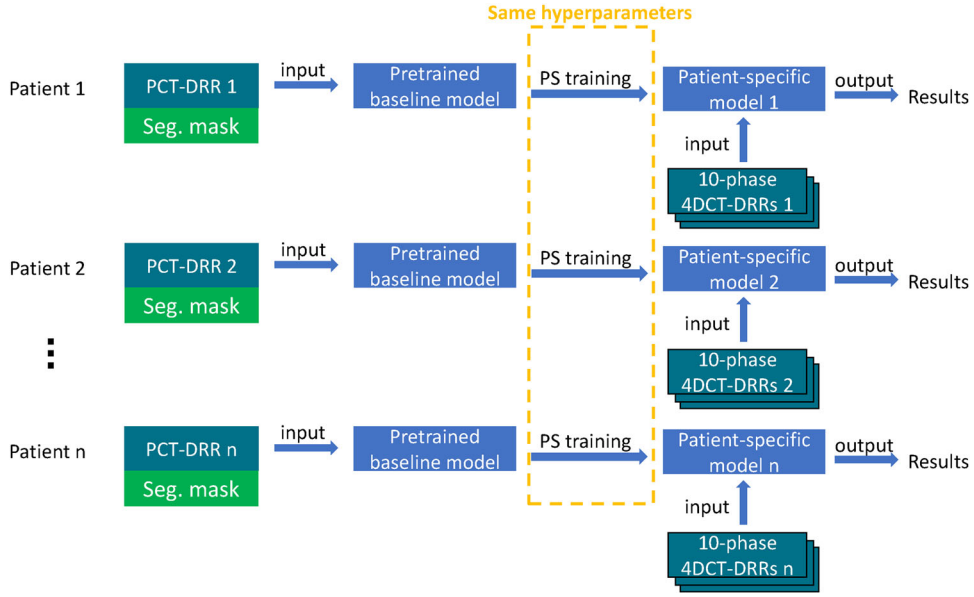

FIGURE 3 The patient-specific training procedure.

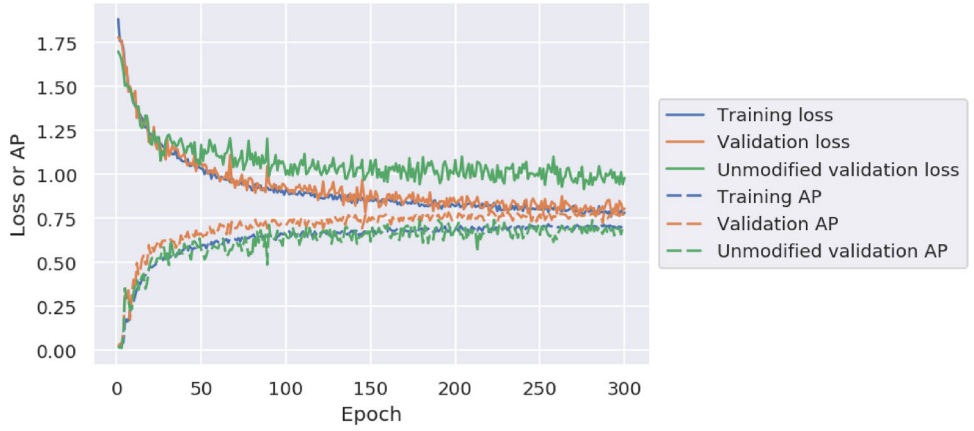

FIGURE 4 Loss curves (solid lines) and AP curves (dashed lines) for the training set (blue), the validation set (orange), and the unmodified validation set (green) over baseline training epochs. AP, anterior-posterior.

identified all three of the nodules present in this patient with relatively high confidence scores. Predictions for the two nodules in the lower part show good agreement with the ground truth for both boxes (2.2 mm and 1.7 mm for COB distance, respectively) and contours (0.78 and 0.77 for DSC, respectively). Prediction for the nodule at the top was slightly shifted upwards compared to the ground-truth, and its confidence score, 79, is lower than those of the other two nodules, which are both very close to the full score of 100. The bottom row shows a failed case where the baseline model missed the true target located at the patient's right lung and incorrectly predicted an oval-shaped structure in the DRR in the lower left lung as target. Among all true positive detections of the 16 LIDC validation patients, the mean absolute COB error was 1.5 and 2.1 mm in left-right and superior-inferior directions, the mean COM distance calculated from the segmentation was 3.0 mm and the mean segmentation DSC was 0.71.

For a consistent training setup of patient-specific (PS) models, the hyperparameter tuning results suggested a learning rate of 1e-5, an alpha of 1500 for deformation augmentation, and a patient-specific training epoch number of 15. With these parameters, 84 PS models were trained for each of the 84 test-set patients and then tested on these patients' respective 10-phase 4DCT-DRRs. The inference time per frame running on an Nvidia GPU (Quadro RTX 8000 with 48 GB of memory) was approximately 170 ms.

All PS models and the baseline model were applied on patients' 10-phase 4DCT-DRRs. Figure 6 shows the failure/success occurrences for baseline and PS models on 84 test-set patients individually, excluding the 13 patients selected for fine tuning experiments in section 2.5.3.

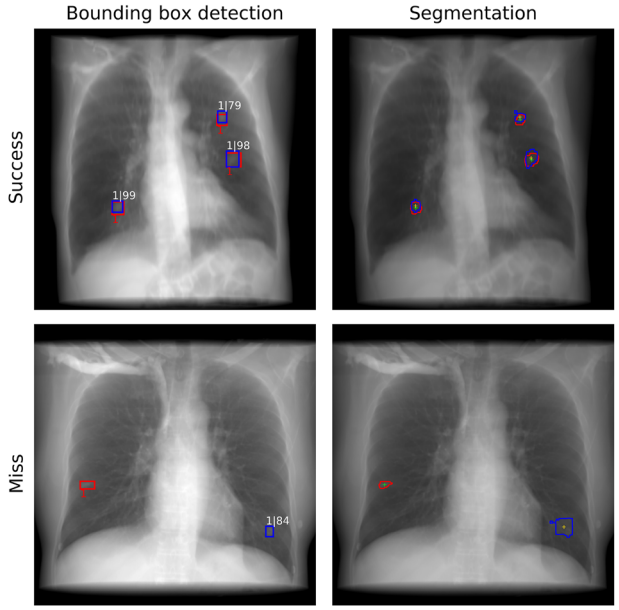

FIGURE 5 Examples of predictions made by the baseline model. Top and bottom rows show a successful case and a missed case, respectively. Ground-truth bounding boxes and segmentation contours are drawn in red; predicted bounding boxes and contours in blue; ground-truth/predicted centers of mass are drawn as greed/yellow crosses respectively. Numbers next to the boxes represent the nodule's class label, which is always 1 in this case. For predicted boxes, the confidence score related to the detection ranging from 0 to 100 is given next to the class label. All images shown in this figure have a size of 1536 × 1536 in pixels.

Specifically in this figure, we defined the failed condition as having the mean box DSC averaged over the patient's 10 4DCT phases lower than or equal to 0.6. One can observe from the figure that while the baseline model failed on some cases, PS models could improve performance for several cases. To determine for which patient PS tracking is successful, we established an IoU threshold of 0.2, above which there were no failed cases in Figure 6. For cases below the threshold, the baseline model failed on all patients except one, and the PS models succeeded in 32 out of 47 patients (68%).

The performance of PS models was further evaluated in terms of the four metrics averaged over the 10 4DCT phases of each patient and over patients within each PCT-DRR IoU interval. The results are presented in Table 1 and Table 2. Table 1 presents results for the test-set patients with PCT-DRR IoU threshold of 0.2 (37 cases). Further details of these 37 patients are summarized in Table B-2 in the supplementary materials. 30 out of 37 patients' PCT-DRR IoUs fell within the middle range between 0.3 and 0.7, with (0.5, 0.6] being the most populated interval, and all patients' IoUs were below 0.9. All seven IoU groups above 0.2 gave similar results. On average, for the 37 cases, the mean COM error was 2.6 mm, mean segmentation DSC was 0.78, mean COB error was 2.7 mm, and mean box DSC was 0.83.

TABLE 1 Testing results of patient-specific models on 4DCT-DRRs for various PCT-DRR IoU groups (above 0.2) of patients. Best results of evaluation metrics are highlighted in bold.

PCT-DRR IoU interval	Number of patients	Mean COM error [mm]	Mean seg. DSC	Mean COB error [mm]	Mean box DSC
(0.2, 0.3]	2	**1.9**	0.74	2.6	0.72
(0.3, 0.4]	6	2.3	0.80	**2.2**	0.84
(0.4, 0.5]	7	2.9	0.76	3.1	0.83
(0.5, 0.6]	12	2.4	0.79	2.5	0.85
(0.6, 0.7]	5	3.4	0.70	3.2	0.76
(0.7, 0.8]	4	2.2	**0.84**	2.6	0.87
(0.8, 0.9]	1	3.4	0.83	3.4	**0.88**
(0.2, 0.9]	37	2.6	0.78	2.7	0.83

Abbreviations: 4DCT, four-dimensional computed tomography; COB, center of box; COM, center of mass; DRR, digitally reconstructed radiograph; DSC, Dice similarity coefficient; IoU, intersection over union; PCT, planning CT.

TABLE 2 Testing results of patient-specific models on 4DCT-DRRs for patients with PCT-DRR IoU below 0.2. Patients were separated into two groups representing failure and success respectively: mean box DSC $\leq$ 0.6 and mean box DSC $>$ 0.6.

Mean box DSC interval	Number of patients	Mean COM error [mm]	Mean seg. DSC	Mean COB error [mm]	Mean box DSC
(0., 0.6]	15	29.2	0.22	61.6	0.22
(0.6, 1.0]	32	3.0	0.70	2.8	0.81
(0., 1.0]	47	11.0	0.54	21.5	0.62

Abbreviations: 4DCT, four-dimensional computed tomography; COB, center of box; COM, center of mass; DRR, digitally reconstructed radiograph; DSC, Dice similarity coefficient; IoU, intersection over union; PCT, planning CT.

Testing results of PS models on the remaining 47 test-set patients with PCT-DRR IoU in the [0, 0.2] interval are presented in Table 2. These cases were further divided into two groups according to their mean box DSC averaged over 4DCT phases representing failure (mean box DSC below 0.6) and success (mean box DSC above 0.6) respectively. The 15 failed cases (DSC $\leq$ 0.6) had notably worse performance for all four metrics. For the other 32 cases (mean box DSC $>$ 0.6) where PS models succeeded, PS models achieved a fairly good performance (second row): mean COM error of 3.0 mm, mean seg. DSC of 0.70, mean COB error of 2.8 mm, mean box DSC of 0.81. The last two box-related scores were even comparable to those for the PCT-DRR IoU interval (0.2, 0.9] (mean COB error at 2.7 mm, mean box DSC at 0.83, as listed in the last row of Table 1).

Since several PS models failed on the low PCT-DRR IoU group [0, 0.2], the rest of the testing results will only be shown for patients with PCT-DRR IoU above 0.2.

The comparison of testing results using the baseline model (in blue) and patient-specific (in orange) models are illustrated in Figure 7. PS models clearly

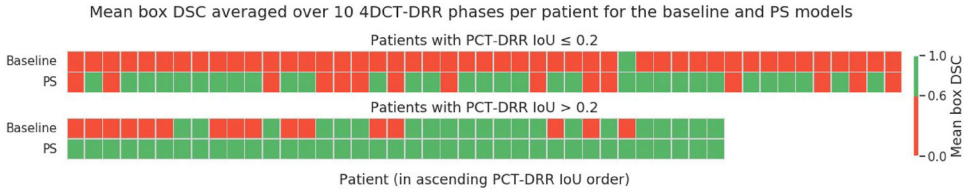

FIGURE 6 Baseline vs PS models in terms of mean box DSC averaged over the 10 4DCT-DRR phases for each patient. Every rectangular tile represents one patient tracked by the baseline or the PS model. Patients are lined up from left to right along the x axis in ascending order of PCT-DRR IoU. Top row represents patients with PCT-DRR IoU ≤ 0.2 (totaling 47) and bottom represents those with PCT-DRR IoU > 0.2 (totaling 37). Patients tracked with mean box DSC > 0.6 were considered success and marked as green tiles; patients tracked with mean box DSC ≤ to 0.6 were considered failures and marked as red tiles. 4DCT, four-dimensional computed tomography; DRR, digitally reconstructed radiograph; DSC, Dice similarity coefficient; IoU, intersection over union; PCT, planning CT; PS, patient-specific.

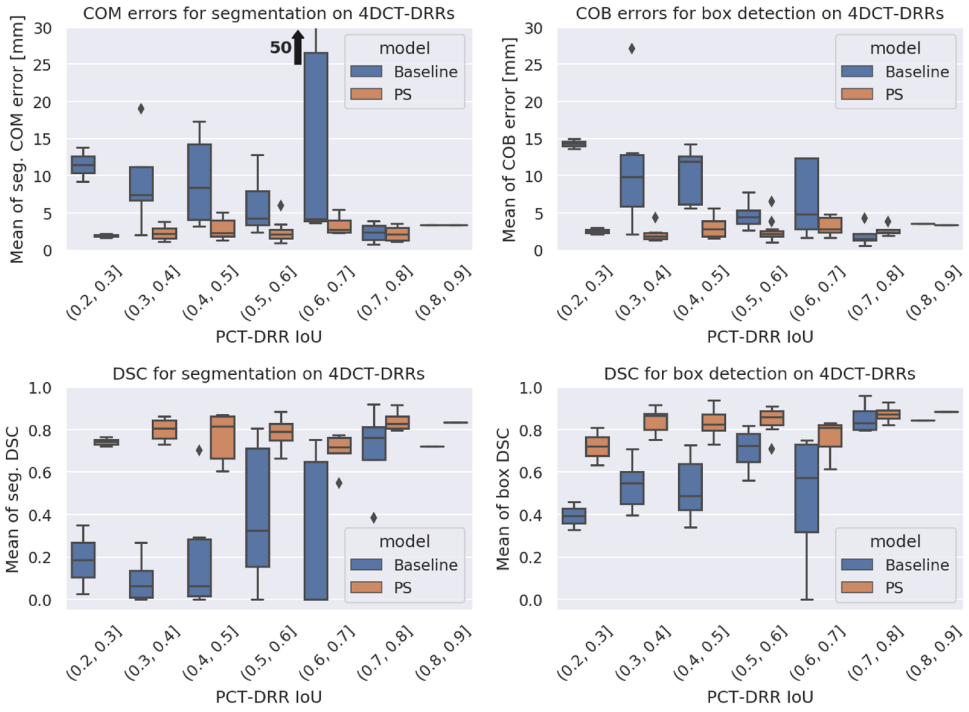

FIGURE 7 Comparison between testing results on 4DCT-DRRs of baseline model (blue) and PS models (orange) by PCT-DRR IoU intervals. Top left: COM errors calculated from segmentations on 4DCT-DRRs, the black arrow next to the blue box at the (0.6, 0.7] interval signifies that the baseline's maximal COM error went up to 50 mm, bottom left: DSC between ground-truth and predicted segmentation, top right: COB errors for bounding box detection, bottom right: DSC between ground-truth and predicted boxes for bounding box detection. 4DCT, four-dimensional computed tomography; COB, center of box; COM, center of mass; DRR, digitally reconstructed radiograph; DSC, Dice similarity coefficient; IoU, intersection over union; PCT, planning CT; PS, patient-specific.

outperformed the baseline model in all aspects, except being slightly worse in the (0.7, 0.8] interval for mean COB errors. The maximal mean COM errors in the interval (0.6, 0.7] ranged up to 50 mm for the baseline model. At higher IoU intervals above 0.7, the baseline model and PS models had comparably good performance on all four metrics. Overall, the baseline model had higher accuracy in box detection (blue in the right column) than in segmentation (blue in the left column), while PS models performed similarly in both tasks.

To demonstrate lung tumor tracking with PS models using 4DCT-DRRs, Figure 8 shows three examples of patients for the start and end of exhalation. These three patients were selected based on the per-patient analysis of box DSC over 10 phases (Figure 9) so that they respectively represent good (Patient 23, mean box DSC: 0.91), average (Patient 30, mean box DSC: 0.81), and poor (Patient 17, mean box DSC: 0.61) box agreement to the ground truth. Patient 23 showed consistently good box alignment with the ground-truth boxes across phases, although its segmentation predictions suffered from artifacts near the end of exhalation around the lower boundary of the tumor. Patient 30, on the other hand, had compromised box agreement

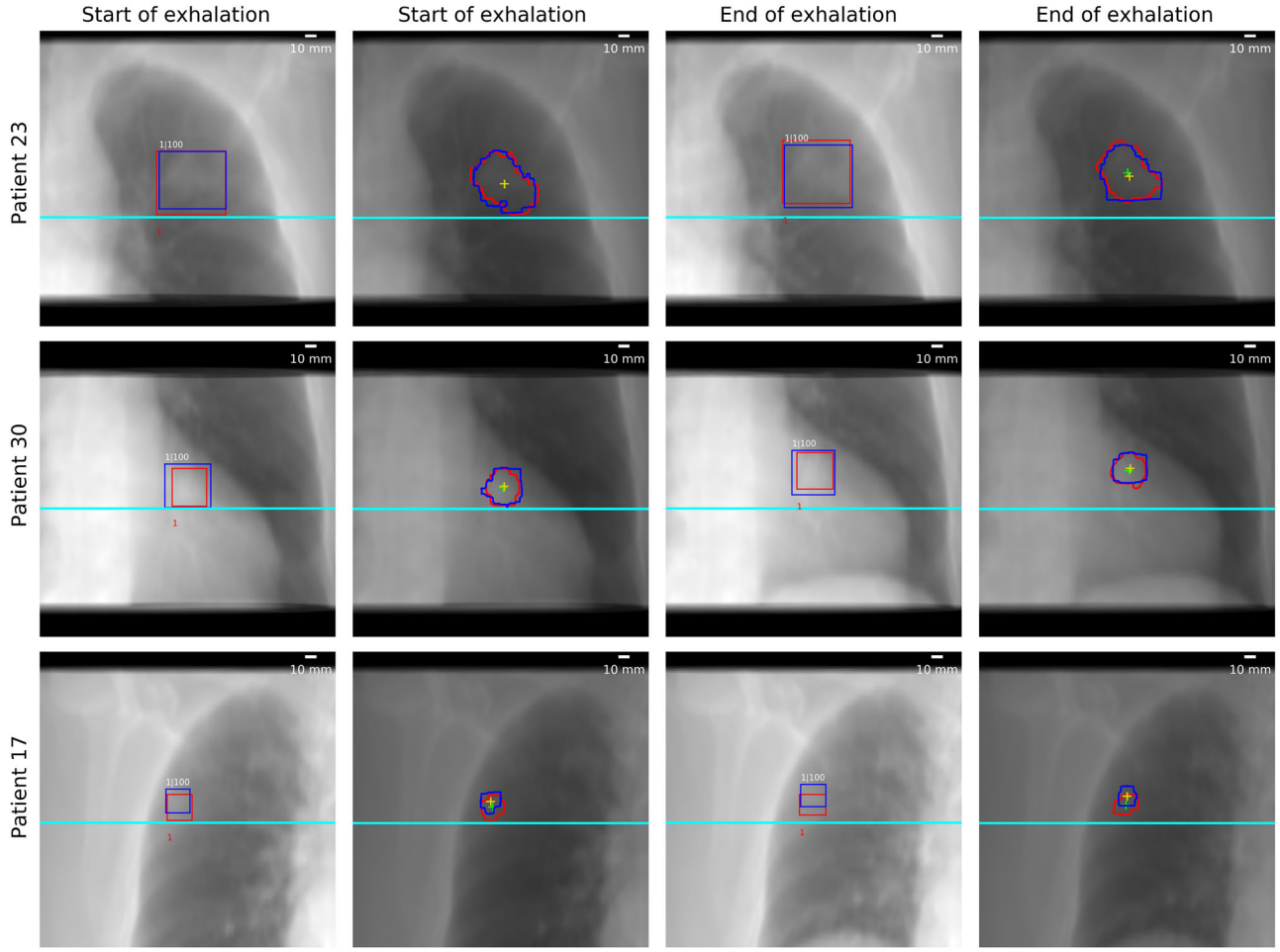

FIGURE 8 Examples of predictions on 4DCT-DRRs made by PS models. DRRs overlaid with bounding boxes (first and third columns) or segmentations (second and fourth columns) of patients 23, 30, and 17, respectively, showcasing good, medium, and bad cases, are plotted in top, middle, and bottom rows. Ground-truth segmentation (boxes) and COM are shown as red contours (boxes) and green crosses; predicted segmentation (boxes) and COM as blue contours (boxes) and yellow crosses. The first two and the last two columns respectively represent the start and end of exhalation phases of the 10-phase 4DCT-DRRs. To aid the illustration of tumor motion, cyan lines were drawn at the same location in each image of the same patient. All images shown in this figure have a size of 672 × 672 in pixels. 4DCT, four-dimensional computed tomography; COM, centers of mass; DRR, digitally reconstructed radiograph; PS, patient-specific.

near the start of exhalation, while its predicted and ground-truth segmentation contours, as well as their centers of mass, consistently followed each other across phases, regardless of the obscured tumor location by the heart. Patient 17 in the bottom row represents one of the worst cases, where predictions of both boxes and segmentations exhibited different degrees of overall upward shift across phases. Sizes of predicted boxes and shapes of predicted segmentation contours in general correspond to those of the ground truth. Full animation of all 10 phases of the three patients are available in the supplementary materials (see Figures B-4, B-5, B-6). Figures reporting the PS models' performance per patient for the other three evaluation metrics are also provided in the supplementary materials (see Figures B-1, B-2, B-3). In general, no correlation between the motion amplitude and the performance of any evaluation metric was found.

Figure 10 compares the performance of PS models and the template matching method for the whole group of test-set patients (top row) and for a subset, 37 in total, of the test-set patients that have been identified before by their PCT-DRR IoU value (>0.2). Over the 84 test-set patients, PS models had slightly better accuracy than the template matching method but the differences were not significant, with its median COB error at 2.7 mm compared to 3.0 mm for template matching and median of box DSC at 0.82 compared to 0.80. Over the 37 test-set patients, PS models showed significantly better ($p <$ 0.001) accuracy than the template matching method for both metrics, with its median COB error at 2.3 mm versus 3.0 mm for the template matching method, and its

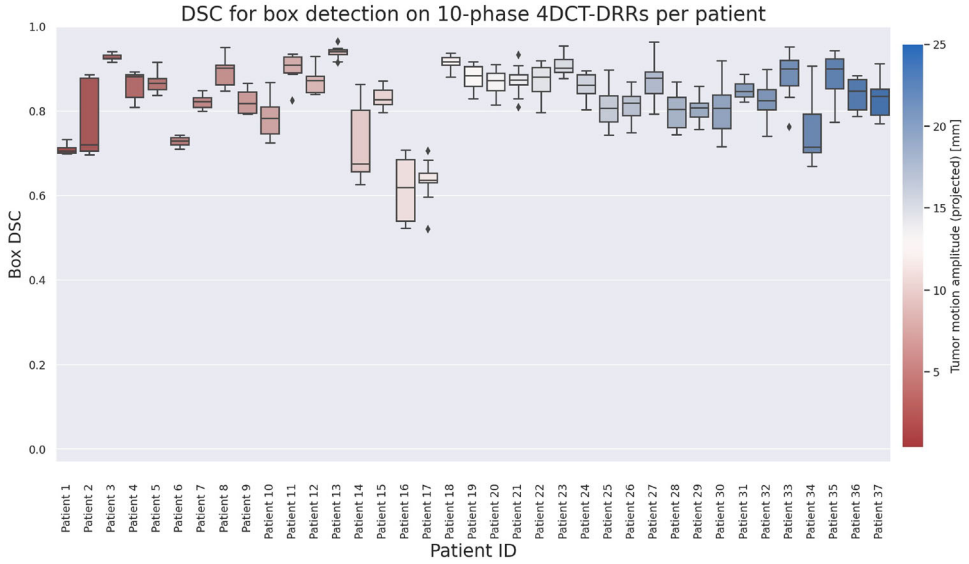

FIGURE 9 Box DSC of PS models evaluated for the 10 phases of 4DCT-DRRs per patient (for patients with PCT-DRR IoU > 0.2). Boxes are colored in terms of patients' motion amplitude in the coronal projection plane (calculated as the Euclidean distance formed of LR motion and SI motion). The patient index ranges over the number of cases which passed the PCT-DRR IoU threshold of 0.2. 4DCT, four-dimensional computed tomography; COB, center of box; COM, center of mass; DRR, digitally reconstructed radiograph; DSC, Dice similarity coefficient; IoU, intersection over union; PCT, planning CT LR, left-right; PS, patient-specific; SI, superior-inferior.

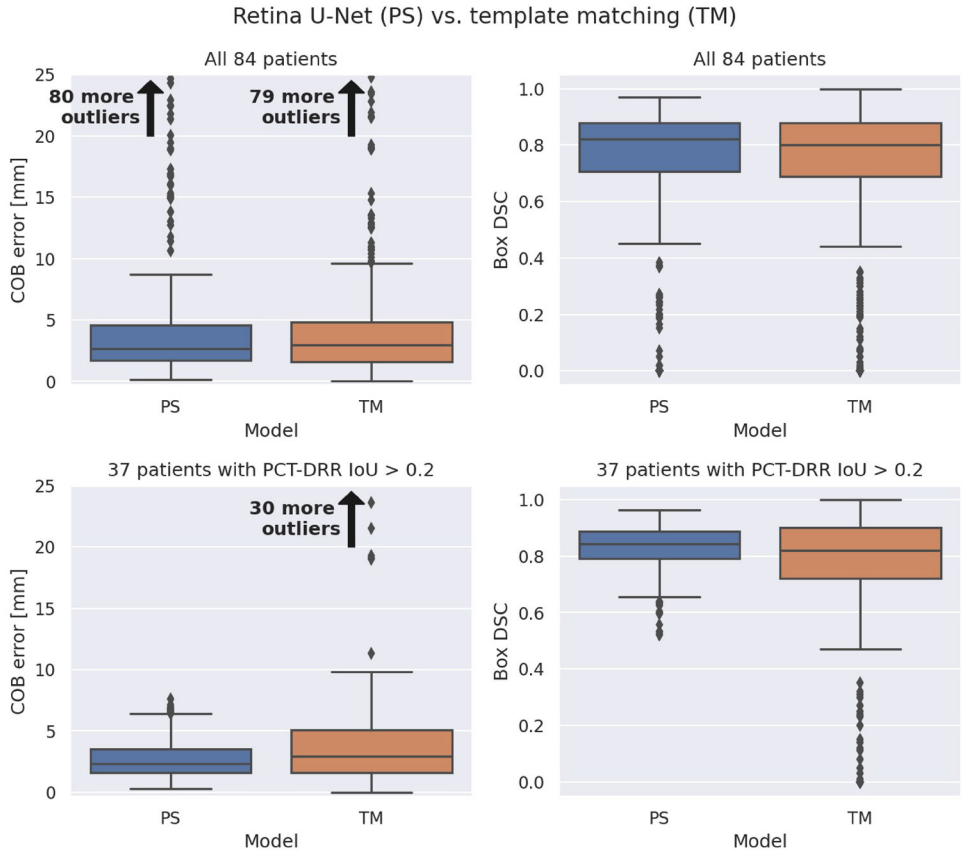

FIGURE 10 Comparison between testing results on 4DCT-DRRs of PS models (blue) and the TM method (orange). Top row: COB error (left) and box DSC (right) per phase and per patient for all 84 test-set patients. Bottom row: COB error (left) and box DSC (right) per phase and per patient for a subset of test-set patients, 37 in total, that had PCT-DRR IoU > 0.2. 4DCT, four-dimensional computed tomography; COB, center of box; DRR, digitally reconstructed radiograph; DSC, Dice similarity coefficient; IoU, intersection over union; PCT, planning CT; PS, patient-specific; TM, template matching.

median box DSC at 0.84 versus 0.82. PS models also achieved a better robustness on this subset than the template matching method, having a smaller spread of data points for both metrics.

4 | DISCUSSION

In this study, we implemented a markerless lung tumor tracking approach that performs simultaneous bounding box detection and target segmentation using retina U-Net, which was never tried before. We trained the model on public data serving as the baseline, then developed a patient-specific refinement procedure of this baseline model to check the applicability of the approach, and finally evaluated the performance of the resulting PS models.

According to the testing results, the PCT-DRR IoU threshold of 0.2 was an effective predictor of PS models' performance on patients' 4DCT-DRRs. It is worth pointing out that the PCT-DRR IoU threshold 0.2 and the matching IoU threshold 0.1 mentioned in section 2.4.3 have different meanings and were used differently. The PCT-DRR IoU value for a patient is an evaluation of the baseline model's performance on this specific patient, and 0.2 is the threshold where we observed consistent performance of PS models on 4DCT-DRRs. The matching IoU threshold, on the other hand, was used during training to determine whether a detection made by the model was true positive. A lower value of matching IoU could lead to a lower accuracy of the box coordinates (COB error) but higher average precision and detection accuracy. The value 0.1 not only was in agreement with Jaeger et al.,[31] but also was chosen partly because it was the best result of our fine tuning and partly because once the baseline managed to roughly locate the target in as many cases as possible, the subsequent patient-specific training could help boost the accuracy, rendering more cases trackable.

The PCT-DRR IoU threshold thus served as a good criterion for selecting patients suitable to track with the PS models. Quantitatively, for patients with PCT-DRR IoU higher than 0.2 and tumor volumes greater than 700 mm^3, PS models could achieve on average a segmentation COM error of 2.6 mm and a segmentation DSC of 0.78. To put the results into perspective, the median inter-observer DSC of multiple LIDC nodule annotations was 0.90, which represents the variation of the nodule annotations randomly selected as ground truth to feed into the baseline model and can be considered the upper limit of the model performance. Since we can identify and exclude patients having PCT-DRR IoU lower than 0.2 prior to treatment, applying the baseline model on patients' PCT-DRR provides certain confidence in the performance of PS models. 37 out of the 84 testing patients had a PCT-DRR IoU higher than 0.2. By adding the 13 validation patients with PCT-DRR IoUs higher than 0.7 chosen for hyperparameter fine tuning, it amounts to 50 patients in total, out of the cohort of 97 lung-cancer patients pre-selected based on their tumor sizes. This means our model was applicable to 52% of the patients with tumor volume within the volume range of LIDC nodules. This percentage is lower than the successful tumor visualization rate (66%) reported by Bahig et al.,[21] but they performed a much more rigorous patient pre-selection than a bare volume filtering, where every patient case was discussed and chosen by a professional group composed of physicians and physicists based on multiple aspects and staff experience. Considering the 37 test patients in the PCT-DRR IoU range above 0.2, patients were primarily concentrated in the middle range between 0.4 and 0.7. This uneven distribution was partly caused by the fact that prior to testing, we reserved 13 high PCT-DRR-IoU (> 0.7) patients for hyperparameters tuning and validation due to limited computational resources. Consequently, it is possible that insufficient data points in the high PCT-DRR IoU range skewed the conclusion on PS models' performance in this range. Additionally, statistics in the low IoU range [0, 0.2] also suggest that many patients (up to 68% or 32 cases) below the threshold of 0.2 could be salvaged in case they could be identified in advance, for example, when X-ray images would be available for testing so that patients' 4DCT-DRRs can be used to verify PS models' performance.

The evaluation results of PS models per IoU interval (Table 1) exhibited an overall similar performance between segmentation and box detection. During the post-processing of segmentation predictions, foreground pixels that were isolated from the predicted bounding box region were filtered out to force agreement between segmentation and box detection. The decision of outweighing the validity of the box detection over that of the segmentation throughout the study was based on the fact that the model architecture, Retina U-Net, was originally proposed for the task of detection and categorization rather than semantic segmentation. The addition of the auxiliary segmentation task was proposed by Jaeger et al.[31] to improve the detection and categorization performance. It is therefore expected that the box detection is slightly more robust than the segmentation over 4DCT phases, which is confirmed by the smaller average standard deviation for the box DSC (at 0.37) across the 37 patients (PCT-DRR IoU above 0.2) compared to that of the segmentation DSC (at 0.45). On the other hand, the readily available box detection can be considered as an alternative choice for tracking, especially when the segmentation fails, which happened once in one phase of a patient (Patient 32 in Figures 7, B-1, B-2, B-3) where the predicted segmentation mask had zero foreground pixel.

The benchmarking experiment against the template matching method revealed that when considering the broader, more generic patient groups without any prior patient differentiation, our approach did not significantly outperform the template matching method.

However, after restricting the patients to PCT-DRR IoU > 0.2, a significant performance gain in box detection with respect to the template matching method was observed. Furthermore, the unique advantage of our approach, which is the capability to identify applicable cases in advance of treatment, assures a higher level of robustness.

Inference-time wise, our PS models take approximately 170 ms per frame of size 672 × 672. Although slower than the above-mentioned studies,[24–26] it has the same order of magnitude as the system latency reported by a dynamic multi-leaf collimator (DMLC)-linac study[10] and several magnetic resonance imaging (MRI)-linac studies[36–38] and is thus still suitable for real-time markerless tracking, especially when considering the potential latency compensation enabled by existing fast target motion prediction algorithms.[28,39]

The good agreement between the AP curves of validation set and unmodified validation set supports that synthesizing LIDC data in the way we reported in section 2.4.2 was effective as a measure of data augmentation for training the baseline model. According to the comparison between PS models and the baseline model (Figure 7), the former ones outperformed the baseline model in all metrics. The baseline model exhibited an unstable performance as suggested by the large spread of its boxes (in blue) in the boxplots, whereas PS models showed great improvement both in accuracy and in robustness after being refined on the PCT-DRR. This signifies the importance of PS training as part of the training strategy. This performance gap, nevertheless, decreased considerably for the five patients with PCT-DRR IoU higher than 0.7 to the extent that the baseline model's performance became comparable to that of PS models in this high IoU range, though more samples are needed in order to draw a definitive conclusion.

One inevitable source of errors in our investigation was the uncertainty of deformable image registration (DIR), which was used for obtaining the GT segmentation masks for the 10 phases of the 4DCT based on the available PCT segmentation. The quality of the obtained GT segmentation masks could directly impact the evaluation of the model's segmentation ability, including the COM error and segmentation DSC. Moreover, the GT segmentations of the LIDC CT scans used for building the baseline model and the PCT scans collected for the study of PS models were drawn by multiple physicians. The diverse GT annotation source implies that the inter-physician variability might also be a factor contributing to the model uncertainty. Additionally, all CT scans used for training and testing PS models had a slice thickness of 3 mm in the SI direction. This resulted in the coarse resolution in motion sampling for DRRs, hence the blurry image quality, and inherently limited the accuracy of PS models, especially because PS models derived their predictions from anterior-posterior DRRs and the major component of tumor motion was mostly SI motion.

Image quality-wise, two aspects should be taken into consideration. (1) DRRs are only a simulation of the X-ray images, and the image quality difference between them caused by beam hardening, scattering and image noise cannot be ignored. Though the proposed approach has been extensively tested on DRRs, how well it performs on actual X-ray images remains to be investigated, which is considered one limitation of this study. An evaluation on kV X-ray images should be performed in future studies. (2) Factors like the target-to-background contrast and the target occlusion caused by other anatomical structures such as heart, mediastinum and diaphragm, could impact the model performance. Moreover, the projection angle of DRRs used in this study was limited to 0°, and other projection angles arising from fixed or gantry-mounted X-ray imaging devices could also impact the target visibility. Analysis of baseline model performance showed that among the 47 failed patients (with PCT-DRR IoU below 0.2), 51% had occlusion, 68% had low contrasts (below 5.0), and 83% had at least one of the two conditions. Still, 9 occlusion-bearing patients and 18 low-contrast patients had PCT-DRR IoUs higher than 0.2, making up 24% and 49% of the 37 successful cases, respectively. Further examining the 37 successful cases, PS models performed better on high contrast (above 5.0) patients than on low contrast patients, and performed comparably for patients with and without tumor occlusion. Overall, it can be concluded that the baseline model worked best on occlusion-free and high-contrast cases, and that although the PS models tend to perform better on high-contrast cases than others, they were efficient for all cases despite low contrast and occlusion once the PCT-DRR IoU had passed 0.2. This confirmed again that the baseline model's PCT-DRR IoU was a good index to estimate the tracking effectiveness of our approach for a specific patient.

In terms of train/test division, our study differs from other studies[22,25,26] in that we did not use patients' 4DCT data in any form during training. Zhao et al.[22,23] and Zhou et al.[26] interpolated frames between consecutive respiratory phases, respectively, by using linearly interpolated motion vector fields and by using a deep learning-based interpolation model, and then randomly divided the generated DRRs into train and test sets. Zhao et al. reached a mean absolute difference (MAD) of 1.58 mm jointly averaged over horizontal and vertical directions for DRRs projected with 0° X-ray tube angle over 10 prostate-cancer patients.[23] This accuracy is roughly equivalent to 2.27 mm in mean 2D error, which is approximated as the Euclidean distance calculated from the average prediction errors in the two dimensions. Another work by Zhao et al. applied the same method on two pancreatic cases and all MADs in both directions were less than 2.60 mm for DRRs at 0° X-ray tube angle,[22] which amounts to 2.52 mm on average for 2D error. The mean segmentation COM error of our study, 2.6 mm, is comparable to those of

these two studies. The best accuracy of all retrospective patient studies was reported by Zhou et al.,[26] with a mean 3D error of 0.29 mm and a mean segmentation DSC of 0.98 over 14 pancreatic cases and 12 different X-ray tube angles. However, it is important to note that the random train/test partition scheme of the densely interpolated 4DCT-DRR data can diminish the independence between train and test sets, the degree of which depends on the interpolation density, because information contained in the test set would have been too closely approximated by information already seen by the model in the training set to truly test the model's generalizing capability. To avoid this drawback, we chose to reserve the single 4DCT scan of each patient for the testing of their PS model. Only the static PCT was used for refining the baseline model, which was trained on population-based data, into a tailored PS model. Nevertheless, the 3D PCT was separately acquired immediately after the 4DCT acquisition. The temporal proximity entails a high anatomical similarity between PCT and 4DCT, which may not represent well the potential variations between a pre-treatment scan and the treatment session.

Distinct from other studies, Sakata et al. trained on all 10 phases of the 4DCT-DRRs and tested on fluoroscopic images acquired during treatment delivery.[25] This train/test division scheme is free of interpolation dependency, and is representative of clinical tracking scenarios. An accuracy of 1.03 mm in mean 3D error (Euclidean distance of COM) was achieved over 8 lung cancer patients. They manually annotated the COM of tumors on all fluoroscopic images as the ground truth, and thus did not provide contour evaluation such as DSC calculation. Their superior COM accuracy compared to ours may be explained by their higher image quality: (1) The 4DCT-DRRs that their model was trained on were derived from 4DCT with a voxel size of 1 mm in SI direction, compared to the 3 mm of our PCT and 4DCT data. (2) They tested on fluoroscopic images which intrinsically had a higher resolution than DRRs simulated from CT data. Furthermore, our model was not aware of the motion pattern of the tumor in question due to the separation of 4DCT-DRRs from training data, while their model was able to fully exploit the 4DCT to actively learn the specific motion pattern of the patient. It is speculated that the performance of our model may be improved by including 4DCT-DRRs as training data, provided that subsequent X-ray images can be acquired for testing. We plan to conduct experiments in the future to explore the gain of training on motion patterns contained in 4DCT-DRRs.

Besides DL-based approaches, numerous other studies have investigated non-DL-based image tracking techniques for markerless lung tumor tracking, including well-established methods such as image registration,[40] template matching,[41] and optical flow[42,43]. Other less common methods like short arc tumor tracking[44] and hidden Markov model[45] have also been proposed. Some studies were exclusively conducted on digital or experimental phantoms with tumor motions that were mechanically controlled following either simulated breathing patterns[42] or patient-measured motion traces.[45,46] They were often able to achieve sub-millimeter accuracy, which might potentially be biased by the often simpler geometry of phantom anatomy and tumor shape/size. Several other studies like Rozario et al,[40] Bruin et al.,[41] Shieh et al.[44] and Ichiji et al.[43] included retrospective studies on real patients' data such as beam's eye view (BEV) images, CBCT projection images, or clinical x-ray image sequences. Rozario et al.[40] tested their image registration-based method on over 5000 frames of MV BEV images of 5 patients and reported rather unstable performance, with tumors' average 2D position deviations at 180 degrees gantry angle for a single fraction ranging from 4.6 mm up to 7.9 mm (6 fractions from 3 patients). Shieh et al.[44] specifically targeted challenging tracking cases by validating on 4DCBCT projection images of 4 patients with central tumors attached to the mediastinum and with very low contrast of tumors in the projection images. Despite the low visibility of tumor in projection images, they were able to track the tumors at all gantry angles in all 11 CBCT scans and achieved a mean 3D tracking error ranging from 2.2-9.9 mm. The time resolution of their approach is however limited by the gantry speed, ranging from 1.5–9 s and leaving room for improvement in order to enable real-time tracking. Bruin et al.[41] used a template matching approach and tested in a qualitative manner due to the lack of ground-truth data on the CBCT projections of 18 patients with 20 tumors in total. 65% tumors were deemed successfully tracked by judging the correspondence of the predicted longitudinal tumor trajectories manually overlaid with the motion of an external Real-time Position Management (RPM) marker. Ichiji et al.[43] proposed a new method that used the optical flow technique to track key points on the tumor, and tested on clinical X-ray image sequences. They achieved an average root mean square error of 2.46 mm for kV X-ray images and 1.53 mm for MV X-ray images, and discovered that the accuracy of their method linearly decreased with the motion range of the target tumor, which has not been observed in our approach. Compared to these non DL-based studies, we evaluated our approach on a larger set (84 testing patients) of patients' data covering a wider range of tumor visibilities, despite using the non-ideal DRRs, and achieved an average 2D tracking error of 2.6 mm (mean COM error) over 37 patients. The accuracy of our approach was independent of the tumor motion amplitude. More importantly, our approach allowed the systematic identification of suitable patients prior to treatment via applying the pre-built baseline model on patients' PCT-DRR to obtain the PCT-DRR IoU value, which was proven to be a strong predictor of PS models' performance.

5 | CONCLUSIONS

We trained a deep learning-based model for markerless lung tumor tracking that is able to perform simultaneous bounding box detection and tumor segmentation. We implemented a novel patient-specific refinement procedure exploiting a pre-trained baseline model and leveraging clinically available pre-treatment PCT data of patients. The patient-specific models achieved an accuracy of 2.6 mm in mean segmentation COM error and 0.78 in mean segmentation DSC at an inference time of approximately 170 ms per frame, rendering the method suitable for real-time tumor tracking. The proposed approach was consistently applicable to 52% of the patients with tumor volumes within the volume range of LIDC nodules.

ACKNOWLEDGMENTS

This work was supported by the funding "Förderprogramm für Forschung und Lehre" issued by the Medical Faculty of Ludwig-Maximilians-Universität München (reg. no. 1113).

CONFLICT OF INTEREST STATEMENT

Dr P. Freislederer is now employed by Brainlab AG. The Department of Radiation Oncology of the LMU University Hospital has a research agreement with Brainlab. Brainlab was not involved and had no influence on this study.

DATA AVAILABILITY STATEMENT

Training data for the baseline model, the LIDC dataset and its synthesized augmented version, can be shared upon reasonable requests. The LMU hospital dataset will not be available due to missing ethics approval for public sharing. Code for the modified version of the Retina U-Net architecture can be shared upon requests.

REFERENCES

1. Ford EC, Mageras GS, Yorke E, Ling CC. Respiration-correlated spiral CT: a method of measuring respiratory-induced anatomic motion for radiation treatment planning. *Med Phys*. 2003;30:88-97.
2. Vedam SS, Keall PJ, Kini VR, Mostafavi H, Shukla HP, Mohan R. Acquiring a four-dimensional computed tomography dataset using an external respiratory signal. *Phys Med Biol*. 2003;48:45-62.
3. Ohara K, Okumura T, Akisada M, et al. Irradiation synchronized with respiration gate. *Int J Radiat Oncol Biol Phys*. 1989;17:853-857.
4. Kubo HD, Wang L. Compatibility of Varian 2100C gated operations with enhanced dynamic wedge and IMRT dose delivery. *Med Phys*. 2000;27:1732-1738.
5. Stromberg JS, Sharpe MB, Kim LH, et al. Active breathing control (ABC) for Hodgkin's disease: reduction in normal tissue irradiation with deep inspiration and implications for treatment. *Int J Radiat Oncol Biol Phys*. 2000;48:797-806.
6. Rosenzweig KE, Hanley J, Mah D, et al. The deep inspiration breath-hold technique in the treatment of inoperable non-small-cell lung cancer. *Int J Radiat Oncol Biol Phys*. 2000;48:81-87.
7. Lax I, Blomgren H, Näslund I, Svanström R. Stereotactic radiotherapy of malignancies in the abdomen. Methodological aspects. *Acta Oncologica (Stockholm, Sweden)*. 1994;33:677-683.
8. Lin L, Souris K, Kang M, et al. Evaluation of motion mitigation using abdominal compression in the clinical implementation of pencil beam scanning proton therapy of liver tumors. *Med Phys*. 2017;44:703-712.
9. Keall PJ, Kini VR, Vedam SS, Mohan R. Motion adaptive x-ray therapy: A feasibility study. *Phys Med Biol*. 2000;46:1-10.
10. Rottmann J, Keall P, Berbeco R. Markerless EPID image guided dynamic multi-leaf collimator tracking for lung tumors. *Phys Med Biol*. 2013;58:4195-4204.
11. Shieh CC, Caillet V, Dunbar M, et al. A Bayesian approach for three-dimensional markerless tumor tracking using kV imaging during lung radiotherapy. *Phys Med Biol*. 2017;62:3065-3080.
12. Zhang X, Homma N, Ichiji K, Takai Y, Yoshizawa M. Tracking tumor boundary in MV-EPID images without implanted markers: a feasibility study. *Med Phys*. 2015;42:2510-2523.
13. Campbell WG, Miften M, Jones BL. Automated target tracking in kilovoltage images using dynamic templates of fiducial marker clusters. *Med Phys*. 2017;44:364-374.
14. Shirato H, Shimizu S, Kunieda T, et al. Physical aspects of a real-time tumor-tracking system for gated radiotherapy. *Int J Radiat Oncol Biol Phys*. 2000;48:1187-1195.
15. Shirato H, Shimizu S, Shimizu T, Nishioka T, Miyasaka K. Real-time tumour-tracking radiotherapy. *Lancet North Am Ed*. 1999;353:1331-1332.
16. Khashab MA, Kim KJ, Tryggestad EJ, et al. Comparative analysis of traditional and coiled fiducials implanted during EUS for pancreatic cancer patients receiving stereotactic body radiation therapy. *Gastrointest Endosc*. 2012;76:962-971.
17. van der Horst A, Wognum S, Dávila Fajardo R, et al. Interfractional position variation of pancreatic tumors quantified using intratumoral fiducial markers and daily cone beam computed tomography. *Int J Radiat Oncol Biol Phys*. 2013;87:202-208.
18. Bhagat N, Fidelman N, Durack JC, et al. Complications associated with the percutaneous insertion of fiducial markers in the thorax. *Cardiovasc Intervent Radiol*. 2010;33:1186-1191.
19. Teske H, Mercea P, Schwarz M, Nicolay NH, Sterzing F, Bendl R. Real-time markerless lung tumor tracking in fluoroscopic video: handling overlapping of projected structures. *Med Phys*. 2015;42:2540-2549.
20. Shieh CC, Caillet V, Dunbar M, et al. A Bayesian approach for three-dimensional markerless tumor tracking using kV imaging during lung radiotherapy. *Phys Med Biol*. 2017;62:3065-3080.
21. Bahig H, Campeau MP, Vu T, et al. Predictive parameters of cyberknife fiducial-less (xsight lung) applicability for treatment of early non-small cell lung cancer: a single-center experience. *Int J Radiat Oncol Biol Phys*. 2013;87:583-589.
22. Zhao W, Shen L, Han B, et al. Markerless pancreatic tumor target localization enabled by deep learning. *Int J Radiat Oncol Biol Phys*. 2019;105:432-439.
23. Zhao W, Han B, Yang Y, et al. Incorporating imaging information from deep neural network layers into image guided radiation therapy (IGRT). *Radiother Oncol*. 2019;140:167-174.
24. Takahashi W, Oshikawa S, Mori S. Real-time markerless tumour tracking with patient-specific deep learning using a personalised data generation strategy: proof of concept by phantom study. *Br J Radiol*. 2020;93:20190420.
25. Sakata Y, Hirai R, Kobuna K, Tanizawa A, Mori S. A machine learning-based real-time tumor tracking system for fluoroscopic gating of lung radiotherapy. *Phys Med Biol*. 2020;65:085 014.
26. Zhou D, Nakamura M, Mukumoto N, Yoshimura M, Mizowaki T. Development of a deep learning-based patient-specific target contour prediction model for markerless tumor positioning. *Med Phys*. 2022;49:1382-1390.

27. Shen C, Nguyen D, Zhou Z, Jiang SB, Dong B, Jia X. An introduction to deep learning in medical physics: advantages, potential, and challenges. *Phys Med Biol*. 2020;65:05TR01.
28. Teo TP, Ahmed SB, Kawalec P, et al. Feasibility of predicting tumor motion using online data acquired during treatment and a generalized neural network optimized with offline patient tumor trajectories. *Med Phys*. 2018;45:830-845.
29. Fedorov A, Hancock M, Clunie D, et al. DICOM re-encoding of volumetrically annotated lung imaging database consortium (LIDC) nodules. *Med Phys*. 2020;47:5953-5965.
30. Rit S, Oliva MV, Brousmiche S, Labarbe R, Sarrut D, Sharp GC. The reconstruction toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the insight toolkit (ITK). *J Phys Conf Ser*. 2014;489:012 079.
31. Jaeger PF, Kohl SAA, Bickelhaupt S, et al. Retina U-Net: embarrassingly simple exploitation of segmentation supervision for medical object detection. *Proceedings of the Machine Learning for Health NeurIPS Workshop*. PMLR; 2020:171-183.
32. Schultheiss M, Schmette P, Bodden J, et al. Lung nodule detection in chest X-rays using synthetic ground-truth data comparing CNN-based diagnosis to human performance. *Sci Rep*. 2021;11:15 857.
33. Isensee F, Jäger P, Wasserthal J, et al. Batchgenerators - a python framework for data augmentation. *Zenodo*. 2020. https://doi.org/10.5281/zenodo.3632567
34. Simard P, Steinkraus D, Platt J. Best practices for convolutional neural networks applied to visual document analysis. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings*. Edinburgh, UK, 2003:958-963. https://doi.org/10.1109/ICDAR.2003.1227801
35. van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. Scikit-image: image processing in Python. *PeerJ*. 2014;2:e453.
36. Glitzner M, Woodhead PL, Borman PTS, Lagendijk JJW, Raaymakers BW. Technical note: MLC-tracking performance on the Elekta unity MRI-linac. *Phys Med Biol*. 2019;64:15NT02.
37. Liu PZY, Dong B, Nguyen DT, et al. First experimental investigation of simultaneously tracking two independently moving targets on an MRI-linac using real-time MRI and MLC tracking. *Med Phys*. 2020;47:6440-6449.
38. Yun J, Wachowicz K, Mackenzie M, Rathee S, Robinson D, Fallone BG. First demonstration of intrafractional tumor-tracked irradiation using 2D phantom MR images on a prototype linac-MR. *Med Phys*. 2013;40:051 718.
39. Lombardo E, Rabe M, Xiong Y, et al. Offline and online LSTM networks for respiratory motion prediction in MR-guided radiotherapy. *Phys Med Biol*. 2022;67(9):095006.
40. Rozario T, Chiu TD, Chen M, et al. A novel markerless lung tumor-tracking method using treatment MV beam imaging. *Appl Sci*. 2018;8:2525.
41. de Bruin K, Dahele M, Mostafavi H, Slotman BJ, Verbakel WFAR. Markerless real-time 3-dimensional kV tracking of lung tumors during free breathing stereotactic radiation therapy. *Adv Radiat Oncol*. 2021;6(4):100705.
42. Teo PT, Crow R, Nest SV, Sasaki D, Pistorius S. Tracking lung tumour motion using a dynamically weighted optical flow algorithm and electronic portal imaging device. *Meas Sci Technol*. 2013;24:074 012.
43. Ichiji K, Yoshida Y, Homma N, et al. A key-point based real-time tracking of lung tumor in x-ray image sequence by using difference of Gaussians filtering and optical flow. *Phys Med Biol*. 2018;63:185 007.
44. Shieh CC, Keall PJ, Kuncic Z, Huang CY, Feain I. Markerless tumor tracking using short kilovoltage imaging arcs for lung image-guided radiotherapy. *Phys Med Biol*. 2015;60:9437.
45. Shinohara T, Ichiji K, Wang J, et al. Improved tumor image estimation in X-ray fluoroscopic images by augmenting 4DCT data for radiotherapy. *J Adv Comput Intell Intell Informat*. 2022;26:471-482.
46. Mueller M, Poulsen P, Hansen R, et al. The markerless lung target tracking AAPM grand challenge (MATCH) results. *Med Phys*. 2022;49:1161-1180.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Huang L, Kurz C, Freislederer P, et al. Simultaneous object detection and segmentation for patient-specific markerless lung tumor tracking in simulated radiographs with deep learning. *Med Phys*. 2023;1-17. https://doi.org/10.1002/mp.16705