IPEM
Institute of Physics and
Engineering in Medicine

PAPER • OPEN ACCESS

Deep learning based direct segmentation assisted by deformable image registration for cone-beam CT based auto-segmentation for adaptive radiotherapy

View the article online for updates and enhancements.

You may also like

Physics in Medicine & Biology

IPEM
Institute of Physics and
Engineering in Medicine

CrossMark

PAPER

Deep learning based direct segmentation assisted by deformable image registration for cone-beam CT based auto-segmentation for adaptive radiotherapy

Xiao Liang [ORCID] , Howard Morgan, Ti Bai [ORCID] , Michael Dohopolski, Dan Nguyen [ORCID] and Steve Jiang*

Medical Artificial Intelligence and Automation Laboratory and Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, United States of America
* Author to whom any correspondence should be addressed.

E-mail: steve.jiang@utsouthwestern.edu

Abstract

Cone-beam CT (CBCT)-based online adaptive radiotherapy calls for accurate auto-segmentation to reduce the time cost for physicians. However, deep learning (DL)-based direct segmentation of CBCT images is a challenging task, mainly due to the poor image quality and lack of well-labelled large training datasets. Deformable image registration (DIR) is often used to propagate the manual contours on the planning CT (pCT) of the same patient to CBCT. In this work, we undertake solving the problems mentioned above with the assistance of DIR. Our method consists of three main components. First, we use deformed pCT contours derived from multiple DIR methods between pCT and CBCT as pseudo labels for initial training of the DL-based direct segmentation model. Second, we use deformed pCT contours from another DIR algorithm as influencer volumes to define the region of interest for DL-based direct segmentation. Third, the initially trained DL model is further fine-tuned using a smaller set of true labels. Nine patients are used for model evaluation. We found that DL-based direct segmentation on CBCT without influencer volumes has much poorer performance compared to DIR-based segmentation. However, adding deformed pCT contours as influencer volumes in the direct segmentation network dramatically improves segmentation performance, reaching the accuracy level of DIR-based segmentation. The DL model with influencer volumes can be further improved through fine-tuning using a smaller set of true labels, achieving mean Dice similarity coefficient of 0.86, Hausdorff distance at the 95th percentile of 2.34 mm, and average surface distance of 0.56 mm. A DL-based direct CBCT segmentation model can be improved to outperform DIR-based segmentation models by using deformed pCT contours as pseudo labels and influencer volumes for initial training, and by using a smaller set of true labels for model fine tuning.

1. Introduction

Online adaptive radiotherapy (ART) is an advanced radiotherapy technology in which the daily treatment plan is adapted to the patient's changing anatomy (e.g. shrinking tumor, losing body weight), typically based on cone beam computed tomography (CBCT) images. The online nature of the treatment demands high efficiency since the patient is immobilized while waiting for treatment to start. The time-consuming process of segmenting the tumor volumes and organs at risk (OARs) has become a major bottleneck for the widespread use of online ART, warranting an urgent need for accurate auto-segmentation tools (Glide-Hurst *et al* 2021).

Auto-segmentation of CBCT images is a very challenging task, mainly due to poor image quality and lack of training labels for deep learning (DL)-based methods. First, the greater presence of noise and artifacts on CBCT images, such as capping, cupping, ring, and streaking artifacts, makes CBCT more difficult than CT for auto-segmentation tasks (Lechuga and Weidlich 2016). Second, contouring of tumor volumes and OARs is not part of

the common applications of CBCT-based image-guided radiotherapy. Therefore, unlike CT for treatment planning, one cannot use the clinical contours generated from routine clinical practice to train DL models for CBCT segmentation. Expert clinicians must retrospectively contour large sets of CBCT images specifically for CBCT segmentation research, which is time consuming and challenging. Due to two major limitations, poor image quality and lack of a well-labeled large set of training data, studies have shown that DL-based direct segmentation on CBCT images produce poor segmentation results (Beekman *et al* 2021, Léger *et al* 2020, Alam *et al* 2021, Dahiya *et al* 2021, Dai *et al* 2021).

Auto-segmentation on CBCT for ART is a unique task when planning CT (pCT) with manual contours are available. Using pCT with manual contours as prior knowledge, some studies have shown that DL-based direct segmentation could achieve improved results. A simple way to take advantage of pCT and its contours is to directly mix them into a limited CBCT training dataset. This cross-domain augmentation of the training set was effective for CBCT augmentation (Léger *et al* 2020). With a more complicated data augmentation strategy, one study generated variant synthetic CBCT (sCBCT) images with one pair of pCT and CBCT of the same patient where the generated data was used to train a CBCT segmentation model (Dahiya *et al* 2021). Similarly, some other studies utilized artifact induction to convert pCT to sCBCT to make use of high quality CT manual contours for CBCT segmentation (Schreier *et al* 2020, Alam *et al* 2021).

All the studies mentioned above use data augmentation methods to mitigate the lack of training labels on CBCT, either by adding pCT with manual contours into the training set directly or by generating sCBCT with contour labels from pCT with manual contours. While auto-segmentation results can be improved to some degree by data augmentation, the biggest drawback of these methods is that neither CT nor sCBCT can truly represent a real CBCT image. Therefore, a more robust and popular way to utilize pCT and its high quality contours for CBCT auto-segmentation is through deformable image registration (DIR) methods. By deforming pCT to CBCT, pCT contours can be propagated to CBCT. While traditional DIR methods including B-spline and Demons algorithms (Gu *et al* 2009, Klein *et al* 2010, Fedorov *et al* 2012) are computationally intensive and time consuming, DL-based DIR methods can be quick at inference, but usually require large amounts of training data (Han *et al* 2021), which can be mitigated with methods like test-time optimization (TTO) (Liang *et al* 2022). The biggest benefit of DIR-based segmentation is that the topological consistency of contours can be preserved by smoothly deforming pCT contours, since most organs have small anatomical changes. However, when anatomical changes are big, DIR-based contour propagation can be biased towards the original pCT contour shape.

A combination of DL-based direct segmentation and DIR-based contour propagation can potentially leverage the advantages of two methods. One way is to firstly use a DL-based model to direct segment easier OARs in CBCT images, then subsequently use the segmentation results to constrain the DIR between pCT and CBCT, and finally to propagate the target volumes and rest of OARs from pCT to CBCT (Archambault *et al* 2020). Another way is through joint learning of segmentation and registration. A typical way for joint segmentation and registration is to predict segmentations on both moving and fixed images using unsegmented moving and fix images as inputs (Estienne *et al* 2019, Xu and Niethammer 2019). Another way for joint learning is to use the moving and fixed image, as well as moving segmentations as inputs to predict segmentations on the fixed image (Beekman *et al* 2021). In those approaches for joint learning, DL-based direct segmentation and DIR-based contour propagation are combined either through parameter sharing between the segmentation and registration model, or through a joint loss function. However, the above-mentioned joint learning approaches cannot significantly outperform DIR-based contour propagation for CBCT segmentation, and still requires a large amount of segmentation labels on fix (CBCT) images for model training (Beekman *et al* 2021).

In this paper, we explore a new method to improve DL-based direct segmentation with the assistance of DIR results, aiming to outperform the DIR-based methods for CBCT segmentation, without requiring a large well-labeled training dataset. We propose to use pseudo labels for initial model training, where the pseudo labels are deformed pCT contours. To help localize OARs and target, we propose to add deformed pCT contours as influencer volumes through additional channels of the segmentation model. We then fine-tune the initially trained model using a small set of training data with true labels.

2. Methods

2.1. Problem definition

In a fully supervised segmentation task, we can denote the training set as $\mathscr{D} = \{(X, Y)\}^D$, where $X \in \mathbb{R}^\Omega$ denotes training images and $Y \in \{0, 1\}^\Omega$ denotes their corresponding pixel-wise labels. Ω denotes its corresponding spatial domain. Given the labeled dataset D, the segmentation task intends to learn a function F with parameter θ to map X to Y by minimizing the standard Dice loss:

$$\min_\theta \mathcal{L}_{\text{Dice}}(\theta) := 1 - \frac{\left[2\sum_{p\in\Omega}y(p)s_\theta(p)\right] + c}{\sum_{p\in\Omega}y(p) + \sum_{p\in\Omega}s_\theta(p) + c}, \tag{1}$$

where $s_\theta = F(x|\theta)$ is the predicted probabilities by the CNNs, and c is a small constant added to prevent dividing by 0. $s_\theta \in [0, 1]^\Omega$, with 0 and 1 denoting background and foreground.

2.2. Pseudo label learning (Model$_{\text{pseudo}}$)

We generated pseudo labels for initial training of the model and firstly deformed pCT to its paired CBCT to get a deformation vector field (DVF). We then used DVF to warp pCT's contours to generate deformed contours, which were the pseudo labels of CBCT. To address the pseudo label noise, we applied multiple DIR algorithms to generate multiple sets of pseudo labels. By randomly selecting one type of pseudo label during each training iteration, we could mitigate random errors coming from DIR algorithms. We applied three different DL-based DIR algorithms: FAIM (Kuang and Schmah 2019), 5-cascaded Voxelmorph (Dalca *et al* 2019), and 10-cascaded VTN (Zhao *et al* 2019) to generate pseudo labels y_1, y_2, and y_3, respectively. Given the dataset $\mathscr{D} = \{(X, Y)\}^D$, where $X = \{x\}$ represents image and $Y = \{y_i\}$ represents pseudo labels, the segmentation model intends to learn a function F with parameter θ_{pl} formulated similarly to equation (1):

$$\min_{\theta_{pl}} \mathcal{L}_{\text{Dice}}(\theta_{pl}) := 1 - \sum_i \frac{\left[2\sum_{p\in\Omega}y_i(p)s_{\theta_{pl}}(p)\right] + c}{\sum_{p\in\Omega}y_i(p) + \sum_{p\in\Omega}s_{\theta_{pl}}(p) + c}, \tag{2}$$

where $s_{\theta_{pl}} = F(x|\theta_{pl})$ is the model prediction and $i \in \{1, 2, 3\}$.

2.3. Influencer volumes (Model$_{\text{influencer}}$)

To deal with the low image quality problem for direct CBCT segmentation, we proposed to add influencer volumes as additional channels of input, as shown in figure 1. The architecture used in this experiment was a typical U-Net architecture. Besides the CBCT image, deformed pCT contours were used as additional input channels to constrain the region of interest for segmentation. The influencer volumes were used for shape and location feature extraction. The shape and location features were independently extracted from the influencer volumes for each level and combined with features extracted from CBCT images by multiplication. The combined features were further concatenated with features from up-sampling layers. The output or labels consisted of multi-organ segmentation masks derived from learned relations between the CBCT images and influencer volumes. It is one network predicting all 19 structures.

Since deformed pCT contours were also used as pseudo labels for training, to avoid using the same deformed pCT contours as input and output at the same time, we proposed to assign two different deformed pCT contours as such by randomly picking two DIR methods to generate different deformed contours during each training iteration. In this case, the input was $X = \{x, y_j\}$ and label was $Y = \{y_i\}$, where $i, j \in \{1, 2, 3\}$ and $i \neq j$. We could formulate the loss function in a similar style to equations (1) or (2):

$$\min_{\theta_{iv}} \mathcal{L}_{\text{Dice}}(\theta_{iv}) := 1 - \sum_i \frac{\left[2\sum_{p\in\Omega}y_i(p)s_{\theta_{iv}}(p)\right] + c}{\sum_{p\in\Omega}y_i(p) + \sum_{p\in\Omega}s_{\theta_{iv}}(p) + c}, \tag{3}$$
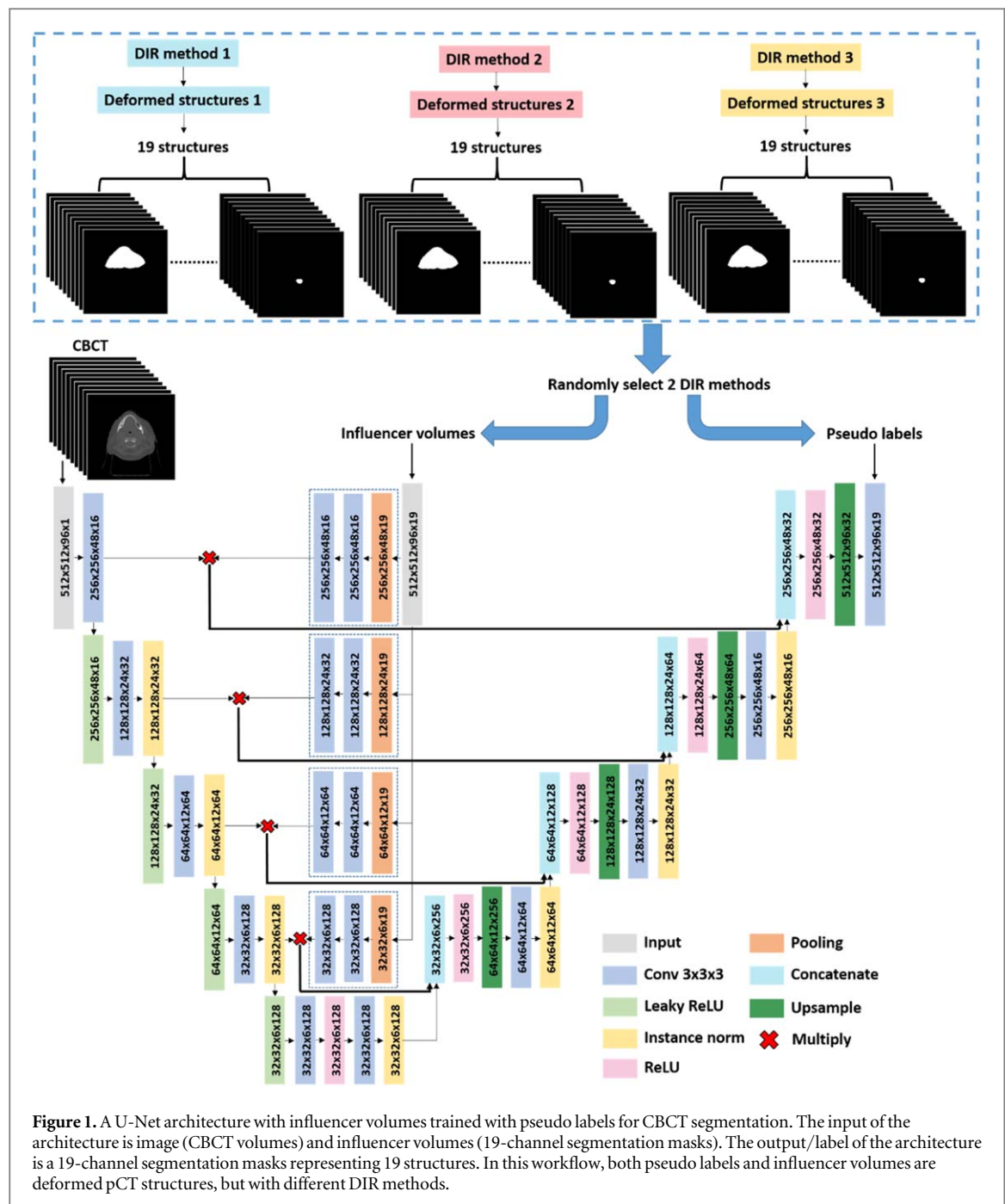
where $s_{\theta_{iv}} = F(x, y_j|\theta_{iv})$ is the model prediction with image x and deformed pCT segmentation y_j as input.

2.4. Fine-tuning with true labels (Model$_{\text{finetune}}$)

To further improve segmentation performance, we propose to fine-tune Model$_{\text{influencer}}$ using a small training dataset with true labels, to mitigate DIR errors that propagate through pseudo labels. And the workflow is shown in figure 2.

2.5. Data

We retrospectively collected data from 137 patients with head and neck (H&N) squamous cell carcinoma treated with conventionally fractionated external beam radiotherapy. Each patient's data included a 3D pCT volume acquired before the treatment course, OARs and target segmentations delineated and approved by radiation oncologists on the pCT, and a 3D CBCT image. The pCT volumes were acquired by a Philips CT scanner (Philips Healthcare, Best, Netherlands) with $1.17 \times 1.17 \times 3.00$ mm^3 voxel spacing. The CBCT volumes were acquired by Varian On-Board Imagers (Varian Medical Systems Inc., Palo Alto, CA, USA) with $0.51 \times 0.51 \times 1.99$ mm^3 voxel spacing and $512 \times 512 \times 93$ dimensions. Among those 137 patients, 39 patients had true segmentation labels on CBCT drawn by a radiation oncology expert. Nineteen structures that were either critical OARs or had large anatomical changes during radiotherapy courses were selected as segmentation targets. These structures

Figure 1. A U-Net architecture with influencer volumes trained with pseudo labels for CBCT segmentation. The input of the architecture is image (CBCT volumes) and influencer volumes (19-channel segmentation masks). The output/label of the architecture is a 19-channel segmentation masks representing 19 structures. In this workflow, both pseudo labels and influencer volumes are deformed pCT structures, but with different DIR methods.

were: left brachial plexus (L_BP), right brachial plexus (R_BP), brainstem, oral cavity, constrictor, esophagus, nodal gross tumor volume (nGTV), larynx, mandible, left masseter (L_Masseter), right masseter (R_Masseter), posterior arytenoid-cricoid space (PACS), left parotid gland (L_PG), right parotid gland (R_PG), left superficial parotid gland (L_Sup_PG), right superficial parotid gland (R_Sup_PG), left submandibular gland (L_SMG), right submandibular gland (R_SMG), and spinal cord. Contour masks have the same size and resolution with CBCT images.

To generate pseudo labels and influencer volumes, image registration was performed between pCT and CBCT for the 98 patients without true CBCT segmentation labels. pCT is first rigid registered to its corresponding CBCT through Velocity (Varian Medical Systems Inc., Palo Alto, CA, USA), and then deformedly registered to the CBCT through our previously proposed DIR methods (Liang *et al* 2022): applying TTO to three different state-of-the-art DIR models including FAIM, 5-cascaded Voxelmorph, and 10-cascaded VTN. Subsequently, the pCT contours were warped accordingly to generate deformed pCT contours as pseudo labels or influencer volumes for training. However due to DIR error in area with large anatomical changes or low image quality, the pseudo labels which is deformed pCT contours are not as accurate as manual contours which is gold standard. The remaining 39 patients with true CBCT labels which were manually delineated were

Figure 2. A U-Net architecture with influencer volumes fine-tuned with true labels for CBCT segmentation. The input of the architecture is image (CBCT volumes) and influencer volumes (19-channel segmentation masks). The output/label of the architecture is a 19-channel segmentation masks representing 19 structures. In this workflow, influencer volumes are deformed pCT structures and true labels come from manual contours on CBCT.

grouped into 30 for model fine-tuning and nine for model testing. CBCT images and contour masks were padded to size of $512 \times 512 \times 96$ and $512 \times 512 \times 96 \times 19$ from $512 \times 512 \times 93$ and $512 \times 512 \times 93 \times 19$, respectively.

2.6. Experiments

Four experiments were performed for this work, and are illustrated in figure 3. First, we trained the U-Net model on the 98 patients with pseudo labels by switching the training label among three types of pseudo labels without any prior knowledge ($Model_{pseudo}$) or adding any influencer volumes to study the performance of the direct segmentation. Then, we added influencer volumes into U-Net and trained the network with pseudo labels to observe the performance gained from adding influencer volumes ($Model_{influencer}$). Both the influencer volumes and pseudo labels were deformed pCT contours, but coming from different DIR algorithms. Finally, we fine-tuned $Model_{influencer}$ on 30 patients with true labels to further improve segmentation accuracy ($Model_{finetune}$). During the fine-tuning stage, we applied early stopping, layer freezing, and lower learning rate to prevent model overfitting. Considering the error of pseudo labels might influence the accuracy of the model trained on pseudo labels, we directly trained a segmentation model ($Model_{true}$) with the same architecture of $Model_{pseudo}$ but with true CBCT segmentation labels. However, due to limited amount of data available, only 30 patients with true CBCT segmentation labels were used to train $Model_{true}$ by starting from scratch. In this work, DIR-based contour propagating was used as the baseline, since it is the most commonly used method in current clinical practice for auto-segmentation in CBCT-based ART. We considered 10-cascaded VTN model with TTO as the

Figure 3. Workflow of experiments.
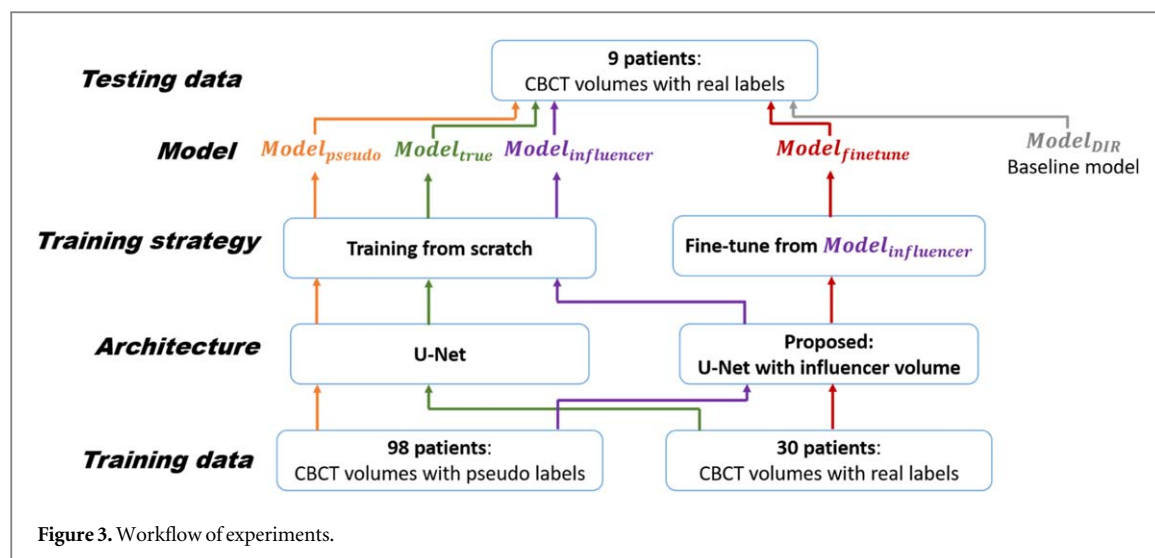
state-of-the-art DIR baseline model (Model$_{DIR}$). All five models were tested on nine patients. Dice similarity coefficient (DSC), Hausdorff distance at the 95th percentile (HD95), and average surface distance (ASD) were used to evaluate segmentation accuracy with statistical tests calculated.

3. Results

3.1. Model trained on pseudo labels without influencer volumes

Model$_{pseudo}$ exhibited worse performance than Model$_{DIR}$, as shown in tables 1, 2, and 3, with all 19 structures achieving lower DSC, higher HD95 and ASD scores with Model$_{pseudo}$. This suggests that models depending on CBCT images alone cannot derive reliable segmentation results. The main reasons for these inferior outcomes are listed below based on our observations.

Firstly and most importantly, CBCT images have many artifacts and low soft tissue contrast compared to CT images. Like the images shown in figure 4(a), some organs including the brainstem, esophagus, parotid gland, and submandibular gland do not have a clear boundary from surrounding tissues, making segmentation more difficult. However, those organs usually do not have significant anatomical changes, and the deformed pCT contours are quite accurate.

Secondly, for some OARs, superior and inferior ends sometimes produce large segmentation errors due to the direct segmentation model's inability to deal with certain geometry information if not provided with guidelines, as shown in figure 4(b). For example, a consensus guideline for CT-based delineation of the constrictor (Brouwer *et al* 2015) specified that the cranial border was defined as the caudal tip of pterygoid plates, and the caudal border as the lower edge of the cricoid cartilage. However, Model$_{pseudo}$ failed to pick up this border data from the training data directly, leading to delineation errors around the superior and inferior borders.

Thirdly, target volumes and some OARs are extremely challenging to segment, even on CT. Target delineation, like nodal clinical tumor volume, is more variable and therefore more difficult to predict than OARs. The brachial plexus is difficult to localize on CT images and is identified with adjacent structures using additional help from anatomic texts or magnetic resonance imaging. It is unsurprising to see that DIR-based segmentation prediction has fairly good performance, since it uses pCT contours as a start point. However, direct segmentation of those extremely difficult structures is prone to failure.

Fourthly, direct segmentation models are prone to inaccurate or incomplete labels in the training dataset. For example, in figure 4(d), the manual contours are actually the left superficial parotid gland, but mislabeled as parotid gland. In addition, some structures may be modified to represent avoidance structures and not necessarily hold fast to the exact anatomic boundaries of that organ, such as with the oral cavity where the portion overlapping the planning tumor volume is sometimes cropped out, leaving the manual segmentation of oral cavity incomplete. We also found that some organs lack complete delineation in the superior–inferior direction in the training dataset, because a complete delineation was unnecessary if a part of the organ was distant from the target volume.

Fifthly, outliers in the testing dataset have a negative impact on model performance. Figure 4(e) shows two outliers that we observed in the testing dataset. One test patient had a tracheostomy tube, however, no such patients are in the training dataset. The presence of the tracheostomy tube leads to incorrect delineation of the

Table 1. Mean and standard deviation of DSC of 9 test patients for different auto-segmentation models. Model$_{DIR}$ is DIR only segmentation, which is baseline model in this study. Model$_{pseudo}$ is direct DL segmentation using pseudo labels for training. Model$_{true}$ is direct DL segmentation using true labels for training. Model$_{influencer}$ is direct DL segmentation using both pseudo labels and influencer volumes for training. Model$_{finetune}$ is derived from fine-tuning Model$_{influencer}$ with true labels. Paired sample *T* tests are conducted between the baseline model (Model$_{DIR}$) and other models (Model$_{pseudo}$, Model$_{true}$, Model$_{influencer}$, Model$_{finetune}$). Numbers in green and red means *P*-value < 0.05, otherwise *P*-value > 0.05, which indicates that the model predicted segmentation with DSC in red is less accurate than Model$_{DIR}$ predicted segmentation, and vice versa for model predicted segmentation with DSC in green.

Structure	Model$_{DIR}$	Model$_{pseudo}$	Model$_{true}$	Model$_{influencer}$	Model$_{finetune}$
L_BP	0.71±0.09	0.34±0.10	0.38±0.08	0.71±0.07	0.72±0.08
R_BP	0.73±0.05	0.40±0.10	0.34±0.10	0.73±0.04	0.75±0.04
Brainstem	0.91±0.04	0.72±0.10	0.71±0.08	0.90±0.04	0.91±0.04
Oral cavity	0.95±0.01	0.60±0.15	0.57±0.21	0.95±0.01	0.95±0.01
Constrictor	0.83±0.03	0.72±0.12	0.70±0.20	0.85±0.02	0.86±0.01
Esophagus	0.83±0.06	0.62±0.12	0.56±0.20	0.83±0.07	0.83±0.07
nGTV	0.83±0.06	0.32±0.21	0.30±0.23	0.84±0.05	0.84±0.06
Larynx	0.89±0.05	0.73±0.04	0.67±0.12	0.89±0.04	0.90±0.04
Mandible	0.89±0.05	0.87±0.04	0.88±0.03	0.90±0.04	0.91±0.04
L_Masseter	0.90±0.03	0.84±0.04	0.80±0.08	0.90±0.02	0.92±0.02
R_Masseter	0.90±0.03	0.83±0.06	0.80±0.07	0.90±0.04	0.90±0.05
PACS	0.78±0.03	0.66±0.09	0.65±0.07	0.79±0.02	0.81±0.02
L_PG	0.89±0.03	0.70±0.14	0.67±0.12	0.88±0.04	0.89±0.03
R_PG	0.91±0.01	0.72±0.10	0.72±0.08	0.90±0.02	0.91±0.01
L_Sup_PG	0.84±0.06	0.60±0.13	0.58±0.13	0.84±0.05	0.84±0.05
R_Sup_PG	0.86±0.03	0.50±0.14	0.60±0.07	0.85±0.02	0.86±0.02
L_SMG	0.84±0.01	0.59±0.20	0.60±0.10	0.85±0.02	0.86±0.01
R_SMG	0.84±0.04	0.55±0.13	0.51±0.15	0.85±0.04	0.85±0.03
Spinal cord	0.86±0.05	0.75±0.06	0.80±0.05	0.86±0.05	0.89±0.04
Mean	0.85	0.63	0.62	0.85	0.86

Table 2. Mean and standard deviation of HD95 of 9 test patients for different auto-segmentation models. Model$_{DIR}$ is DIR only segmentation, which is baseline model in this study. Model$_{pseudo}$ is direct DL segmentation using pseudo labels for training. Model$_{true}$ is direct DL segmentation using true labels for training. Model$_{influencer}$ is direct DL segmentation using both pseudo labels and influencer volumes for training. Model$_{finetune}$ is derived from fine-tuning Model$_{influencer}$ with true labels. Paired sample T tests are conducted between the baseline model (Model$_{DIR}$) and other models (Model$_{pseudo}$, Model$_{true}$, Model$_{influencer}$, Model$_{finetune}$). Numbers in green and red means P-value < 0.05, otherwise P-value > 0.05, which indicates that the model predicted segmentation with DSC in red is less accurate than Model$_{DIR}$ predicted segmentation, and vice versa for model predicted segmentation with DSC in green. The unit of HD95 is mm.

Structure	Model$_{DIR}$	Model$_{pseudo}$	Model$_{true}$	Model$_{influencer}$	Model$_{finetune}$
L_BP	3.11±1.72	16.44±7.21	17.62±10.70	3.04±1.33	3.08±1.49
R_BP	4.23±4.15	14.25±3.19	15.96±10.80	4.08±4.18	3.86±3.79
Brainstem	1.78±0.79	6.29±1.51	6.18±2.32	1.56±0.44	1.56±0.50
Oral cavity	2.50±0.79	23.27±7.13	24.73±10.00	2.61±0.68	2.60±0.66
Constrictor	1.84±0.10	6.03±2.62	11.94±5.00	1.84±0.87	1.99±1.34
Esophagus	2.50±1.11	10.38±7.41	12.28±11.44	2.49±1.07	2.52±1.20
nGTV	2.68±0.94	51.79±23.57	45.21±28.85	2.67±0.87	2.52±0.87
Larynx	3.23±1.34	9.02±2.53	12.49±4.38	2.96±0.97	2.91±0.89
Mandible	2.15±1.59	4.34±3.94	4.76±3.14	2.10±1.56	2.00±1.63
L_Masseter	1.63±0.38	2.99±0.50	4.40±1.26	1.59±0.33	1.48±0.34
R_Masseter	1.84±0.32	3.82±1.27	4.89±2.73	1.66±0.39	1.74±0.29
PACS	3.32±0.78	6.24±2.68	8.14±2.97	3.00±0.86	2.79±0.88
L_PG	2.29±0.48	8.96±4.03	10.15±2.15	2.13±0.24	2.07±0.42
R_PG	2.43±1.52	9.03±5.53	12.51±10.80	2.44±1.51	1.93±0.21
L_Sup_PG	2.58±0.48	11.41±3.49	10.24±4.53	2.35±0.25	2.00±0.12
R_Sup_PG	2.77±0.72	15.25±5.42	11.73±2.46	2.58±0.69	2.16±0.35
L_SMG	2.46±0.58	13.51±17.11	7.38±2.83	2.30±0.54	2.31±0.35
R_SMG	2.48±0.57	9.16±3.15	10.94±5.18	2.52±0.61	2.50±0.67
Spinal cord	2.82±2.89	15.25±5.42	5.63±5.02	2.79±2.74	2.50±2.12
Mean	2.56	12.50	12.48	2.46	2.34

Table 3. Mean and standard deviation of ASD of 9 test patients for different auto-segmentation models. $Model_{DIR}$ is DIR only segmentation, which is baseline model in this study. $Model_{pseudo}$ is direct DL segmentation using pseudo labels for training. $Model_{true}$ is direct DL segmentation using true labels for training. $Model_{influencer}$ is direct DL segmentation using both pseudo labels and influencer volumes for training. $Model_{finetune}$ is derived from fine-tuning $Model_{influencer}$ with true labels. Paired sample T tests are conducted between the baseline model ($Model_{DIR}$) and other models ($Model_{pseudo}$, $Model_{true}$, $Model_{influencer}$, $Model_{finetune}$). Numbers in green and red means *P*-value < 0.05, otherwise *P*-value > 0.05, which indicates that the model predicted segmentation with DSC in red is less accurate than $Model_{DIR}$ predicted segmentation, and vice versa for model predicted segmentation with DSC in green. The unit of ASD is mm.

| Structure	$Model_{DIR}$	$Model_{pseudo}$	$Model_{true}$	$Model_{influencer}$	$Model_{finetune}$
L_BP	0.88±0.45	2.99±1.05	2.39±1.66	0.80±0.27	0.72±0.34
R_BP	0.78±0.20	2.41±0.66	2.53±0.85	0.69±0.15	0.59±0.14
Brainstem	0.41±0.21	1.61±0.55	1.58±0.48	0.39±0.19	0.37±0.18
Oral cavity	0.72±0.13	8.00±3.00	7.16±2.73	0.69±0.17	0.70±0.13
Constrictor	0.55±0.40	1.36±0.50	0.88±2.10	0.51±0.38	0.46±0.31
Esophagus	0.68±0.42	2.91±2.49	3.49±5.48	0.65±0.42	0.65±0.42
nGTV	0.69±0.27	12.30±7.94	13.66±13.00	0.70±0.27	0.65±0.23
Larynx	0.93±0.39	2.48±0.40	3.68±2.00	0.94±0.39	0.89±0.31
Mandible	0.36±0.17	0.79±0.58	0.97±0.69	0.32±0.15	0.28±0.15
L_Masseter	0.39±0.06	0.85±0.22	1.05±0.45	0.39±0.05	0.32±0.04
R_Masseter	0.41±0.05	0.91±0.20	1.04±0.33	0.43±0.13	0.49±0.17
PACS	0.75±0.21	1.75±0.75	2.01±0.57	0.77±0.19	0.74±0.15
L_PG	0.52±0.12	2.19±1.58	2.45±1.32	0.53±0.13	0.51±0.10
R_PG	0.51±0.11	2.05±1.12	2.41±0.89	0.52±0.16	0.44±0.13
L_Sup_PG	0.54±0.12	2.06±0.85	3.03±1.18	0.55±0.08	0.56±0.07
R_Sup_PG	0.59±0.14	2.13±1.07	2.26±0.73	0.64±0.25	0.61±0.18
L_SMG	0.64±0.19	3.30±4.63	2.08±0.52	0.63±0.22	0.58±0.17
R_SMG	0.70±0.22	2.65±1.05	2.56±0.84	0.61±0.18	0.58±0.17
Spinal cord	0.69±0.53	1.21±0.38	0.85±0.26	0.67±0.52	0.50±0.38
Mean	0.62	2.84	2.95	0.60	0.56
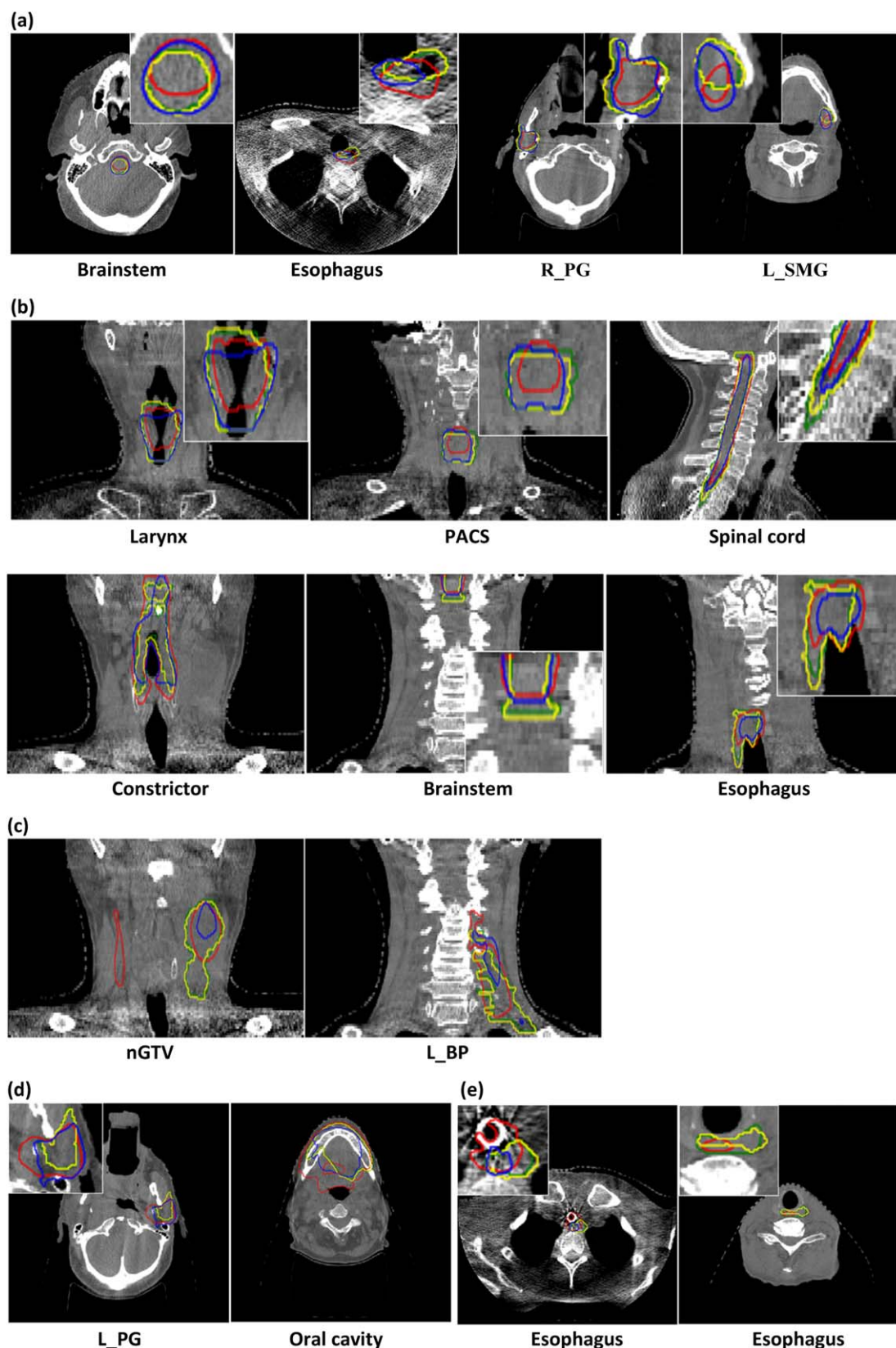
Figure 4. Categories of reasons that cause poor performance of direct segmentation without any prior knowledge on CBCT. Green lines are manual contours drawn by a radiation oncology expert, yellow lines are DIR propagated contours (Model$_{DIR}$), red lines come from direct segmentation without influencer volumes trained on pseudo labels (Model$_{pseudo}$), and blue lines are direct segmentation without influencer volumes trained on true labels (Model$_{true}$).

esophagus by the direct segmentation model. Another test patient had the esophagus pushed away to his left side, but no similar patient exists in the training dataset. Model$_{pseudo}$ usually has poor performance on outliers, while Model$_{DIR}$ is more accurate by preserving shape information from pCT contours.

3.2. Model trained on real labels without influencer volumes

Model$_{true}$ also exhibited much worse performance than Model$_{DIR}$, as shown in tables 1, 2, and 3, with all 19 structures in DSC, HD95, and ASD evaluation. The main reasons for the inferior outcomes are the same with the reasons listed above in section 3.1, except there are no inaccurate labels in the true label dataset. With much smaller size of data (30 patients) to train a 3D U-Net model, Model$_{true}$ also suffers overfitting problem. This indicates that, with limited data available, models depending on CBCT images alone cannot derive reliable segmentation results.

3.3. Model trained on pseudo labels with influencer volumes

Tables 1, 2, and 3 show that Model$_{DIR}$ and Model$_{influencer}$ have similar DSC, HD95, and ASD scores over all 19 structures. With the use of influencer volumes from pseudo labels, the performance of the DS model can be significantly improved to the level of DIR-based segmentation. It is not surprising that the performance of Model$_{influencer}$ does not surpass that of Model$_{DIR}$, since Model$_{influencer}$ used the pseudo labels generated by DIR for training.

3.4. Model fine-tuned on real labels with influencer volumes

When Model$_{influencer}$ is fine-tuned using true labels, the DIR errors contained in pseudo labels for training could potentially be corrected, allowing for the model performance to surpass that of Model$_{DIR}$. Tables 1, 2, and 3 shows that DSC, HD95, and ASD scores of Model$_{finetune}$ are better than or equal to those of Model$_{DIR}$ and Model$_{influencer}$. Paired sampled T tests were performed between Model$_{DIR}$ and other models for each structure. Significant difference ($P < 0.05$) were colored in red or green (red: predicted segmentation is less accurate, green: predicted segmentation is more accurate). Model$_{finetune}$ outperforms Model$_{DIR}$ for 8, 9, and 8 structures in DSC, HD95, and ASD evaluation respectively. The average DSC over 19 structures by Model$_{finetune}$ is 0.86, with a minimum DSC of 0.72 for L_BP and maximum DSC of 0.95 for the oral cavity. The average HD95 and ASD over 19 structures by Model$_{finetune}$ are 2.34 mm and 0.56 mm, with a minimum HD95 of 1.48 mm for L_Masseter, maximum HD95 of 3.86 mm for R_BP, minimum ASD of 0.28 mm for mandible, and maximum ASD of 0.72 mm for L_BP. Examples of segmentation from axial, frontal, and sagittal views by Model$_{DIR}$, Model$_{influencer}$, and Model$_{finetune}$ are shown in figure 5 for visual evaluation. We can see that Model$_{finetune}$ not only can maintain shape characteristics of the prior segmentation in pCT, it can also eliminate the errors caused by the prior segmentation.

4. Discussion and conclusions

Based on this work, it is evident that direct segmentation on CBCT images without prior knowledge is infeasible, mainly due to the poor image quality, superior and inferior border uncertainty, delineation complexity, outliers, inaccurate or incomplete labels, and also lack of true labels. With pCT and its corresponding contours available in the ART workflow as prior knowledge, the accuracy of DL-based direct CBCT segmentation can be greatly improved.

Different from CT auto-segmentation tasks where large amount of labels are usually available for training, manual labels are not common on CBCT images. To solve this lack of a large well-labelled training dataset, we proposed to use deformed pCT as pseudo labels for the initial DL model training, as the performance of the initially trained model is far inferior to the DIR-based methods. We then proposed to use deformed pCT from another DIR algorithm as influencer volumes in the network. By adding influencer volumes as new channels to the model to constrain shape and localization, the model performance can be dramatically improved, reaching the level of DIR-based methods. To outperform the DIR-based methods, the DL-based direct segmentation model initially trained with pseudo labels and influencer volumes can be fine-tuned using a small set of training data with true labels.

Fine-tuning with true labels could mitigate DIR errors contained in the pseudo labels since in the fine-tuning stage, there are several ways to prevent overfitting (Ying 2019). Reducing overfitting by training the network on more datasets is not considered in this work since we have only have a small amount of data with true labels. Reducing overfitting by changing the complexity of the network is another way to prevent overfitting. For example, the model could be tuned via freezing some layers and only updating parameters of the remaining layers. Another simple alternative to avoid overfitting is to improve regularization (Goodfellow *et al* 2016) by early stopping via monitoring model performance on a validation set and stopping training when performance degrades. Meanwhile, adding regularization requires a smaller learning rate.

This application is designed to be used in CBCT-based ART for CBCT auto-segmentation. In the ART workflow, pCT and physician contours are always available before CBCT scan. However, OARs that need to be delineated are different according to target location. That's why not all 19 structures exist on pCT in our training
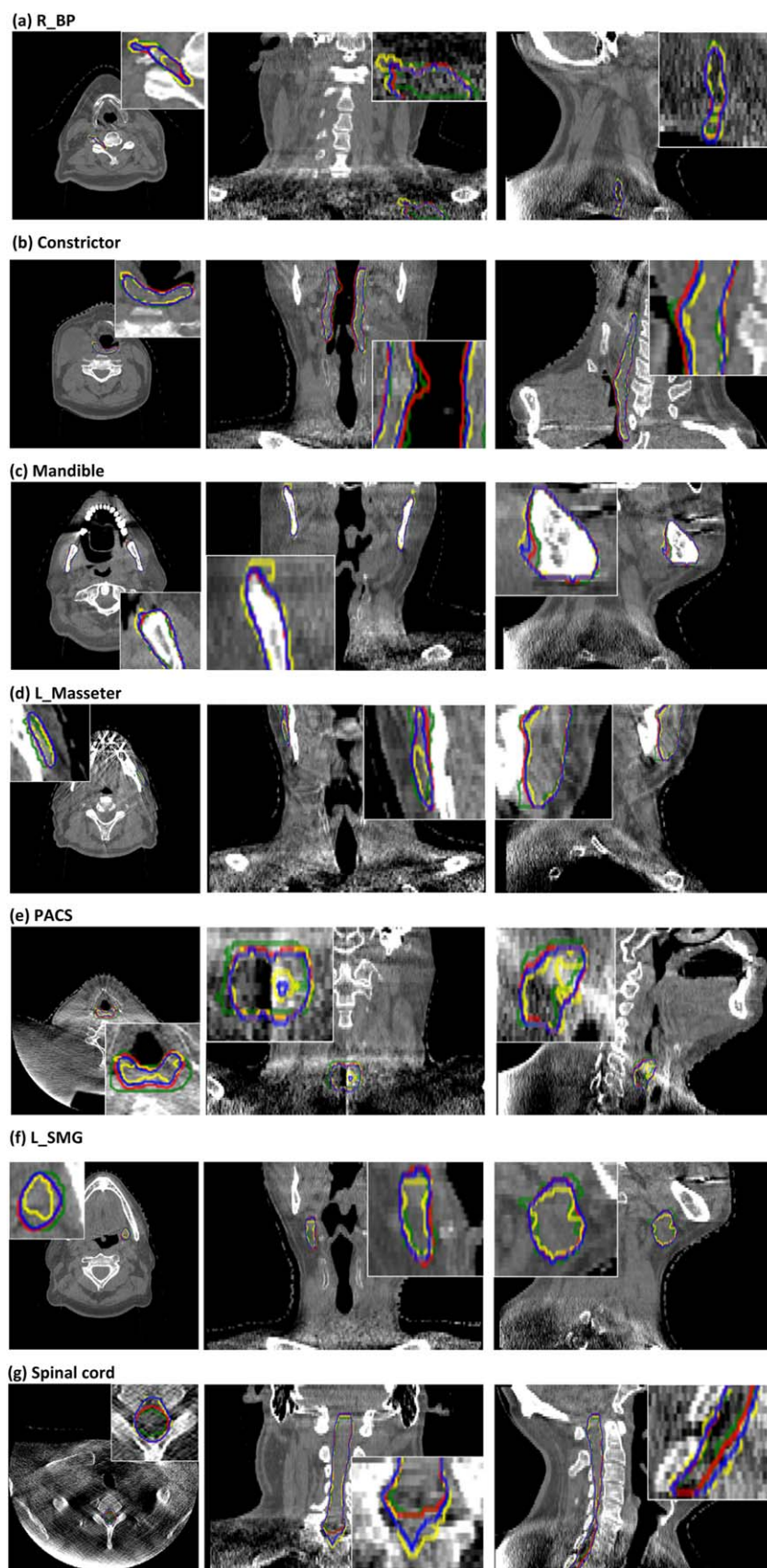
Figure 5. Segmentation examples in axial, frontal, and sagittal views. Green lines are manual contours drawn by a radiation oncology expert, yellow lines are DIR-based segmentation (Model$_{DIR}$), blue lines are segmentations with influencer volumes trained with pseudo labels (Model$_{influencer}$), and red lines are segmentations from fine-tuned direct segmentation with influencer volumes (Model$_{finetune}$).

dataset. The OARs that need to be delineated on CBCT are always the same with the OARs that have delineated on pCT by a physician. Therefore those missing OARs on pCT do not need to be contoured on CBCT in the ART workflow. The missing OARs will not affect model prediction of the other OARs.

We used H&N patients to test our models, since CBCT-based ART is often used for this disease site and since segmentation is more challenging. The same approach can be easily expanded to and tested on datasets from other disease sites.

In summary, to overcome the two major issues related to CBCT-based image segmentation for online ART, such as poor image quality and lack of well-labelled large training datasets, we developed a method to use DIR-propagated contours as pseudo labels and influencer volumes for initial training and subsequently fine-tuned the model using a small set of a training dataset with true labels. The method has been tested with a cohort of H&N cancer patients and demonstrated superior segmentation accuracy to the commonly used DIR-based methods.

Acknowledgments

Conflict of interest statement

The authors declare no competing financial interest. The authors confirm that all funding sources supporting the work and all institutions or people who contributed to the work, but who do not meet the criteria for authorship, are acknowledged. The authors also confirm that all commercial affiliations, stock ownership, equity interests or patent licensing arrangements that could be considered to pose a financial conflict of interest in connection with the work have been disclosed.

Ethical statement

UT Southwestern Institutional Review Boards approved the research with approval ID of STU-082013-008.

ORCID iDs

Xiao Liang ⬤ https://orcid.org/0000-0002-2472-2396
Ti Bai ⬤ https://orcid.org/0000-0002-6697-7434
Dan Nguyen ⬤ https://orcid.org/0000-0002-9590-0655

References

Alam S R, Li T, Zhang P, Zhang S-Y and Nadeem S 2021 Generalizable cone beam CT esophagus segmentation using physics-based data augmentation *Phys. Med. Biol.* **66** 065008
Archambault Y, Boylan C, Bullock D, Morgas T, Peltola J, Ruokokoski E, Genghi A, Haas B, Suhonen P and Thompson S 2020 Making on-line adaptive radiotherapy possible using artificial intelligence and machine learning for efficient daily re-planning *Med. Phys. Intl. J.* **8** 77–86
Beekman C, van Beek S, Stam J, Sonke J-J and Remeijer P 2021 Improving predictive CTV segmentation on CT and CBCT for cervical cancer by diffeomorphic registration of a prior *Med. Phys.* **49** 1701–11
Brouwer C L *et al* 2015 CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines *Radiother. Oncol.* **117** 83–90
Dahiya N, Alam S R, Zhang P, Zhang S-Y, Li T, Yezzi A and Nadeem S 2021 Multitask 3D CBCT-to-CT translation and organs-at-risk segmentation using physics-based data augmentation *Med. Phys.* **48** 5130–41
Dai X, Lei Y, Wang T, Dhabaan A H, McDonald M, Beitler J J, Curran W J, Zhou J, Liu T and Yang X 2021 Head-and-neck organs-at-risk auto-delineation using dual pyramid networks for CBCT-guided adaptive radiotherapy *Phys. Med. Biol.* **66** 045021
Dalca A V, Balakrishnan G, Guttag J and Sabuncu M R 2019 Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces *Med. Image Anal.* **57** 226–36
Estienne T *et al* 2019 U-ReSNet: Ultimate Coupling of Registration and Segmentation with Deep Nets *MICCAI 2019: Medical Image Computing and Computer Assisted Intervention—MICCAI 2019* (Shenzhen, China) pp 310–9
Fedorov A *et al* 2012 3D Slicer as an image computing platform for the quantitative imaging network *Magn. Reson. Imaging* **30** 1323–41
Glide-Hurst C K *et al* 2021 Adaptive radiation therapy (ART) strategies and technical considerations: a state of the ART review from NRG oncology *Int. J. Radiat. Oncol.*Biol.*Phys.* **109** 1054–75
Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge, MA: MIT press)
Gu X, Pan H, Liang Y, Castillo R, Yang D, Choi D, Castillo E, Majumdar A, Guerrero T and Jiang S B 2009 Implementation and evaluation of various demons deformable image registration algorithms on a GPU *Phys. Med. Biol.* **55** 207–19

Han X *et al* 2021 Deep-learning-based image registration and automatic segmentation of organs-at-risk in cone-beam CT scans from high-dose radiation treatment of pancreatic cancer *Med. Phys.* **48** 3084–95

Klein S, Staring M, Murphy K, Viergever M A and Pluim J P W 2010 a toolbox for intensity-based medical image registration *IEEE Trans. Med. Imaging* **29** 196–205

Kuang D and Schmah T 2019 Faim—a convnet method for unsupervised 3d medical image registration *Int. Workshop on Machine Learning in Medical Imaging* (Berlin: Springer) pp 646–54

Lechuga L and Weidlich G A 2016 Cone Beam CT versus Fan Beam CT: a comparison of image quality and dose delivered between two differing CT imaging modalities *Cureus* **8** e778–778

Léger J, Brion E, Desbordes P, De Vleeschouwer C, Lee J A and Macq B 2020 Cross-domain data augmentation for deep-learning-based male pelvic organ segmentation in cone beam CT *Appl. Sci.* **10** 1154

Liang X, Chun J, Morgan H, Bai T, Nguyen D, Park J C and Jiang S 2022 Segmentation by test-time optimization (TTO) for CBCT-based adaptive radiation therapy *Med. Phys.* Early view (https://doi.org/10.1002/mp.15960)

Schreier J, Genghi A, Laaksonen H, Morgas T and Haas B 2020 Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam CT and CT *Radiother. Oncol.* **145** 1–6

Xu Z and Niethammer M 2019 DeepAtlas: joint semi-supervised learning of image registration and segmentation *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019* ed D Shen *et al* (Cham: Springer International Publishing) pp 420–9

Ying X 2019 An overview of overfitting and its solutions *J. Phys. Conf. Ser.* **1168** 022022

Zhao S, Dong Y, Chang E I and Xu Y 2019 Recursive cascaded networks for unsupervised medical image registration *Proc. of the IEEE/CVF Int. Conf. on Computer Vision* pp 10600–10