One shot PACS: <u>P</u>atient specific <u>A</u>natomic <u>C</u>ontext and <u>S</u>hape prior aware recurrent registration-segmentation of longitudinal thoracic cone beam CTs

Jue Jiang,

Harini Veeraraghavan

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, NY, New York, USA.

Abstract

Image-guided adaptive lung radiotherapy requires accurate tumor and organs segmentation from during treatment cone-beam CT (CBCT) images. Thoracic CBCTs are hard to segment because of low soft-tissue contrast, imaging artifacts, respiratory motion, and large treatment induced intra-thoracic anatomic changes. Hence, we developed a novel Patient-specific Anatomic Context and Shape prior or PACS-aware 3D recurrent registration-segmentation network for longitudinal thoracic CBCT segmentation. Segmentation and registration networks were concurrently trained in an end-to-end framework and implemented with convolutional long-short term memory models. The registration network was trained in an unsupervised manner using pairs of planning CT (pCT) and CBCT images and produced a progressively deformed sequence of images. The segmentation network was optimized in a one-shot setting by combining progressively deformed pCT (anatomic context) and pCT delineations (shape context) with CBCT images. Our method, one-shot PACS was significantly more accurate (p <0.001) for tumor (DSC of 0.83 ± 0.08, surface DSC [sDSC] of 0.97 ± 0.06, and Hausdorff distance at 95th percentile [HD95] of 3.97 ± 3.02mm) and the esophagus (DSC of 0.78 ± 0.13, sDSC of 0.90 ± 0.14, HD95 of 3.22 ± 2.02) segmentation than multiple methods. Ablation tests and comparative experiments were also done.

Keywords

One-shot learning; CBCT lung tumor and esophagus segmentation; multi-modality registration; recurrent network; anatomic context and shape prior

I. Introduction

Adaptive image-guided radiation treatments (AIGRT) of lung cancers require accurate segmentation of tumor and organs at risk (OAR) such as the esophagus from during treatment cone-beam CT (CBCT) [1]. Tumors are difficult to segment due to very low soft-tissue contrast on CBCT, imaging artifacts, radiotherapy (RT) induced radiographic and size changes, and large intra- and inter-fraction motion [1]. Normal organ like the esophagus

jiangj1@mskcc.org.

is also hard to segment due to low soft-tissue contrast and displacements exceeding 4mm between treatment fractions [2].

Cross-modality deep learning methods have used structure constraints from planning CT (pCT) [3], as well as MRI contrast as prior knowledge to improve pelvic organ [4] and lung tumor [5] segmentation from CBCT. However, expert delineated CBCTs needed for training are not routinely segmented and suffer from high inter-rater variability [6].

Atlas-based image registration methods overcome the issue of limited segmented datasets by directly propagating segmentations [7], [8] as well as by providing synthesized images as augmented data for segmentation training [9]-[11]. Cross-domain adaptation based synthetic CBCT generation based data augmentation [12] is another promising approach used for pelvic organs segmentation. A hybrid approach [13] combining data augmentation using cross-domain adaptation of pCT, MRI, and CBCT with multi-modality registration was used for liver segmentation from during treatment CBCTs.

Multi-task networks [10], [14]-[16] handle limited datasets by using implicit data augmentation available from the different tasks through the losses to jointly optimize registration and segmentation. Notably, these methods have shown feasibility for one and few-shot normal organ segmentation [10], [14], [17], using CT-to-CT or MRI-to-MRI registration. Planning CT to CBCT registration is harder because of low soft-tissue contrast and narrow field of view (FOV) on CBCT [13].

Prior works applied to CBCT registration aligned large moving organs [4], [10], [13], [18], [19]. We tackle more challenging lung tumor segmentation from CBCT for longitudinal response assessment in tumors undergoing treatment. Previously, longitudinal tracking in anatomy depicting large changes, such as growing infant brains [8] and pre- and post-surgical brains [20] aligned same modalities with high contrast (MRI-to-MRI). Cross-domain adaptation based synthetic CT [21] as well as deep network combined with scale invariant feature transform (SIFT) detected features [19], and surface points registration of multiple modalities [13] are example approaches used to handle low soft-tissue contrast on CBCT. We address multiple challenges, namely, multi-modal registration (pCT to CBCT), longitudinal registration of highly deforming diseased and healthy tissues with altered appearance from treatment, and segmentation on low soft-tissue contrast CBCT.

In order to tackle all these challenges, our approach combines an end-to-end trained joint recurrent registration network (RRN) and recurrent segmentation network (RSN). Our approach models large local deformations by computing progressive deformations that incrementally improves regional alignment. This is accomplished by using convolutional long-short term memory (CLSTM) [22] to implement the recurrent units of RRN and RSN. CLSTM models long-range temporal deformation dynamics, needed to model the progressive deformations in regions undergoing large deformations. The convolutional layers used in CLSTM models the spatial dynamics of a dense 3D flow field compared to 1D information computed by LSTM [23]. Our approach increases flexibility to capture longitudinal size and shape changes in tumors compared to the Recurrent Registration Neural Network (R2N2) [24], which computes parameterized local deformations. Finally, in

order to handle low contrast on CBCT, RSN combines progressively aligned anatomical context (pCT) and shape (pCT delineation) prior produced by a jointly trained RRN. Hence, we call our approach patient-specific anatomic context and shape prior or PACS-aware registration-segmentation network. We show that our approach is more accurate than multiple methods.

The RRN is trained in an unsupervised way using only pairs of target and moving images without structure guidance from segmented pCT or CBCT. RSN is optimized with a single segmented CBCT example and combines progressively warped pCT images and delineations produced by the RRN with CBCT for one-shot training. Our contributions are:

- Multi-modal recurrent joint registration-segmentation approach to handle large anatomic changes in tumors undergoing treatment and highly deforming esophagus.

- One-shot segmentation with patient-specific anatomic context and shape priors, which handles tumor segmentation despite varying size and locations. To our best knowledge, this is the first one-shot learning approach for longitudinal segmentation of lung tumors from CBCT.

- A recurrent registration network that interpolates dense flow field using only a pair of pCT and CBCT images and optimized with unsupervised training.

- Comprehensive comparison, ablation, and network design experiments to study accuracy.

II. Related Works

A. Medical image registration-based segmentation

One-shot and few-shot learning strategies extract a model from only a single or a few labeled examples. Hence, these are attractive options for medical image analysis where large number of expert segmented cases are not available. As elucidated by Wang et.al [25], a key difference between few shot learning applied to natural vs. medical images is that, in the former, learning is concerned with extracting a model to recognize a new class based on appearance similarities to previously learned classes. Medical image analysis methods are concerned with better modeling the anatomic similarity between subjects using few examples where all classes are available. The challenge is to extract a representation that is robust to imaging and anatomic variability among patients.

Iterative registration methods sidestep the issue of learning by using the available segmented cases as atlases [26]. Assuming that the atlases represent the variability in the patient anatomy, these methods provide reasonably accurate tissue segmentations [7]. Unsupervised deep learning-based synthesis of realistic training samples [9] as augmented datasets are a more robust option, because they do not suffer from catastrophic failures. Also, once trained the network produces computationally fast segmentations unlike iterative registration. Nonetheless, accuracy is reduced due to poor image quality and large anatomical changes [27].

Joint registration-segmentation methods [14]-[16], [28] are more accurate than registration-based segmentation. This is because, these methods model the interaction between registration and segmentation features and improve accuracy. Furthermore, these approaches are amenable to training with few segmented examples including semi-supervised learning. For example, a registration network was used to segment unlabeled data for training [11], [14] as well as create augmented samples through random perturbations in the warped images [10]. These methods benefit by using the segmentation network to provide additional regularization losses to optimize the registration network training such as through segmentation consistency [14] and cycle consistency losses [11]. However, the one-step registration computed by these methods may not handle very large deformations.

Multi-stage [29] and recursive cascade registration [30] methods have shown that incrementally refined outputs produced from intermediate steps as inputs to subsequent steps increases accuracy to model large deformations but require large number of parameters. R2N2 [24] improves on these methods by using gated recurrent units with local Gaussian basis functions and captures organ deformations occurring in a breathing cycle from MRI.

B. Shape and spatial priors for registration-segmentation

Template shape constraints [31] as well as population level anatomical priors learned using a generative model [32], [33] were previously used to regularize same modality registrations. Segmentation as auxiliary supervised information has shown to provide more accurate registrations [14], [34]-[36]. Structure guidance from pCT [3] and anatomical priors priors [32], [33] have shown to improve accuracy. Improving on prior works that used either shape [3], [31], [34]-[36] or anatomical priors [32], [33], we use both priors and show accuracy gains, even in the one-shot segmentation training scenario, where a single segmented CBCT example is used for training. Our joint recurrent registration-segmentation network computes a dense 3D flow field but uses fewer parameters than cascaded methods [29], [30] and is more accurate than R2N2 [24].

Rationale for combining pCT as anatomical context and and delineations as shape prior: The pCT has a higher soft-tissue contrast than the CBCT scans, which can provide a spatially aligned anatomic context to improve inference from lower-contrast CBCT. We also expect that the pCT segmentations used as patient-specific priors to segment CBCT will be informative of the tumor and organ shape for segmentation.

III.   Method

A.   Background

Problem setting and approach: Given a single segmented CBCT example $\{x_{cb}^e, y_{cb}^e\}$, it's corresponding delineated pCT $\{x_c^e, y_c^e\}$, as well as several unsegmented CBCTs $x_{cb} \in X_{CBCT}$ and their corresponding segmented pCTs $\{x_c, y_c\} \in \{X_C, Y_C\}$, our goal is to construct a model to generate tumor and esophagus segmentations from weekly CBCTs.

Our approach uses a joint recurrent registration-segmentation network (Fig. 1). The recurrent registration network (RRN) g and recurrent segmentation network (RSN) s are implemented

using CLSTM [22]. RRN aligns x_c to x_{cb} and produces progressively deforming $\{x_c^t, y_c^t\}$ in N CLSTM steps, where $1 \leq t \leq N$. RSN computes a segmentation y_{cb} for x_{cb} by combining progressively warped $\{x_c^t, y_c^t\}$ produced by the RRN using $N+1$ CLSTM steps. As shown in Fig. 1, the first CLSTM step ($t = 0$) of RSN uses input pCT and it's delineation (x_c^0, y_c^0) with x_{cb}. On the other hand, the CLSTM steps $t \geq 1$ of RSN use the outputs of RRN (x_c^t, y_c^t) with x_{cb}.

Classical dynamic system vs. basic recurrent network vs. LSTM vs. CLSTM
—A classical dynamic system (CDS) uses shared feature layers to produce outputs sequentially [37], represented as, $x^t = f(x^{t-1}; \theta)$. A basic recurrent neural network (RNN) also includes a hidden state [37], ($x^t = f(x^{t-1}, h^{t-1}; \theta)$) in order to blend information from preceding temporal step. LSTM and CLSTM are a type of RNN, which use feedback through forget gate and memory cells to capture long range temporal information. CLSTM uses convolution layers to model dense spatial dynamics whereas LSTM models 1D dynamics through fully connected layers [22]. A diagram of the differences are shown in Supplementary Fig 2.

Convolutional long short term memory network: CLSTM is a recurrent network that was introduced to model the dynamics within 2D spatial region [22] via convolution. We extended CLSTM to model deformation dynamics in a 3D spatial region. Moreover, our approach models the large deformation dynamics as an interpolated temporal deformation sequence (for RRN) or an interpolated segmentation sequence (for RSN), given only the start (pCT and its contour) and end images (CBCT image to be segmented) of the sequence.

A CLSTM unit is composed of a memory cell c^t, which accumulates the state x^t at step t, a forget gate f^t that keeps track of relevant state information from the past, a hidden state h^t, which encodes the state, as well as input state i^t, and output gate o^t. The CLSTM components are updated as:

$$
\begin{aligned}
f^t &= \sigma(W_{xf} * x^t + W_{hf} * h^{t-1} + b_f)\\
i^t &= \sigma(W_{xi} * x^t + W_{hi} * h^{t-1} + b_i)\\
\widetilde{c}^t &= tanh(W_{x\widetilde{c}} * x^t + W_{h\widetilde{c}} * h^{t-1} + b_{\widetilde{c}})\\
o^t &= \sigma(W_{xo} * x^t + W_{ho} * h^{t-1} + b_o)\\
c^t &= f^t \odot c^{t-1} + i^t \odot \widetilde{c}^t\\
h^t &= o^t \odot tanh(c^t),
\end{aligned}
\tag{1}
$$

where, σ is the sigmoid activation function, $*$ the convolution operator, $\odot$ the Hadamard product, and W the weight matrix.

B. Planning CT to CBCT deformable image registration

The RRN g computes a deformation of x_c to x_{cb}, expressed as $g(x_c, x_{cb}) : \theta_g(x_c) \rightarrow x_c^{cb}$ as a sequence of deformation vector fields (DVF) using N steps: $\phi_c^{cb} = \phi^1 \circ \phi^2 ... \circ \phi^N$. $\phi^t : I + u^t$, where I is the identity and u is the displacement vector field to deform pixels from a currently warped pCT $x_c^t \in \mathbb{R}^{L \times Q \times P}$ closer to the coordinates

of the CBCT image as x_c^{t+1}. No structure guidance from segmented pCT or CBCT is used and the RRN is optimized in an unsupervised manner using pairs of pCT and CBCT images $\{x_c, x_{cb}\}$.

The inputs to the RRN consist of channel-wise concatenated image pairs and hidden state, $\{x_c^{t-1}, x_{cb}, h_g^{t-1}\}$, where h_g^0 is initialized to 0 and $x_c^0 = x_c$ (Fig. 1). The intermediate CLSTM steps $t \ge 1$ receive the hidden state h_g^{t-1} produced from prior CLSTM step $t-1$. RRN at CLSTM step t outputs a warped pCT image x_c^t and the updated hidden state h_g^t. With $x_c^0 = x_c$, it's delineation, $y_c^0 = y_c$, and $\phi^t = g(x_c^{t-1}, x_{cb}, h_g^{t-1})$, the warped pCT and its delineation are computed as:

$$
\begin{aligned}
x_c^t &= x_c^{t-1} \circ \phi^t \\
y_c^t &= y_c^{t-1} \circ \phi^t.
\end{aligned}
\tag{2}
$$

RRN is optimized using image similarity L_{sim} and smoothness loss L_{smooth} measured from the flow field gradient, with a tradeoff parameter λ_{smooth} to control image similarity and deformation smoothness. L_{sim} is computed using Normalized Cross-Correlation (NCC) between the CBCT x_{cb} and the N warped pCTs x_c^t produced by CLSTM steps. NCC was computed locally using window of 5×5×5 centered on each voxel to improve robustness to CT and CBCT intensity differences [34]. L_{sim} was computed as:

$$
L_{sim} = -\sum_{t=1}^{N} L_{sim}^t = -\sum_{t=1}^{N} NCC(x_c^t, x_{cb}) \, / \, N,
\tag{3}
$$

$NCC(x_c^t, x_{cb})$ at each step t is an average of all local NCC calculations. The smoothness loss term is computed as:

$$
L_{smooth} = \sum_{t=1}^{N} L_{smooth}^t = \sum_{t=1}^{N} \sum_{p \in \Omega} \| \nabla \phi^t(p) \|^2 \, / \, N.
\tag{4}
$$

The total registration loss is computed as $L_{reg} = L_{sim} + \lambda_{smooth} \times L_{smooth}$.

C. One shot Patient-specific anatomic context and shape prior-aware CBCT segmentation

The CLSTM step t of RSN uses a channel-wise concatenated input $\{x_c^t, y_c^t, x_{cb}, h_s^t\}$ to compute a segmentation $y_{cb}^t = s(x_c^t, y_c^t, x_{cb}, h_s^t)$. x_c^t and y_c^t are produced by the RRN at CLSTM step t and h_s^t is the hidden state of s from CLSTM at time t, $1 \le t \le N$ (Eqn. 2). As shown in Fig. 1, the first CLSTM step of RSN t = 0 is initialized with a weak prior from undeformed pCT inputs $(x_c^0, y_c^0, x_{cb}, h_s^0)$. CBCT segmentation y_{cb} is produced after $N+1$ CLSTM steps.

In one-shot training, only a single exemplar segmented CBCT $\{x_{cb}^e, y_{cb}^e\}$ is available. RSN learns a mapping $\theta_S(.)$ of an image to it's segmentation, $s(x_c, y_c, x_{cb}^e) : \theta_s(x_{cb}^e) \rightarrow y_{cb}^e$. The segmentation loss is computed from segmentations computed in all CLSTM steps $0 \quad t \quad N$ of RSN as:

$$L_{seg} = \sum_{t=0}^{N} L_{seg}^t = \sum_{t=0}^{N} log P(y_{cb}^e \mid s(x_c^t, y_c^t, x_{cb}^e, h_s^t)).\tag{5}$$

The losses, $L_{seg}^0, \dots L_{seg}^{N-1}$ provide deep supervision to train RSN.

1) Online Hard Example Mining (OHEM) Loss: OHEM loss was previously used to improve stability of training in the presence of highly imbalanced classes in few-shot learning [17], [38]. Training stability is improved because the pixels considered for gradient computation change based on model output, and which acts as a form of online bootstrapping. OHEM loss focuses the network towards pixels that are hard to classify in a minibatch. Hard pixels are those that are associated with a small probability of producing the correct classification, or $p_{m,c} < \tau$, where $p_{m,c}$ is the probability associated to a class c for a pixel m, and τ is the probability threshold for selecting the hard pixels. We set $\tau = 0.7$ and $K = 10,000$, the minimum number of hard pixels to be used within each mini-batch. The OHEM loss is computed as:

$$L_{seg}^{ohem} = \sum_{t=0}^{N} \sum_{m=1}^{M} \sum_{c=1}^{C} 1\{p_{m,c} < \tau\} log \, P_{m,c}(y_{cb}^e \mid s(x_c^t, y_c^t, x_{cb}^e, h_s^t)),\tag{6}$$

where $\tau \in (0,1]$ is a threshold; $1\{*\}$ equals to one when the condition inside holds; C is the total class number; M is the total number of voxels inside one mini-batch.

Algorithm 1: One-shot PACS method.

---

input : Unlabeled CT and CBCT dataset $(x_c, x_{cb}) \in \{X_c, X_{cb}\}^K$,

Exemplar segmented CBCT and corresponding segmented

pCT, $(\{x_{cb}^e, y_{cb}^e, x_c, y_c\})^1 \in \{X_c, Y_c, X_{cb}, Y_{cb}\}^K$, K is the

number of pCT, CBCT pairs, 1 refers to a single example.

output: Registration model θ_g to align x_c to x_{cb} and segmentation

model θ_s to segment x_{cb}

1 Upsample one-shot CBCT example (x_{cb}^e, y_{cb}^e) and its corresponding

   pCT (x_c, y_c) to product $K/10$ examples;

2 $\theta_g, \theta_s \leftarrow$ initialize ;

3 **for** $Epoch\ id \leq Maximum\ Epoch$ **do**

4   **for** $Iter \leq Maximum\ Iter$ **do**

5   **if** $Registration\ Flag$ **then**

6              $x_c, x_{cb} \leftarrow$ sample mini-batch from $\{X_c, X_{cb}\}$;

7              $L_{reg} \leftarrow$ calculated using (3) (4);

8              $\theta_g \xleftarrow{+} -\Delta_{\theta_g}(L_{reg})$ (Gradient update);

9   **else**

10   load the example $(x_c, y_c, x_{cb}^e, y_{cb}^e)$;

11              $x_c^t, y_c^t \leftarrow$ calculated using (2) ;

12              $L_{seg}^{ohem} \leftarrow$ calculated using (6);

13              $\theta_s \xleftarrow{+} -\Delta_{\theta_s}(L_{seg}^{ohem})$ (Gradient update);

14   **end**

15   **end**

16 **end**

---

2) Joint registration-segmentation network optimization: The RRN network parameters are fixed when training the RSN and vice versa. The number of training examples available to optimize RRN $\mathscr{K} \gg 1$, where 1 is the number of examples available to optimize RSN. We replicated the one-shot exemplar $0.1 \times \mathscr{K}$ times through online augmentation to improve training stability. The number of replications was determined experimentally. This example replication is akin to upsampling (not to be confused with image upsampling) strategy used in machine learning for improving model generalizability. The training examples are shuffled to randomize the order of network updates. RRN g is updated using the gradient $-_{\theta_g}(L_{reg})$. RSN s is updated using the gradient $-\Delta_{\theta_s}(L_{seg}^{ohem})$.

The detailed training procedure for one-shot PACS registration-segmention method is in Algorithm 1.

D. Implementation details

All networks were implemented using Pytorch library and trained on Nvidia GTX V100 with 16 GB memory. The networks were optimized using ADAM algorithm with an initial learning rate of 2e-4 for the first 30 epochs and then decayed to 0 in the next 30 epochs and a batch size of 1. We set λ_{smooth}=30 experimentally. Eight CLSTM steps were used for both RRN and RSN. GPU memory limitation was addressed using truncated backpropagation through time (TBPTT) [39] after every 4 CLSTM steps.

The RSN was constructed with 3D Unet with the CLSTM placed on the encoder layers. Each convolutional block was composed of two convolution units, ReLU activation, and max-pooling layer. This resulted in feature sizes of 32,64,128,256, and 512. The RRN extended the Voxelmorph architecture [34] with CLSTM implemented in the encoder layers. Diffeomorphic deformation was ensured by using a diffeomorphic integration layer [40] following the 3-D flow field output of the CLSTM. The last layer of the RRN was composed of a spatial transformation function based on spatial transform networks [41] to convert the feature activations into DVF. The 3D networks architecture details are in the Supplementary document Table I and II.

IV.   Experiments and Results

A.   Dataset and Experiments:

A retrospective dataset of 369 fully anonymized weekly 4D-CBCT acquired from 65 patients with locally advanced non-small cell lung cancer and treated with intensity modulated radiotherapy using conventional fractionation with a single 4D pCT and up to 6 4D CBCTs acquired weekly during treatment were analyzed. Thirteen out of 65 patients were sourced from an external institution cohort [42]. Mid phase CTs and CBCTs were analyzed. The scans had an image resolution that ranged from 0.98 to 1.17mm in-plane and 3mm slice thickness. CBCT scans were acquired on a commercial CBCT scanner (On-board Imager^{TM}, Varian Medical Systems Inc,) using truebeam with (external: a peak kilovoltage (kVp) of 125kVp, tube current of 50 mAs; internal: 100kVp and 20 mAs) and reconstructed using Ram-Lak filter.

The open-source dataset [42] provided expert delineations. In the internal dataset, the gross tumor volume and the esophagus contours were produced on the pCT and CBCT by an experienced radiation oncologist and these represented the ground truth [5]. The esophagus contours were outlined on CBCT below the level of cricoid junction to the entrance of the stomach. CBCT and pCT scans were rigidly aligned using bony anatomy to bring them in the same spatial coordinates. FOV differences were addressed by resampling CBCT images to the same voxel size as the pCTs and the body mask was extracted through automatic thresholding (800HU) for soft tissue and the extracted region used as region of interest as done by other prior works [43], [44].

Metrics: Segmentation was evaluated using the Dice similarity coefficient (DSC), surface DSC (sDSC), and Hausdorff distance at 95th percentile (HD95) on testing set. The tolerance value of 4.38mm for computing sDSC was obtained using two physician segmentations

[5]. Inter-rater accuracy comparisons were done using the DSC metric. Registration was evaluated using segmentation accuracy, measures of deformation smoothness, namely, standard deviation of the Jacobian determinant and folding fraction ($|J_\phi|$ 0 (%)) [29], [30] computed from 95,551,488 voxels on the test set, and target registration error (TRE) using 3D scale invariant feature transform (SIFT) features [43], [45] identified on the pCT and CBCT [43], [44]. In the first step, keypoints are located by applying convolutions with the difference of Gaussians (DoG) function; in the second step, the feature descriptors consisting of 768 dimensional feature are constructed using geometric moment invariants to characterize the keypoints. Correspondences of the detected SIFT features (547 on average) in the pCT and CBCT scans were established using random sample consensus followed by manual verification [43]. On average, 22 corresponding features were used for TRE computation per image pair.

Experimental comparisons: One-shot PACS segmentations were compared against affine image registration, symmetric diffeomorphic registration (SyN) [26], deep learning segmentation only methods 3D Unet [46], Mask-RCNN [47], cascaded segmentation [48], and multiple deep registration based segmentation, Voxelmorph [34], recursive cascaded registration using 10 cascades [30], R2N2 [24], and coupled registration and segmentation network U-ResNet [49]. The Voxelmorph was regularized using segmentation losses from CBCT segmentations [34]. Full-shot PACS segmentation and full-shot PACS registration based segmentation were computed to establish upper bounds in accuracy and compare one-shot PACS segmentation against registration-based segmentation.

Statistical analysis : Statistical comparisons between one-shot PACS and other methods was done using the DSC metrics computed on the testing sets using pairwise, two-sided Wilcoxon signed rank tests at 95% significance level. The effect of treatment based tumor changes on the longitudinal accuracy of CBCT tumor segmentations was measured using one-way repeated measures ANOVA for DSC and HD95. Only $p < 0.05$ were considered to be significant.

Experiments: Separate networks are trained for tumor and esophagus segmentation, because tumors, which are abnormal structures are likely to have variable spatial and feature characteristics. Due to GPU limitation, the lung tumor segmentation model was computed from image volumes containing the entire thorax from the apex of lung to the level of diaphragm with a size of 192×192×60 obtained by resizing a volume of interest (VOI) of size 300×300×90. The full extent of the chest was visible on each slice. The esophagus model was computed from images of size 160×160×80 after resizing a VOI of size 256×256×110 enclosing the entire chest on each slice, starting from the cricoid junction till the entrance of the stomach. All methods were trained with 3-fold cross-validation using 9,800 VOI pairs obtained from 315 scans. The best model was applied on the independent test set of 54 CBCTs. One shot training was done using a randomly selected CBCT scan in the training set. Separately, robustness of tumor segmentation according to the choice of the one-shot CBCT example was evaluated using tumor location (apex: n = 109; inferior: n = 93; and middle: n = 80) and tumor size (small [5cc]: n = 54; medium [5cc to 10cc]: n = 80; and large [> 10cc]: n = 146) [50] as selection criteria. Separate models were trained

for each example and tested on a set aside data consisting of 28 apex, 29 inferior, 30 middle or centrally located tumors and 29 small, 28 medium, and 32 large tumors. Network design and ablation tests were performed to evaluate the impact of various losses, joint vs. two-step training, and the utility of CLSTM on tumor segmentation accuracy.

Network training convergence: Fig. 2 shows the training loss curves at every 500 iterations for the various losses when training the one-shot PACS network. As shown, the segmentation loss and NCC loss progressively decrease indicating training convergence. Increasing smoothness loss indicates increased image deformation.

B. Registration smoothness and accuracy

One-shot PACS produced the lowest TRE of 1.84±0.76 mm and smooth deformations that were within the accepted range of 1% of the folding fraction [29], [30] (Table. I). It required fewer parameters than recursive [30], but more than the R2N2 registration [24]. An example slice with corresponding SIFT features from warped pCT produced using different methods overlaid on the CBCT image is shown in Fig. 3.

Fig. 5 shows example registrations with the progressively changing DVFs and the warped pCTs produced by the CLSTM steps. A mirror flipped view of the pCT and it's corresponding CBCT before and after the registration, depicting the qualitative alignment of the images is shown. Brighter colored DVF curves correspond to the large deformations occurring in the regions corresponding to the shrinking tumor as well as the boundary of lung due to respiration differences. Registration performance for a representative case near descending aorta is shown in Supplementary Fig 6, which shows good alignment. Additional deformation results are in Supplementary Fig 1.

C. Segmentation accuracy

1) Tumor segmentation: Table II shows the segmentation accuracies produced using various methods. There was no difference in accuracy between the one-shot and full-shot PACS segmentation (DSC p=0.16). One-shot PACS segmentation was significantly more accurate (p<0.001) than all other methods, including the full-shot PACS registration-based segmentation (DSC p=1.2e-9). Fig. 4 shows segmentations produced by the various methods on randomly selected and representative cases from the external institution dataset. One shot PACS closely approximated expert's segmentations, despite imaging artifacts, indicating feasibility for tumor segmentation.

Inter-observer variability: Robustness to two rater tumor segmentations was measured for 9 patients. One-shot PACS produced a DSC of 0.82±0.08 and 0.84±0.09 for raters 1 and 2. The inter-rater DSC was 0.83±0.06. Fig. 6 shows three examples with one-shot PACS and two rater segmentations.

2) Esophagus segmentation: Table III shows the accuracies for segmenting the esophagus on CBCT images. One-shot PACS was similarly accurate as the full-shot PACS segmentation (DSC p=0.07). It was also significantly more accurate (p<0.001) than all other methods. Fig. 7 shows esophagus segmentation produced by various methods.

D. Longitudinal response assessment

The mean of maximum HD95 distance per patient from the weekly scans was 4.98mm. The median and inter-quartile range (IQR) of maximum HD95 from different patients were 4.97 mm and 3.90mm to 6.11mm. Longitudinal accuracy evaluation was done on 30 test patients who had CBCTs from all 6 weeks. The percent slope of DSC accuracy was −0.3% and HD95 was −8.4% using HD95 from week 1 to week 6 (Fig. 9). A one way repeated measures ANOVA with lower-bound corrections determined that CBCT tumor segmentations did not differ between weekly time points (DSC: $F_{(5, 1.35)} = 0.0033$, $p = 0.26$; HD95: $F_{(5, 0.56)} = 1.56$, $p = 0.46$). There was no significant interaction of tumor location and time on accuracy (DSC: $F_{(5, 0.82)} = 0.0002$, $p = 0.37$; HD95: $F_{(5, 1.89)} = 5.22$, $p = 0.18$). These results indicate that the one-shot PACS produced reliable segmentations on weekly CBCT. The one-way repeated measures ANOVA analysis for the esophagus also did not show a significant effect with time ($p = 0.26$). The longitudinal accuracy graphs for esophagus are in Supplementary Fig. 8.

Fig. 8 shows the representative example case with volumetric segmentations produced using one-shot PACS and the expert for tumor and esophagus on weekly scans. Our method closely followed the expert delineations.

E. Network design and ablation experiments

1) Robustness of one-shot tumor segmentation to selected CBCT training example: Kruskal-Wallis test showed no difference in the accuracy between one-shot and full shot models for tumor sizes (small: p=0.98; medium: p=0.62; large: p=0.73). Similarly, there was no significant difference between one-shot models trained with examples from different locations and the full-shot model (apex: p=0.29 ; middle: p=0.90; inferior: p=0.99). Summary of mean DSC accuracies as done in [17], produced by the various one-shot models tested on different locations (Fig. 10(a)) and sizes (Fig. 10(b)), shows similar accuracies for all models. Results for full-shot training is also shown for comparison. Larger variability in accuracy was seen for tumors abutting mediastinum (or middle) and small tumors for all models. Qualitative results on representative cases segmented with one-shot models trained with different tumor locations and sizes show good agreement between algorithm and expert (Fig. 11).

2) Impact of CLSTM in RRN: We compared the segmentation accuracy when the CLSTM was removed and implemented with convolutional layers, which converted it into a classical dynamic system (CDS) with shared weights for different steps [37]. Fig. 12 shows the feature activations (steps 1, 2, and 3) produced from layers 2, 3, and 4 (see Supplementary Table I) of the RRN that was trained as a CDS (Fig. 12(a)) and with CLSTM (Fig. 12(b)). The step outputs in the case of CLSTM correspond to hidden feature h_t of CLSTM, whereas for the CDS corresponds to the feature output after step t. Fig. 12 shows alignment of pCT with CBCT image with a centrally located shrinking tumor. Stronger feature activations with a consistent progression in the deformations around the tumor region are seen when using CLSTM (Fig. 12(b)) compared to the CDS network (Fig. 12(a)). Correspondingly, the mean feature activations in all the layers are higher in the CLSTM network (Fig. 12(c)). The CLSTM network also produced higher accuracy than the CDS

network trained without CLSTM (Fig. 12(d)). Concretely, the CDS network produced an accuracy of 0.78 ± 0.10, compared to the CLSTM network of 0.81 ± 0.08.

3) Impact of number of CLSTM steps on accuracy: We analyzed the accuracy and the computational times with increasing number of recurrent steps from 1 to 12. Segmentation accuracy increased and saturated beyond 8 steps (Supplementary Fig. 4). The computational times increased linearly. Therefore, we chose 8 CLSTM steps for our application. Fig. 13 shows the progressive improvement in the tumor and esophagus segmentations for a representative case from the various recurrent steps of the RSN trained using one-shot PACS-aware method. Moreover, one-shot PACS took 6.65 secs for training per iteration and 1.47 secs for testing per image pair.

4) Early vs. intermediate fusion of anatomic context and shape priors into RSN: The default PACS-aware approach combines progressively warped pCT and delineations as additional input channels with the CBCT image into the individual recurrent units placed in the encoder layers of RSN. We tested whether an intermediate fusion strategy, wherein separate encoders are used to compute features from CBCT and the final warped pCT and its delineation, and combined together at the decoder layer of RSN improved accuracy. The schematic of both methods is depicted in Supplementary Fig. 5. Our results showed that the intermediate fusion was less accurate (DSC of 0.72 ± 0.15 vs. 0.83 ± 0.08) than the default early fusion approach. This result indicates that combining the progressively warped pCT and its delineations with CBCT through the recurrent network improves accuracy.

5) Different weights for CLSTM steps: We studied whether assigning larger weights to segmentation losses from the later CLSTM steps had a greater impact on accuracy. For this purpose, the weights on the CLSTM steps in RSN were linearly increased ($w=t/(N+1)$). This approach had marginal impact on accuracy and resulted in a DSC accuracy of 0.82 ± 0.10 compared to 0.83 ± 0.08 for the default method.

6) Ablation experiments: We analyzed the accuracies when removing the different components of the one-shot PACS-aware network, including (I) shape context prior, (II) the anatomic context, (III) deep supervision of the segmentations produced from the intermediate recurrent steps of the segmentation network. We also measured the accuracy (IV) when CLSTM is removed from the segmentor s, (V) when training without the OHEM but with regular cross-entropy loss (Eqn. 5). The default one-shot PACS segmentor results are shown in (VI). As shown in Table IV, removing the shape context prior led to a clear lowering of accuracy indicating it's importance for one-shot segmentation training. The very low accuracy when removing shape context is because the reported results are for one-shot segmentor and not the registration-based segmentation. The shape context was also more relevant than the anatomic context with a significant difference ($p < 0.001$) in accuracies. Similarly, training without the OHEM loss led to lowering of segmentation accuracy. The segmentations produced using the afore-mentioned training scenarios for a representative performance is shown in Supplementary document Fig. 7.

7) **Influence of adding contour loss:** We measured the accuracy when a contour consistency loss was used to regularize the RRN by minimizing the difference in RRN generated and CBCT segmentation. This experiment was performed in the full-shot PACS mode as CBCT segmentations are needed for training both RRN and RSN. This approach produced a DSC of 0.84±0.08 from RSN and 0.81±0.08 from RRN-based segmentation, which is the same as the full-shot method trained without consistency loss, indicating equivalence of the two approaches in terms of accuracy. In contrast, one-shot PACS, which is similarly accurate as the full-shot method does, requires one segmented CBCT example for training.

8) **Joint training or separate training:** We tested whether optimizing the registration network first followed by segmentation network, as a two-step optimization, to provide shape and anatomic context to the segmentor improved accuracy over a jointly trained network. The two-step method produced a tumor segmentation accuracy of 0.81±0.09 DSC compared to 0.83 ± 0.08 using one-shot PACS and 0.82 ± 0.08 compared to 0.84±0.08 in the full-shot setting, indicating the multi-tasked approach is beneficial over two-step optimization.

V.   Discussion

We introduced a one-shot recurrent and joint registration-segmentation approach to longitudinally segment thoracic CBCT scans with large intra-thoracic changes occurring during radiotherapy. Our approach, which incorporates patient-specific anatomic context from higher contrast pCT and shape prior from delineated contours on pCT produced more accurate segmentations than multiple methods. Subset analysis showed that our approach was similarly accurate as two raters indicating feasibility of our approach to reduce inter-rater variability in CBCT segmentations. The shape context as well as the anatomic context prior were essential to improving segmentation network's accuracy in the one-shot setting as shown in the ablation experiments. Our approach was more accurate than cross-modality distillation [5], which incorporated MRI information for improving CBCT segmentation (DSC of 0.83 ± 0.08 using one-shot PACS vs. 0.73 ± 0.10 using MRI-based distillation) on the same dataset, underscoring the importance of spatial and anatomic priors for segmentation. Our method was similarly accurate as full-shot training for tumors, and robust to the chosen one-shot training example by location and size. However, larger variation in accuracies were observed for small tumors and centrally located tumors for all models including the full-shot model. Previously, we showed lowering of accuracy for centrally located tumors and smaller sized lung tumors [50] with standard CT scans. Addition of contour consistency loss in the registration did not improve accuracy, indicating that our one-shot PACS method is a reasonable alternative to full-shot training when large number of segmented examples are unavailable for training.

Furthermore, combining the RSN with RRN was significantly more accurate than RRN propagated segmentations, confirming prior findings [14]-[16], [28] that multi-tasked methods are more accurate than registration-based segmentation. We also found that the multi-tasked approach was more accurate than a two-step optimization.

Our approach handles large anatomical and appearance changes to diseased and healthy tissues during treatment by computing progressive deformations as a sequence by using a 3D CLSTM. CLSTM, which was introduced to model the sequential dynamics of 2D images [22], uses convolutions to compute a dense flow, which adds flexibility compared to parametric LSTM methods [23]. We found that our approach was significantly more accurate than R2N2 [24], even when it also computes progressive deformations. R2N2 [24] employs gated recurrent units with the local deformations computed using Gaussian basis functions, which was insufficient to handle the large anatomic changes common in longitudinal CBCT. Analysis of the recurrent component of our network by replacing the CLSTM with a standard classical system showed that CLSTM produced stronger and consistently progressing activations in local regions (e.g. tumors) undergoing large deformations. Also, the accuracy of the network without CLSTM was similar to the recursive [30], but the architectures are different. Our network uses a shared feature weights in all steps, whereas a recursive cascade [30] method uses different models trained jointly for the cascade steps. Finally, as shown in the ablation experiments, inclusion of CLSTM in the segmentation allowed the network to use progressively warped pCT (anatomic context) and pCT delineation (shape context) to improve accuracy further.

Finally, tumor segmentation using our method showed no significant changes in accuracy with treatment time (due to tumor shrinkage) or tumor location, indicating robustness of the approach for longitudinal response assessment.

We also evaluated our approach for esophagus segmentation. Our approach can be extended to simultaneously segment multiple organs by feeding the multi-channel organ probability map as shape prior and image as contextual prior to produce multi-channel output for segmentation [51].

CBCT images also have much lower FOV compared to the corresponding pCT, exacerbating the problem of robust registration. One prior approach by Zhou et.al [13] explicitly handled this issue by performing random crops of the pCT images for improving alignment. We like others [43], [44] handled this issue through pre-processing using extraction of chest region and resampling of CBCT and pCT images.

Our approach has the following limitations. We did not address the issue of motion averaging for precisely defining the gross tumor margin by aligning with all phases of CBCT acquired in a breathing cycle because the goal was longitudinal response assessment. Although our approach showed feasibility to segment tumors with similar variability as two raters, artifacts are not explicitly handled. Accuracy could be improved further by using SIFT features computed from gradients of the deep feature [19] or surface points [13] for registration. Our approach of computing dense flow fields may be adversely impacted especially for abdominal organs which have uniform density internally. In such cases, surface points as used in [13] or a MRI-based feature distillation [4], [5] could potentially be incorporated with the recurrent network formulation. Nonetheless, to our best knowledge, ours is the first to handle longitudinal segmentation of hard to segment lung tumors undergoing radiographic appearance and size changes from during treatment CBCTs.

VI. Conclusion

We introduced a one shot patient-specific anatomic context and shape prior aware multimodal recurrent registration-segmentation network for segmenting on treatment CBCTs. Our approach showed promising longitudinal segmentation performance for lung tumors undergoing treatment and the esophagus on one internal and one external institution dataset.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

REFERENCES

[1]. Sonke J-J, Aznar M, and Rasch C, "Adaptive radiotherapy for anatomical changes," Semin Radiat Oncol, vol. 29, no. 3, pp. 245–257, 2019. [PubMed: 31027642]

[2]. Wang J, Li J, Wang W, Qi H, Ma Z, Zhang Y et al. , "Detection of interfraction displacement and volume variance during radiotherapy of primary thoracic esophageal cancer based on repeated four-dimensional CT scans," Radiat Oncol, vol. 8, no. 224, 2013.

[3]. Kuckertz S, Papenberg N, Honegger J, Morgas T, Haas B, and Heldmann S, "Learning deformable image registration with structure guidance constraints for adaptive radiotherapy," in Biomedical Image Registration, 2020, pp. 44–53.

[4]. Fu Y, Lei Y, Wang T, Tian S, Patel P, Jani A, Curran W, Liu T, and Yang X, "Pelvic multi-organ segmentation on cone-beam CT for prostate adaptive radiotherapy," Med Phys, vol. 47, no. 8, pp. 3415–3422, 2020. [PubMed: 32323330]

[5]. Jiang J, Riyahi Alam S, Chen I, Zhang P, Rimner A, Deasy JO, and Veeraraghavan H, "Deep cross-modality (MR-CT) educed distillation learning for cone beam CT lung tumor segmentation," Med Phys, 2021.

[6]. Altorjai G, Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R, Georg D, and Dieckmann K, "Cone-beam CT-based delineation of stereotactic lung targets: The influence of image modality and target size on interobserver variability," Int. J Radiat Oncol Bio Phys, vol. 82, no. 2, pp. e265–e272, 2012. [PubMed: 21620581]

[7]. Haq R, Berry S, Deasy J, Hunt M, and Veeraraghavan H, "Dynamic multiatlas selection-based consensus segmentation of head and neck structures from CT images," Med Phys, vol. 46, no. 12, pp. 5612–5622, 2019. [PubMed: 31587300]

[8]. Dong P, Wang L, Lin W, Shen D, and Wu G, "Scalable joint segmentation and registration framework for infant brain images," Neurocomputing, vol. 229, pp. 54–62, 2017. [PubMed: 29416227]

[9]. Zhao A, Balakrishnan G, Durand F, Guttag JV, and Dalca AV, "Data augmentation using learned transformations for one-shot medical image segmentation," in CVPR, 2019, pp. 8535–8545.

[10]. He Y, Li T, Yang G, Kong Y, Chen Y, Shu H, Coatrieux J-L, Dillenseger J-L, and Li S, "Deep complementary joint model for complex scene registration and few-shot segmentation on medical images," in ECCV, vol. 1, 2020.

[11]. Wang S, Cao S, Wei D, Wang R, Ma K, Wang L, Meng D, and Zheng Y, "LT-Net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation," in CVPR, 2019. pp. 9162–9171.

[12]. Jia X, Wang S, Liang X, Balagopal A, Nguyen D et al., "Cone-beam computed tomography (cbct) segmentation by adversarial learning domain adaptation," in MICCAI, 2019, pp. 567–575.

[13]. Zhou B, Augenfeld Z, Chapiro J, Zhou SK, Liu C, and Duncan JS, "Anatomy-guided multimodal registration by learning segmentation without ground truth: Application to intraprocedural CBCT/MR liver segmentation and registration," Medical Image Anal., vol. 71, p. 102041, 2021.

[14]. Xu Z and Niethammer M, "Deepatlas: Joint semi-supervised learning of image registration and segmentation," in MICCAI, 2019, pp. 420–429.

[15]. Estienne T, Vakalopoulou M, Christodoulidis S, Battistela E, Lerousseau M, Carre A, Klausner G, Sun R, Robert C, and Mougiakakou S, "U-resnet: Ultimate coupling of registration and segmentation with deep nets," inInternational conference on medical image computing and computer-assisted intervention. Springer, 2019, pp. 310–319.

[16]. Beljaards L, Elmahdy MS, Verbeek F, and Staring M, "A crossstitch architecture for joint registration and segmentation in adaptive radiotherapy," in Med. Imaging with Deep Learning, 2020, pp. 62–74.

[17]. Cui H, Wei D, Ma K, Gu S, and Zheng Y, "A unified framework for generalized low-shot medical image segmentation with scarce data," IEEE Trans on Med Imaging, vol. 14, no. 8, 2020.

[18]. Foote MD, Zimmerman BE, Sawant A, and Joshi SC, "Real-time 2d-3d deformable registration with deep learning and application to lung radiotherapy targeting," in IPMI, vol. 11492, 2019, pp. 265–276.

[19]. Kearney V, Haaf S, Sudhyadhom A, Valdes G, and Soldberg T, "An unsupervised convolutional neural network-based algorithm for deformable image registration," Phys Med Biol, vol. 63, no. 18, p. 185017, 2018. [PubMed: 30109996]

[20]. Chitphakdithai N and Duncan JS, "Non-rigid registration with missing correspondences in preoperative and postresection brain images," in MICCAI, 2010, pp. 367–374.

[21]. Fu Y, Lei Y, Liu Y, Wang T, Curran WJ, Liu T, Patel P, and Yang X, "Cone-beam computed tomography (CBCT) and CT image registration aided by CBCT-based synthetic CT," in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 11313, 2020, p. 113132U.

[22]. Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K, and W.-c. Woo, "Convolutional LS TM network: A machine learning approach for precipitation nowcasting," arXiv preprint arXiv:1506.04214, 2015.

[23]. Wright R, Khanal B, Gomez A, Skelton E, Matthew J, Hajnal JV, Rueckert D, and Schnabel JA, "Lstm spatial co-transformer networks for registration of 3d fetal us and mr brain images," in Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis, 2018, pp. 149–159.

[24]. Sandkühler R, Andermatt S, Bauman G, Nyilas S, Jud C, and Cattin PC, "Recurrent registration neural networks for deformable image registration," NeurIPS, vol. 32, pp. 8758–8768, 2019.

[25]. Wang S, Cao S, Wei D, Xie C, Ma K, Wang L, Meng D, and Zheng Y, "Alternative baselines for low-shot 3d medical image segmentation—an atlas perspective," AAAI, 2021.

[26]. Avants BB, Epstein CL, Grossman M, and Gee JC, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," Medical Image Anal., vol. 12, no. 1, pp. 26–41, 2008.

[27]. Brock K, Mutic S, McNutt T, Li H, and Kessler M, "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM radiation therapy committee task group no. 132," Med Phys, vol. 44, no. 7, pp. e43–e76, 2017. [PubMed: 28376237]

[28]. Elmahdy M, Jagt T, R.T Z, Qiao Y, Shahzad R et al. , "Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer," Med Phys, vol. 46, no. 8, pp. 3329–3343, 2019. [PubMed: 31111962]

[29]. de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, and Išgum I, "A deep learning framework for unsupervised affine and deformable image registration," Medical Image Anal., vol. 52, pp. 128–143, 2019.

[30]. Zhao S, Dong Y, Chang EI, Xu Y et al., "Recursive cascaded networks for unsupervised medical image registration," in CVPR, 2019, pp. 10600–10610.

[31]. Lee MCH, Petersen K, Pawlowski N, Glocker B, and Schaap M, "Tetris: Template transformer networks for image segmentation with shape priors," IEEE Trans Med Imaging, vol. 38, no. 11, pp. 2596–2606, 2019. [PubMed: 30908196]

[32]. Dalca AV, Guttag J, and Sabuncu MR, "Anatomical priors in convolutional networks for unsupervised biomedical segmentation," in IEEE CVPR, 2018, pp. 9290–9299.

[33]. Zhou Y, Li Z, Bai S, Chen X, Han M, Wang C, Fishman E, and Yuille A, "Prior-aware neural network for partially-supervised multiorgan segmentation," in IEEE ICCV, 2019, pp. 10671–10680.

[34]. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, and Dalca AV, "Voxelmorph: a learning framework for deformable medical image registration," IEEE Trans. Med Imaging, vol. 38, no. 8, pp. 1788–1800, 2019.

[35]. Zhang W, Yan P, and Li X, "Estimating patient-specific shape prior for medical image segmentation," in ISBI. IEEE, 2011, pp. 1451–1454.

[36]. Lee MC, Oktay O, Schuh A, Schaap M, and Glocker B, "Image-and-spatial transformer networks for structure-guided image registration," in MICCAI, 2019, pp. 337–345.

[37]. Goodfellow I, Bengio Y, and Courville A, Deep Learning. MIT Press, 2016, http://www.deeplearningbook.org.

[38]. Wu Z, Shen C, and Hengel v. d., Anton, "High-performance semantic segmentation using very deep fully convolutional networks," arXiv preprint arXiv:1604.04339, 2016.

[39]. Jaeger H, Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the" echo state network" approach. GMD-Forschungszentrum Informationstechnik Bonn, 2002, vol. 5, no. 01.

[40]. Dalca AV, Balakrishnan G, Guttag J, and Sabuncu MR, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," Medical Image Anal., vol. 57, pp. 226–236, 2019.

[41]. Jaderberg M, Simonyan K, Zisserman A, and Kavukcuoglu K, "Spatial transformer networks," arXiv preprint arXiv:1506.02025, 2015.

[42]. Hugo GD, Weiss E, Sleeman WC, Balik S, Keall PJ, Lu J, and Williamson JF, "A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer," Med Phys, vol. 44, p. 762–771, 2017. [PubMed: 27991677]

[43]. Park S, Plishker W, Quon H, Wong J, Shekhar R, and Lee J, "Deformable registration of ct and cone-beam ct with local intensity matching," Physics in Medicine & Biology, vol. 62, no. 3, p. 927, 2017. [PubMed: 28074785]

[44]. Landry G, Nijhuis R, Dedes G, Handrack J, Thieke C, Janssens G, Orban de Xivry J, Reiner M, Kamp F, Wilkens JJ et al. , "Investigating CT to CBCT image registration for head and neck proton therapy as a tool for daily dose recalculation," Med Phys, vol. 42, no. 3, pp. 1354–1366, 2015. [PubMed: 25735290]

[45]. Rister B, Horowitz MA, and Rubin DL, "Volumetric image registration from invariant keypoints," IEEE Trans Image Processing, vol. 26, no. 10, pp. 4900–4910, 2017.

[46]. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, and Ronneberger O, "3d U-Net: learning dense volumetric segmentation from sparse annotation," in MICCAI, 2016, pp. 424–432.

[47]. He K, Gkioxari G, Dollár P, and Girshick R, "Mask R-CNN," in Proc. IEEE ICCV, 2017, pp. 2961–2969.

[48]. Cao Z, Yu B, Lei B, Ying H, Zhang X, Chen DZ, and Wu J, "Cascaded se-resunet for segmentation of thoracic organs at risk," Neurocomputing, vol. 453, pp. 357–368, 2021.

[49]. Estienne T, Vakalopoulou M, Christodoulidis S, Battistella E, Lerousseau M et al., "U-ReSNet: Ultimate coupling of registration and segmentation with deep nets," in MICCAI, 2019, pp. 310–319.

[50]. Jiang J, Hu Y-C, Liu C-J, Halpenny D, Hellmann MD, Deasy JO, Mageras G, and Veeraraghavan H, "Multiple resolution residually connected feature streams for automatic lung tumor segmentation from ct images," IEEE Trans Med Imaging, vol. 38, no. 1, pp. 134–144, 2018. [PubMed: 30040632]

[51]. Jiang J, Hu Y-C, Tyagi N, Rimner A, Lee N, Deasy JO, Berry S, and Veeraraghavan H, "Psigan: joint probabilistic segmentation and image distribution matching for unpaired cross-modality adaptation-based mri segmentation," IEEE Trans Med Imaging, vol. 39, no. 12, pp. 4071–4084, 2020. [PubMed: 32746148]
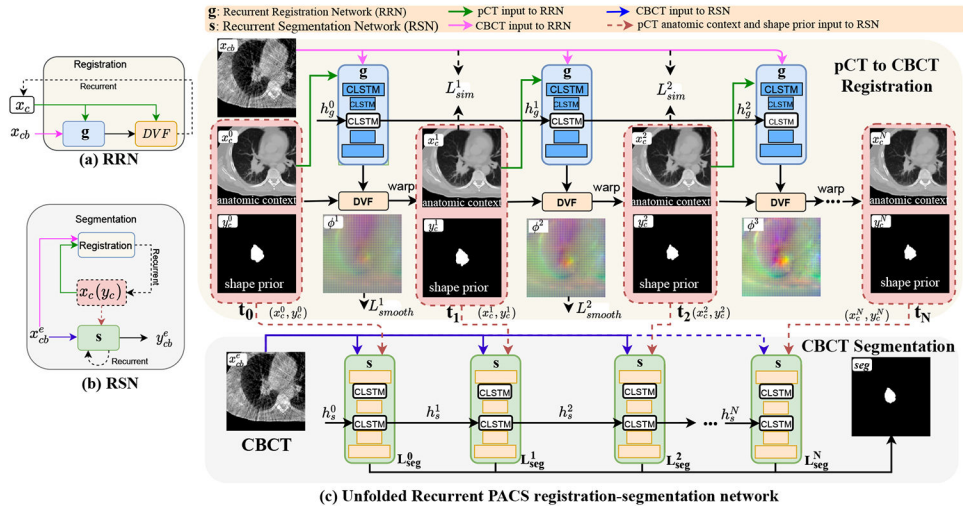
Fig. 1.

One-shot PACS registration-segmentation network. RRN g uses N CLSTM steps to align x_c to x_{cb}. RSN s uses N+1 CLSTM steps to segment x_{cb}, where $\{x_c^t, y_c^t\}$, $1 \le t \le N$ produced by RRN are used in the RSN to provide spatial and anatomic priors.
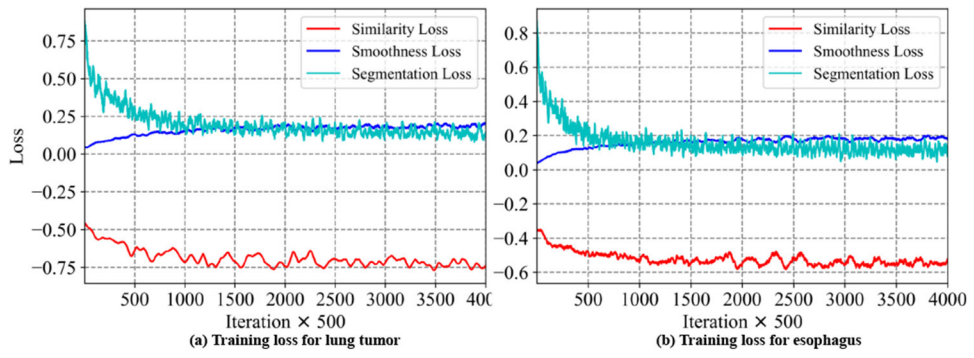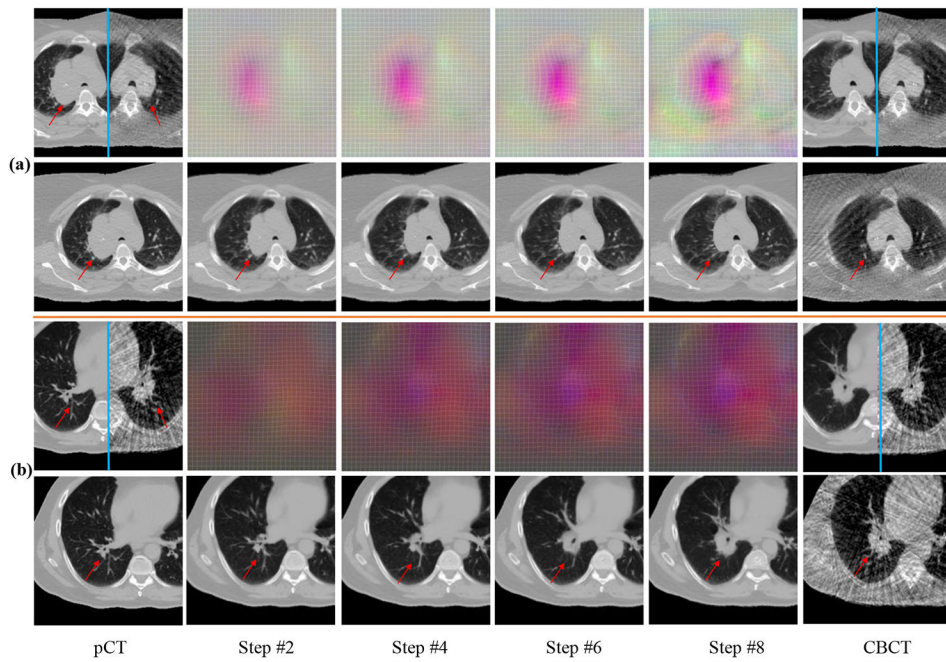
**Fig. 2.**

Training loss curves for one-shot PACS method.
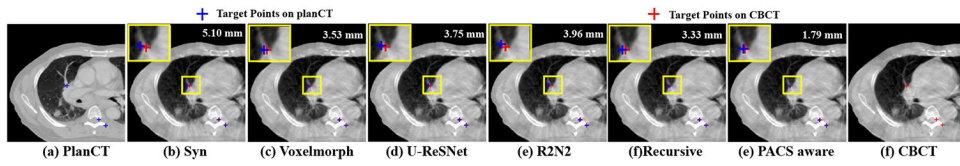
Fig. 3.

SIFT detected targets and the corresponding deformed targets produced by various methods overlaid on CBCT.
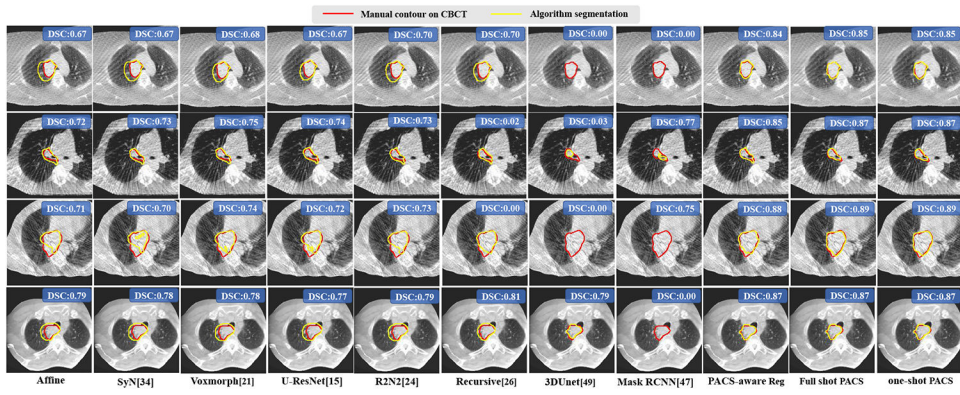
Fig. 4.

Tumor segmentation from CBCT produced by various methods. DSC accuracies for the volume are also shown.

Fig. 5.
Progressive DVFs with warped pCTs (rows 2 , 4) for a shrinking tumor (row 1), and out of plane rotation (row 3). Mirror flipped view of pCT and the CBCT before and after alignment are shown. DVF colors indicate displacements in x (0mm to 10.13mm) (black to red),y (0mm to 7.76mm) (black to green), and z (0mm to 13.50mm) (black to blue) directions. Red arrow identifies the tumor.
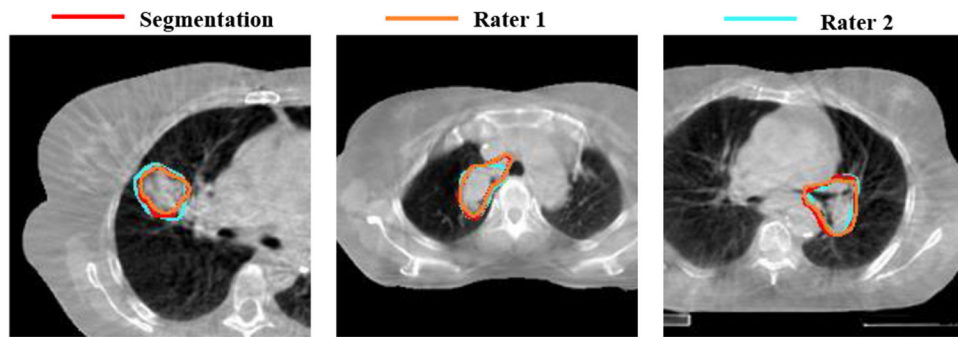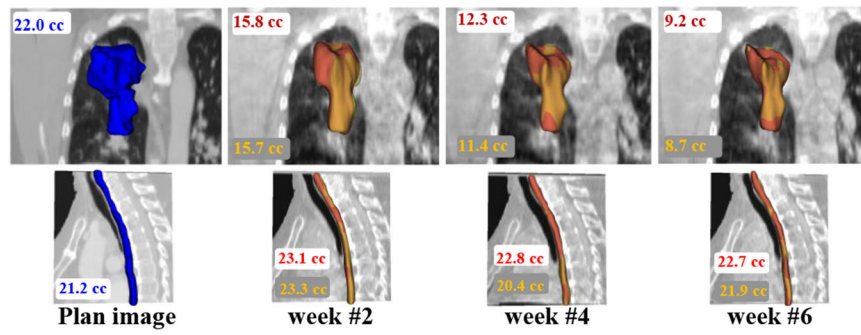
Fig. 6.
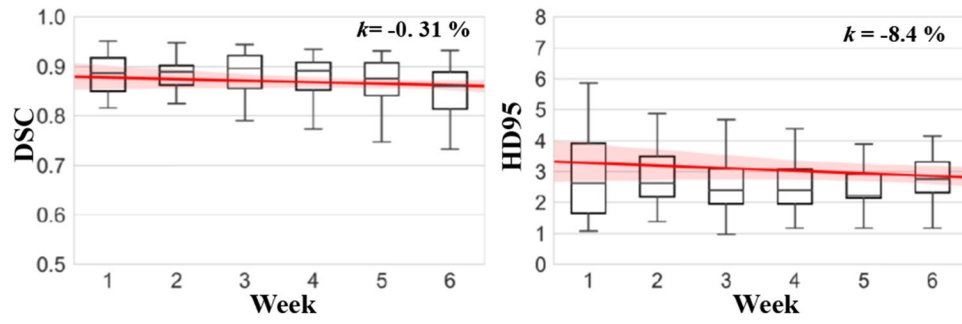One-shot PACS segmentation compared to two raters.
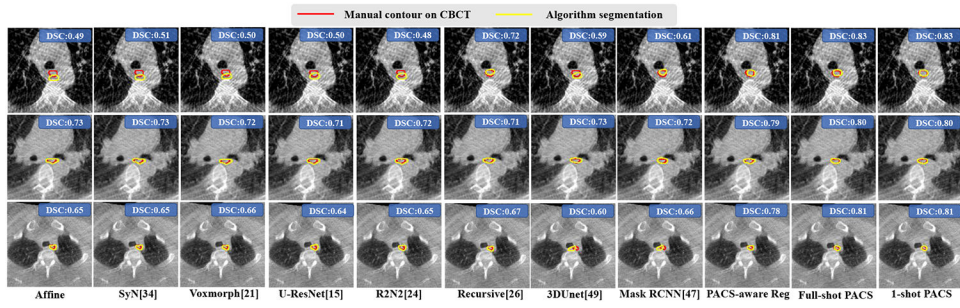
Fig. 7.

Esophagus segmentation from CBCT produced by various methods. Volumetric DSC accuracies are also shown.

Fig. 8.
Longitudinal segmentation using algorithm (red) and expert (yellow) on weekly CBCT for tumor (top row) and the esophagus (bottom row). The pCT delineation is shown in blue.
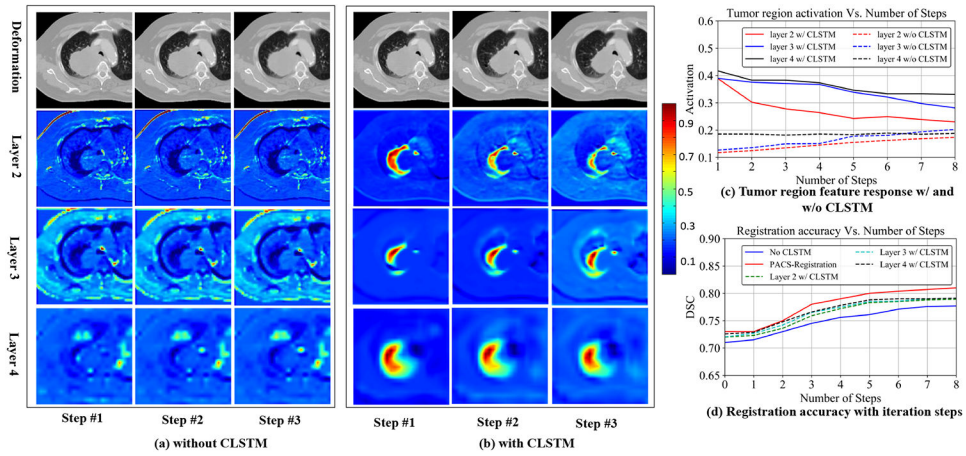
Fig. 9.
Segmentation accuracy at different weeks with percent slope change in accuracy for DSC and HD95 metrics.
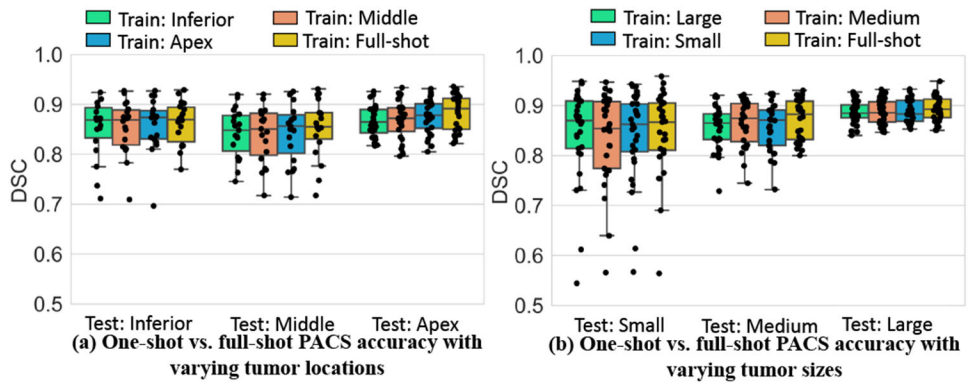
Fig. 10.
Testing set segmentation accuracy with models trained using different one-shot examples by (a) location and (b) size.
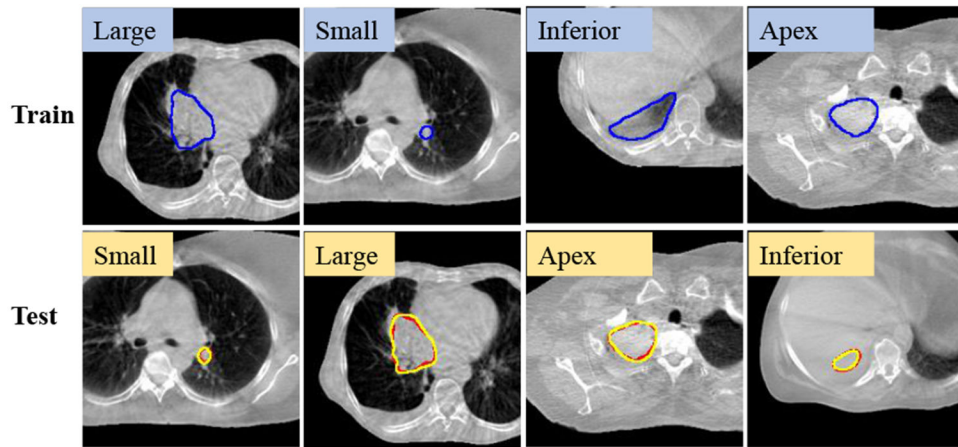
Fig. 11.

Example one-shot PACS segmentor (yellow) results trained with different sizes and locations. Red: expert contour.

Fig. 12.
(a) Feature activations produced in CNN layers 2, 3, and 4 for steps 1, 2, 3 without CLSTM, and (b) with CLSTM. (c) shows mean feature activations in the layers 2, 3, and 4. (d) shows DSC accuracy with increasing number of recurrent steps.
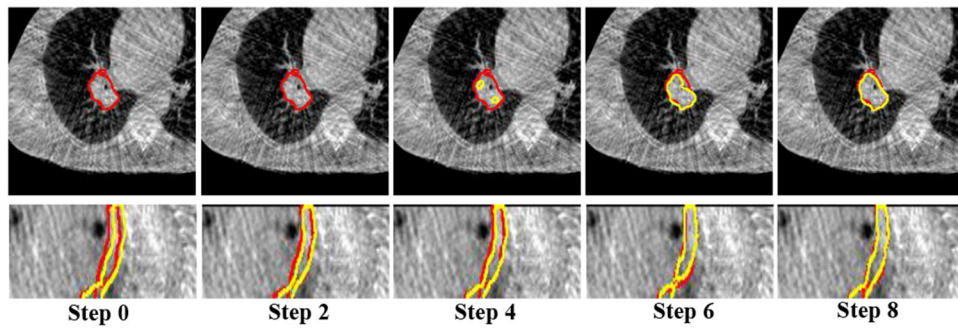
Fig. 13.
Segmentations produced using the one-shot PACS segmentor with increasing number of CLSTM steps for tumor and esophagus. Red is expert, yellow is algorithm contour.

TABLE I

Registration metrics of various methods.

Method	SD Jacobian	J_ϕ	0 (%)	Parameters	TRE (mm)
SyN [26]	0.04±0.01		0.0022±0.0066	N/A	3.94±1.55
Voxelmorph [34]	0.05±0.01		0.0042±0.011	301,411	3.13±1.50
Recursive [30]	0.08±0.02		0.013±0.030	42,418,491	2.77±1.53
U-ReSNet [15]	0.04±0.01		0.021±0.015	4,753,035	3.45±1.62
R2N2 [24]	0.04±0.01		0.039±0.012	39,183	3.25±1.91
PACS-aware	0.13± 0.02		0.020±0.037	522,723	1.84±0.76

TABLE II

Tumor segmentation accuracies produced by various methods. Reg - registration, Seg - segmentation.

Method	Testing (Number=54)		
	DSC	**sDSC**	**HD95** *mm*
Affine Reg	0.71±0.14	0.87±0.14	7.31±3.65
SyN [26]	0.72±0.14	0.88±0.13	7.03±3.52
Voxelmorph [34]	0.75±0.13	0.92±0.11	5.62±3.05
R2N2 [24]	0.74±0.13	0.91±0.10	6.12±2.85
U-ReSNet [15]	0.73±0.14	0.90±0.12	6.44±3.33
Recursive [30]	0.77±0.11	0.93±0.08	5.38±2.57
3D Unet [46]	0.61±0.15	0.83±0.15	16.72±23.50
Mask RCNN [47]	0.64±0.16	0.82±0.14	20.53±23.29
Cascaded Net [48]	0.63±0.16	0.81±0.14	22.61±23.42
PACS-aware Reg	0.81±0.08	0.97±0.05	4.15±1.82
Full-shot PACS seg	**0.84±0.08**	**0.98±0.04**	**3.33±2.02**
One-shot PACS seg	**0.83±0.08**	**0.97±0.06**	**3.97±3.06**

TABLE III

Esophagus segmentation accuracies. Reg - registration, Seg - segmentation.

Method	Testing (Number=54)		
	DSC	**sDSC**	**HD95** *mm*
Affine Reg	0.67±0.16	0.77±0.18	5.11±3.38
SyN [26]	0.69±0.17	0.79±0.19	4.96±3.43
Voxelmorph [34]	0.72±0.15	0.83±0.17	4.40±2.89
R2N2 [24]	0.73±0.15	0.84±0.17	4.33±3.14
U-ReSNet [15]	0.72±0.15	0.83±0.17	4.47±3.20
Recursive [30]	0.73±0.15	0.85±0.16	4.11 ±2.15
3D Unet [46]	0.57±0.18	0.64±0.17	6.79±2.87
Mask RCNN [47]	0.61±0.17	0.68±0.16	6.67±2.67
Cascaded Net [48]	0.60±0.15	0.66±0.15	7.58±2.38
PACS-aware reg	0.76±0.12	0.88±0.13	3.88±2.83
Full shot PACS seg	**0.79±0.13**	**0.91±0.12**	**3.10±2.16**
One shot PACS seg	**0.78±0.13**	**0.90±0.14**	**3.22±2.02**

TABLE IV

Ablation experiments. DS: Deep supervision

	Tumor Segmentation						
	pCT	Shape	CBCT	DS	CLSTM Seg	OHEM	DSC
I	✓	×	✓	✓	✓	✓	0.02±0.00
II	×	✓	✓	✓	✓	✓	0.79±0.11
III	✓	✓	✓	×	✓	✓	0.80±0.10
IV	✓	✓	✓	✓	×	✓	0.81±0.11
V	✓	✓	✓	✓	✓	×	0.80±0.11
VI	✓	✓	✓	✓	✓	✓	**0.83±0.08**