Medical Image Analysis 68 (2021) 101909

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

DeepTarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy

Dakai Jin[a,1,*], Dazhou Guo[a,1], Tsung-Ying Ho[b,*], Adam P. Harrison[a], Jing Xiao[c], Chen-kan Tseng[b], Le Lu[a]

[a] PAII Inc., Bethesda, MD, USA
[b] Chang Gung Memorial Hospital, Linkou, Taiwan, ROC
[c] Ping An Technology, Shenzhen, Guangdong, China

ARTICLE INFO

Article history:
Received 23 March 2020
Revised 10 September 2020
Accepted 13 November 2020
Available online 19 November 2020

Keywords:
Esophageal cancer
Radiotherapy
Gross tumor volume
Clinical target volume
RTCT
PET/CT
Multi-modality fusion
Segmentation
Delineation
Distance transform

ABSTRACT

Gross tumor volume (GTV) and clinical target volume (CTV) delineation are two critical steps in the cancer radiotherapy planning. GTV defines the primary treatment area of the gross tumor, while CTV outlines the sub-clinical malignant disease. Automatic GTV and CTV segmentation are both challenging for distinct reasons: GTV segmentation relies on the radiotherapy computed tomography (RTCT) image appearance, which suffers from poor contrast with the surrounding tissues, while CTV delineation relies on a mixture of predefined and judgement-based margins. High intra- and inter-user variability makes this a particularly difficult task. We develop tailored methods solving each task in the esophageal cancer radiotherapy, together leading to a comprehensive solution for the target contouring task. Specifically, we integrate the RTCT and positron emission tomography (PET) modalities together into a two-stream chained deep fusion framework taking advantage of both modalities to facilitate more accurate GTV segmentation. For CTV segmentation, since it is highly context-dependent—it must encompass the GTV and involved lymph nodes while also avoiding excessive exposure to the organs at risk—we formulate it as a deep contextual appearance-based problem using encoded spatial distances of these anatomical structures. This better emulates the margin- and appearance-based CTV delineation performed by oncologists. Adding to our contributions, for the GTV segmentation we propose a simple yet effective progressive semantically-nested network (PSNN) backbone that outperforms more complicated models. Our work is the first to provide a comprehensive solution for the esophageal GTV and CTV segmentation in radiotherapy planning. Extensive 4-fold cross-validation on 148 esophageal cancer patients, the largest analysis to date, was carried out for both tasks. The results demonstrate that our GTV and CTV segmentation approaches significantly improve the performance over previous state-of-the-art works, e.g., by 8.7% increases in Dice score (DSC) and 32.9 mm reduction in Hausdorff distance (HD) for GTV segmentation, and by 3.4% increases in DSC and 29.4 mm reduction in HD for CTV segmentation.

1. Introduction

Esophageal cancer ranks sixth in mortality amongst all cancers worldwide, accounting for 1 in 20 cancer deaths (Bray et al., 2018). As it is usually diagnosed at late stages, radiotherapy (RT) is often one of the primary treatment (Pennathur et al., 2013). The most critical and challenging tasks in RT planning are the gross tumor volume (GTV) and clinical target volume (CTV) delineations, where high radiation doses are applied to those regions to kill cancer cells (Burnet et al., 2004). As shown in Fig. 1, GTV and CTV are correlated yet different regions. While the GTV represents the visible gross tumor region, the CTV outlines the area that covers the microscopic tumorous region, i.e., sub-clinical disease. Spatially, CTV boundaries must contain the GTV and also any involved lymph nodes (LNs). Physician's delineation principles of GTV and CTV are quite different. The determinant of GTV mainly relies on image appearance clues. Yet, the estimation of CTV requires to first have the GTV, followed by measuring the sub-clinical disease margins through the judgment of both image appearance and spatial distances from the GTV and other involved targets, i.e., LN and organ at risk (OAR). Current clinical protocols rely on manual GTV and CTV delineation, which is time and labor consuming and subject to high inter- and intra-observer variability

[*] Corresponding authors.
E-mail addresses: jindakai376@paii-labs.com (D. Jin), tyho@cgmh.org.tw (T.-Y. Ho).
[1] Co-first author.

https://doi.org/10.1016/j.media.2020.101909
1361-8415/© 2020 Elsevier B.V. All rights reserved.

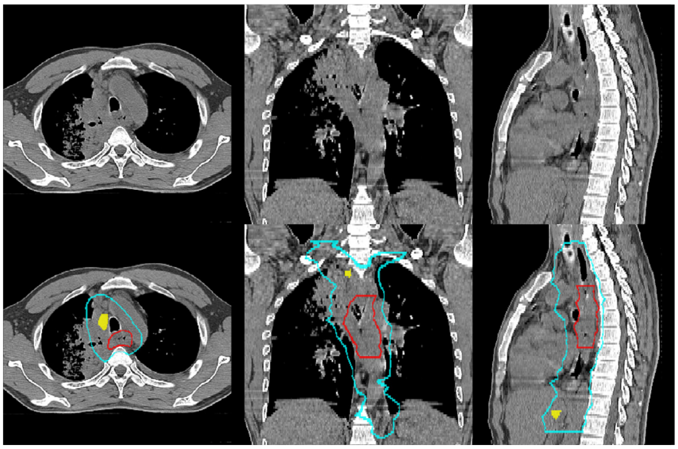

Fig. 1. An example of esophageal cancer GTV and CTV contours in axial, coronal and sagittal views (top: original CT, bottom: contours overlaid on CT). Red, and cyan indicate the GTV and CTV contours, respectively. Yellow represents the involved LNs. Note that CTV margins must cover the GTV, microscopic tumorous regions, and any involved LNs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(Tai et al., 1998; Eminowicz and McCormack, 2015; Nowee et al., 2019). This motivates automated approaches for GTV and CTV segmentation, which could potentially increase the target contouring accuracy and consistency, as well as significantly shorten the planning time allowing for timely treatment.

Both GTV and CTV delineation offer their own distinct challenges. The assessment of esophageal GTV by radiotherapy computed tomography (RTCT) alone has been shown to be error prone, due to the poor contrast between the GTV and surrounding tissues (Muijs et al., 2010). Within the clinic, these shortfalls are often addressed by correlating with the patient's positron emission tomography/computed tomography (PET/CT) scan, when available. These PET/CTs are taken on an earlier occasion to help stage the cancer and decide treatment protocols. Despite misalignments between the PET/CT and RTCT, positron emission tomography (PET) still provides highly useful information to help manually delineate the GTV on the RTCT, thanks to its high contrast in highlighting malignant regions (Leong et al., 2006). As shown in Fig. 2, RTCT and PET can each be crucial for accurate GTV delineation, due to their complementary strengths and weaknesses. Yet, leveraging both diagnostic PET and RTCT requires contending with the unavoidable misalignments between the two scans acquired at different times.

Turning to CTV delineation, its quality depends highly on physician's experience due to its judgement-based characteristics. For esophageal cancer, this is even more challenging because tumors may potentially spread along the entire esophagus and metastasize up to the neck or down to the upper abdominal LNs, and present adjacent to several OARs, such as the lung (Jin et al., 2018) and airway (Jin et al., 2017). Recent works on automated CTV segmentation mostly operate based on the RTCT appearance alone (Men et al., 2017; 2018; Wong et al., 2020). However, as shown in Fig. 3, CTV delineation depends on the radiation oncologist's visual judgment of both the appearance *and* the spatial configuration of the GTV, LNs, and OARs, suggesting that only considering the RTCT makes the problem ill-posed (Men et al., 2017; 2018; Wong et al., 2020).

By considering their different characteristics and challenges, we propose tailored methods for GTV and CTV delineation to solve each task. Together, these result in a combined workflow that provides a comprehensive solution (named as DeepTarget) for target contouring in esophageal cancer radiotherapy. Specifically, there are four major contributions in our work.

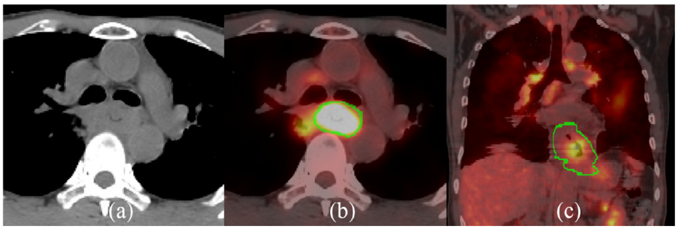

Fig. 2. Esophageal GTV examples in RTCT and PET images, where the green line indicates the ground truth boundary. (a)-(b): although the GTV boundaries are hardly distinguishable in RTCT, it can be reasonably inferred with the help of PET. (c) PET can be noisy with false positive high-uptake regions, as well as the limited uptakes in the GTV. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1. For the GTV segmentation, we introduce a new two-stream chained deep network fusion method to incorporate the joint RTCT and PET information for accurate esophageal GTV segmentation (see Fig. 6). One of the streams is trained using the RTCT alone, while the other stream uses both RTCT and registered PET. The former exploits the anatomical appearance features in computed tomography (CT), while the latter takes advantage of PET's sensitive, but sometimes spurious and overpoweringly strong contrast. The two streams explore tumor characteristics from different perspectives, hence, their predictions can be further deeply fused with the original RTCT to generate a final robust GTV prediction. The misalignment between RTCT and PET/CT is alleviated by a deformable registration with a robust anatomy-guided initialization.
2. For the GTV segmentation, we also introduce a simple yet surprisingly powerful progressive semantically-nested network (PSNN) segmentation model, which incorporates the strengths of both UNet (Ronneberger et al., 2015) and PHNN (Harrison et al., 2017) by using deep supervision to progressively propagate high-level semantic features to lower-level, but higher resolution features. The PSNN achieves superior performance in the tumor segmentation task as compared to prior arts, e.g., DenseUNet (Yousefi et al., 2018) and PHNN (Harrison et al., 2017).
3. For the CTV segmentation, we introduce a novel spatial context encoded deep CTV delineation framework. Instead of expecting the CNN to learn distance-based margins from the GTV, LNs, and OARs binary masks, we provide the CTV delineation network with the 3D signed distance transform maps (SDMs) of these structures. Specifically, we include the SDMs of the GTV, LNs, lung, heart and spinal canal with the original RTCT volume as inputs to the network. From a clinical perspective, this allows the CNN to emulate the oncologists manual delineation, which uses the distances of GTV, LNs vs. the OARs as a key constraint in determine the CTV boundaries.
4. We demonstrate in the extensive experiments that both our GTV and CTV segmentation methods can significantly improve the performance over prior state-of-the-art (SOTA): 8.7% increase in absolute Dice score (DSC) (from 70.3% to 79.0%) and 8.5 mm reduction in average surface distance (ASD) (from 14.2 mm to 5.7 mm) as compared to Yousefi et al. (2018) for GTV segmentation, and 3.4% increase in DSC (from 79.2% to 82.6%) and 3.3 mm reduction in ASD (from 7.7 mm to 4.4 mm) as compared to Cardenas et al. (2018a) for CTV segmentation.

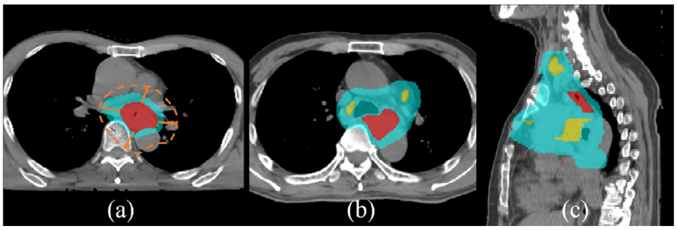

Fig. 3. Esophageal CTV delineation illustration, where red, yellow, and cyan indicate the GTV, involved LNs and CTV, respectively. (a) shows that the CTV is not a uniform margin expansion (brown-dotted line) from the GTV, while (b)-(c) shows how delineation becomes more complicated when involved LNs are present. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The initial results of this work were presented in two conference papers each focusing on separately GTV Jin et al. (2019a) and CTV Jin et al. (2019b) segmentation. The current manuscript extends the two previous works in three aspects.

1. We provide a more comprehensive literature review for the GTV and CTV segmentation works, as well as for PET/CT co-segmentations, which could not be included in our conference papers because of page limits.
2. We expand our esophageal RT dataset to 148 patients with paired PET/CT and RTCT images from the original 110. We conduct extensive 4-fold cross-validation for both GTV and CTV segmentation using the same splits at the patient level. For the GTV experiment, we additionally compared against three recent SOTA PET/CT co-segmentation methods (Zhong et al., 2019; Zhao et al., 2019; Kumar et al., 2020).
3. We integrate the GTV and CTV segmentation together, reporting, for the first time, a combined and more complete esophageal target contouring workflow. In doing so, we study the impact of using our GTV predictions as input into the CTV task, characterizing the performance of this integrated workflow and demonstrating its potential clinical value.

2. Related work

2.1. GTV segmentation

A handful of studies have addressed automated esophageal GTV segmentation (Hao et al., 2017; Tan et al., 2017; Yousefi et al., 2018). Tan et al. (2017) developed an adaptive region-growing algorithm with a maximum curvature strategy to segment the esophageal tumor in PET alone and evaluated using phantom images. However, due to the misalignments between PET and RTCT and their different imaging principles, even a dedicated cross-modality registration algorithm may not achieves satisfactory results (Mattes et al., 2003). Hao et al. (2017); Yousefi et al., 2018 developed esophageal GTV segmentation models for RTCT using 2D and 3D convolutional neural networks (CNNs), respectively. To compensate for 2D CNN limitations, Hao et al. (2017) applied the graph cut algorithm (Boykov et al., 2001) to further refine the segmentation results. In contrast, Yousefi et al., 2018 designed a 3D UNet (Çiçek et al., 2016) equipped with densely connected convolutional modules (Huang et al., 2017), i.e., DenseUNet, and achieved the SOTA esophageal GTV segmentation performance. GTV segmentation has also been extensively studied in other cancers, such as lung cancer (Lian et al., 2019; Zhong et al., 2019; Zhao et al., 2019; Kumar et al., 2020) and head & neck cancer (Guo et al., 2019; Lin et al., 2019; Ma et al., 2019), where CT, PET or MRI modalities are combined or separately adopted depending on specific tasks. Our work differs from the prior art by proposing a two-stream deep fusion strategy to fully leverage the complementary information from the CT and PET modalities, which we show is superior to both using CT in isolation and combining CT and PET in co-segmentation approaches (discussed below).

2.2. CT/PET tumor co-segmentation

CT/PET based tumor co-segmentation has been a popular topic in recent years and is closely relevant to our GTV two-stream deep fusion method. Huang et al., 2018; Guo et al. (2019) applied CT/PET early fusion to segment the GTV in head & neck cancer, where UNet or Densely connected UNet are adopted as the segmentation networks. Xu et al. (2018) designed a cascaded pipeline to sequentially use CT and PET images for the whole-body bone lesion segmentation. Zhong et al. (2019); Zhao et al. (2019); Kumar et al. (2020) designed different CT/PET fusion methods to segment lung tumors and achieved encouraging results. Zhong et al. (2019) proposed a segmentation method using two coupled UNets (Çiçek et al., 2016), one that takes CT alone as input and one that takes PET alone, where each UNet shares connections with each other to promote the complementary information. Zhao et al. (2019) trained two parallel VNets (Milletari et al., 2016) to extract the PET- and CT-based probability maps where a final fusion module produced the final segmentation. Kumar et al. (2020) introduced a co-learning module to derive a spatially varying fusion map that quantifies the relative importance of each modality, which is then multiplied with PET and CT specific features to obtain a representation of the complementary CT/PET information at different image locations. We compared our tumor segmentation with these prior arts (Zhong et al., 2019; Zhao et al., 2019; Kumar et al., 2020) to demonstrate the effectiveness of our proposed CT/PET fusion strategy. These prior works all assume that the PET and CT images are well-aligned, which may not be true in practice.

2.3. CTV segmentation

No prior work, CNN-based or not, has been reported for esophageal cancer CTV segmentation. Studies on CTV segmentation of other cancer types mostly focus on the RTCT appearance alone. Yang et al. (2014) applied an atlas-based method to segment the CTVs of head & neck cancer, where the atlas was built based on a dual-force demons registration algorithm (Pennec et al., 1999). Men et al. (2017, 2018) applied different deep network structures in a fully convolutional network (CNN) framework, such as dilated convolution (Yu and Koltun, 2016) or ResNet (He et al., 2016), to directly segment the CTV from RTCT in rectal and breast cancer, respectively. Recently, Cardenas et al. (2018b) used the pure distance information from the GTV and several OARs to conduct voxel-wise classification to identify if a voxel belongs to the oropharyngeal CTV, demonstrating considerable improvement compared to pure CT appearance based methods. Notably, Cardenas et al. (2018a) later showed that considering the GTV and LN binary masks together with the RTCT can achieve the SOTA oropharyngeal CTV delineation performance. However, binary masks do not explicitly provide distances to the model. In contrast, we explicitly provide spatial distance information from the GTV, involved LNs and OARs to segment the CTV. This provides much greater context for the network to exploit.

3. Methods

The overall workflow of our DeepTarget system is depicted in Fig. 4, which consists of three major components: (1) image preprocessing to register PET to RTCT and to perform prerequisite anatomy segmentation in RTCT, i.e., the involved LN

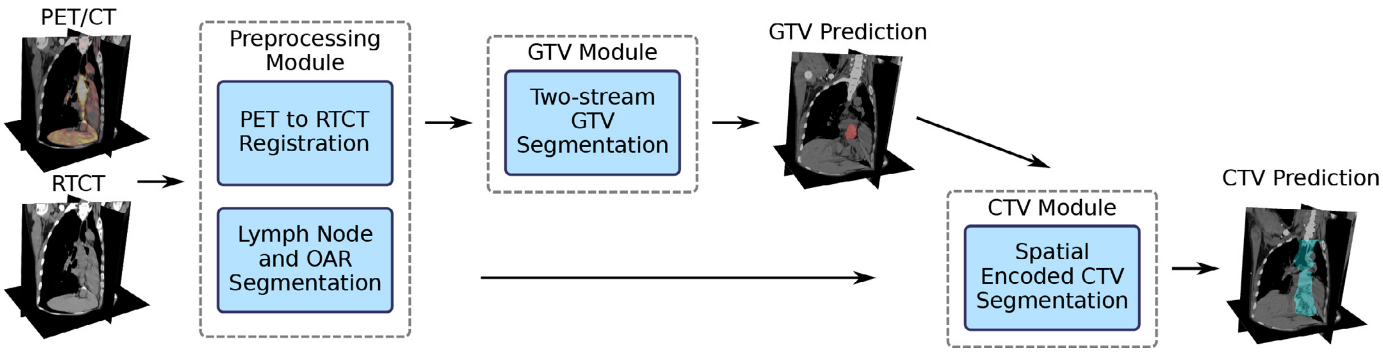

Fig. 4. The overall workflow of our DeepTarget system that segment both GTV and CTV in esophageal cancer radiotherapy.

and OAR segmentation; (2) GTV segmentation using a two-stream chained 3D deep fusion method and the new proposed progressive semantically-nested network (PSNN); (3) CTV segmentation using a deep contextual- and appearance-based method, involving the RTCT and the 3D spatial distance maps from the GTV, involved LNs and OARs. Our workflow can accept different backbones. In the results section, we demonstrate that PSNN is an effective architecture for the appearance-based GTV segmentation. However, for the CTV segmentation different networks have limited impact since CTV delineation relies mainly on the problem formulation (appearance-based or spatial context-based).

3.1. Image preprocessing

In our system, preprocessing consists of PET to RTCT registration and also the segmentation of prerequisite anatomical structures in RTCT.

3.1.1. PET to RTCT registration

To effectively leverage the complementary information from PET and RTCT we generate an aligned PET/RTCT pair by registering the former to the latter. Direct PET to RTCT registration suffers from large errors due to their completely different imaging principles (Mattes et al., 2003). To overcome this difficulty, we register the diagnostic CT (accompanying the PET) to RTCT and use the resulting deformation field to align the PET to RTCT. This intra-patient image registration from same imaging modality is a much easier task, where many deformable registration methods have demonstrated very good performance (Murphy et al., 2011).

To register the diagnostic CT to RTCT we apply a cubic B-spline based free-form deformable registration algorithm (Rueckert et al., 1999) in a coarse to fine multi-scale deformation process implemented by Klein et al. (2010). We choose this option due to its good capacity for shape modeling and efficiency in capturing local non-rigid motions. However, to perform well, the registration algorithm must have a robust rigid initialization to manage patient pose, scanning range, and respiratory differences in the two CT scans (Fig. 5(a)). To accomplish this, we use the 3D lung mass centers from the two CT scans as the initial matching positions. We compute these mass centers from mask predictions produced by our prerequisite region segmentation step (see Section 3.1.2). This leads to a reliable initial matching for the chest and upper abdominal regions, helping the success of the registration. The resulting deformation field is then applied to the diagnostic PET to align it to the RTCT at the planning stage. One registration example is shown in Fig. 5(b).

3.1.2. Prerequisite anatomy segmentation

Our CTV segmentation module relies on incorporating the spatial context/distance of the anatomical structures of interest, i.e., the GTV, involved LNs and OARs. We assume manual segmentations for the involved LNs are available, considering that even the most recent LN detection performance is prone to large false-positive rates (Zhu et al., 2020; Yan et al., 2020). We leave the metastasis LN identification itself as a separate research topic. For the OARs, we do not make this assumption. Indeed, missing OAR segmentations (25%) is common in our dataset. We consider three major organs: the lung, heart, and spinal canal, since most esophageal CTVs are closely integrated with them. Unlike densely-distributed and large amounts of OARs in the head and neck region (Guo et al., 2020), the three organs we considered are easier to segment. Hence, we simply trained a 2D PSNN model to segment them using available organ labels in our dataset. Robust performance is achieved with validation Dice scores for the lung, heart and spinal canal as high as 97%, 95% and 80%, respectively.

In contrast to the OARs, we chose to use the ground truth GTVs in the CTV segmentation training phase. This is because segmenting tumors is a much harder problem than segmenting healthy organs. Although our esophageal GTV segmentation method produces significant improvements as compared to previous SOTA, the automated GTV predictions may still present errors that harm the CTV segmentation training stability. However, during inference, we do not make this assumption. As shown in the results Section 5.3, we extensively studied the CTV segmentation performance by using the predicted GTVs from different methods and demonstrate that our CTV method performs well regardless of the origin of GTV predictions.

3.2. Esophageal GTV segmentation

Fig. 6 depicts an overview of the proposed two-steam chained 3D deep fusion method for esophageal GTV segmentation. As mentioned, we aim to effectively exploit the complementary information within the PET and CT images. Assuming N data instances, we denote the training data as $S^{\mathrm{GTV}} = \left\{\left(X_n^{\mathrm{CT}}, X_n^{\mathrm{PET}}, Y_n^{\mathrm{GTV}}\right)\right\}_{n=1}^{N}$, where X_n^{CT}, X_n^{PET}, and Y_n^{GTV} represent the input RTCT, registered PET, and binary ground truth GTV images, respectively. For the segmentation backbone, we use our new PSNN model, which is described in Section 3.2.3.

3.2.1. CT stream and early fusion stream

Dropping n for clarity and using j to denote voxel index, we first use two separate streams to generate segmentation maps using X^{CT} and $[X^{\mathrm{CT}}, X^{\mathrm{PET}}]$ as network input channels:

$$\hat{y}_j^{\mathrm{CT}} = p_j^{\mathrm{CT}}\left(y_j^{\mathrm{GTV}} = 1 | X^{\mathrm{CT}}; \mathbf{W}^{\mathrm{CT}}\right), \tag{1}$$

$$\hat{y}_j^{\mathrm{EF}} = p_j^{\mathrm{EF}}\left(y_j^{\mathrm{GTV}} = 1 | X^{\mathrm{CT}}, X^{\mathrm{PET}}; \mathbf{W}^{\mathrm{EF}}\right), \tag{2}$$

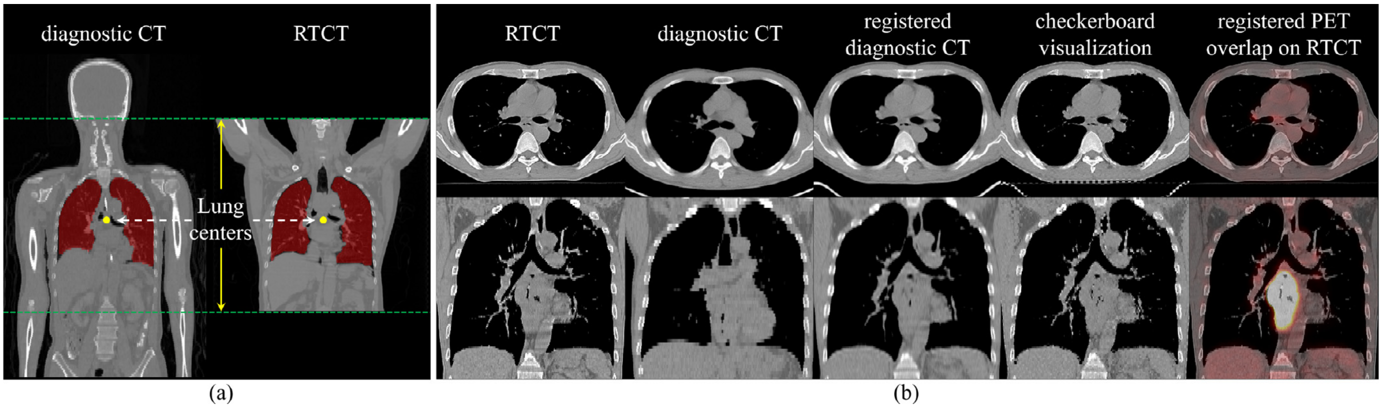

Fig. 5. (a) Illustration of our rigid initialization using the anatomy-based criteria. Note the considerable patient pose differences between diagnostic CT and RTCT, which necessitates a robust rigid deformation initialization. (b) shows an example of the deformable registration results for a patient in axial and coronal views. From left to right are the RTCT image; the diagnostic CT image before and after the registration, respectively; a checkerboard visualization of the RTCT and registered diagnostic CT images; and finally the overlapped PET image, transformed using the diagnostic CT deformation field, on top of the RTCT.

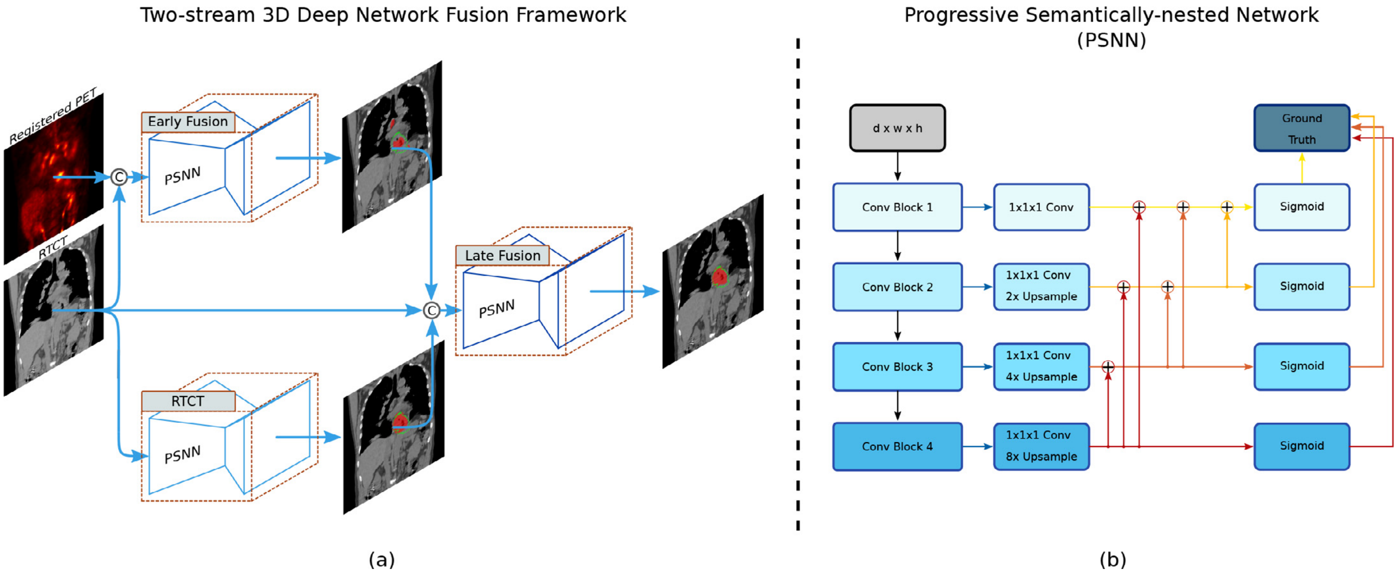

Fig. 6. (a) depicts our two-stream chained esophageal GTV segmentation method consisting of early fusion (EF) and late fusion (LF) networks. The © symbol denotes concatentation along the image channel. Although 3D inputs are used, we depict 2D images for clarity. (b) illustrates our proposed 3D PSNN model, which employs deep supervision at different scales within a parameter-less high-to-low level image segmentation decoder. In the implementation, we use four 3D convolutional blocks. The first two and last two blocks are composed of two and three 3 × 3 × 3 convolutional + BN + ReLU layers, respectively.

where $p_j^{(\cdot)}(\cdot)$ and $\hat{y}_j^{(\cdot)}$ denote the CNN functions and output segmentation maps, respectively, $\mathbf{W}^{(\cdot)}$ represents the corresponding CNN parameters, and y_j indicates the ground truth GTV binary values. Eq. (2) can be seen as an (EF) of CT and PET, taking advantage of the high spatial resolution in CT and high tumor-intake contrast properties in PET, respectively. On the other hand, the CT stream in Eq. (1) provides predictions based on CT appearance alone, which can be particularly helpful in circumventing the biased influence from noisy non-malignant high uptake regions, which are not uncommon in PET.

3.2.2. Two-stream chained late fusion

As Fig. 6(a) illustrates, we harmonize the outputs from Eqs. (1) and (2) by concatenating them together with the original RTCT as the inputs to a third network:

$$\hat{y}_j^{\mathrm{LF}} = p_j^{\mathrm{LF}}\left(y_j^{\mathrm{GTV}} = 1|X^{\mathrm{CT}}, \hat{Y}^{\mathrm{CT}}, \hat{Y}^{\mathrm{EF}}; \mathbf{W}^{\mathrm{CT}}, \mathbf{W}^{\mathrm{EF}}, \mathbf{W}^{\mathrm{LF}}\right). \qquad (3)$$

In this way, the formulation in (3) can be seen as a (LF) of the aforementioned two streams of the CT and EF models. We use the Dice score (DSC) loss to first train the CT, EF, and LF models separately. Then, we combine them all and conduct the fine-tuning end-to-end until convergence (see implementation details in Section 4.2).

3.2.3. PSNN segmentation model

For conventional segmentation tasks in medical image, e.g., the tumor or organs, various networks often exhibit different performance. The difficulty mainly lies in the poor appearance contrast of targets. Networks equipped with high abstraction ability and well-handling of the resolutions often achieve better performance. This is actually aligned with physicians' reasoning process, whose delineation heavily rely upon high-level semantic context to locate the target and disambiguate the boundaries. In certain respects, the UNet (Ronneberger et al., 2015) architecture follows this intuition, where it gradually decodes high-level features *down* to lower-level space. Nonetheless, the decoding path in UNet consumes a great deal of parameters, adding to its complexity. On the other hand, models like progressive holistically-nested network (PHNN) (Harrison et al., 2017; George et al., 2017) use deep supervision (Lee, Xie, Gallagher, Zhang, Tu, 2015) to connect lower

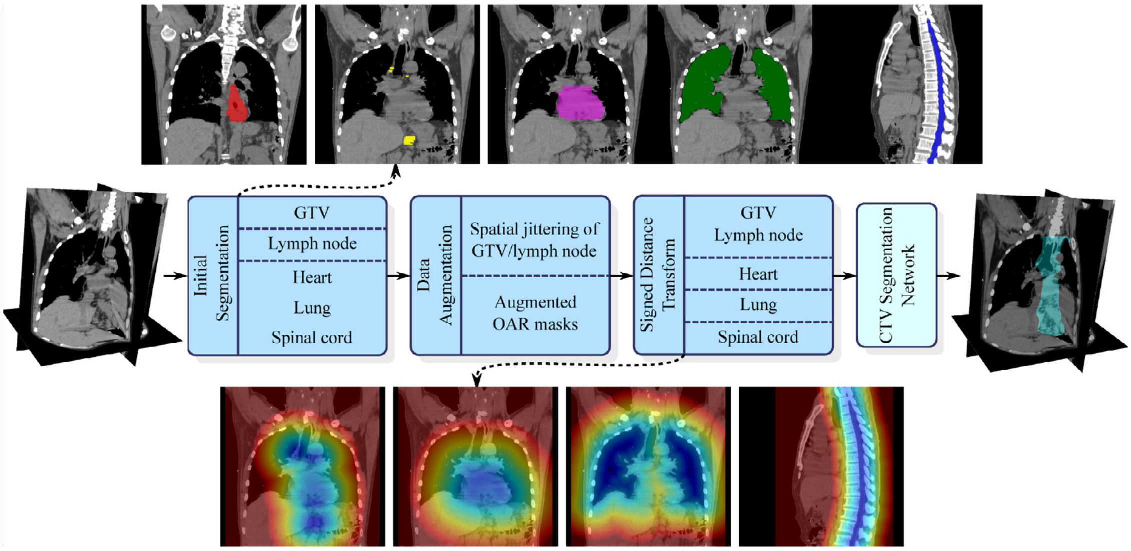

Fig. 7. Overall workflow of our spatial context encoded CTV delineation method. The top and bottom rows depict various masks and SDMs, respectively, overlayed on the RTCT. From left to right are the GTV, LNs, heart, lung, and spinal canal. The GTV and LNs share a combined SDM.

and higher-level features together using parameter-less pathways. However, unlike UNet, PHNN propagates lower-level features *up* to high-level layers, which is counter-intuitive. Instead, a natural and simple means to combine the strengths of both PHNN and UNet is to use essentially the same parameter-less blocks as PHNN, but reverse the direction of the deeply-supervised pathways, to allow high-level information to propagate *down* to lower-level space. We denote such a backbone progressive semantically-nested networks (PSNNs), indicating the progressive aggregation of higher-level semantic features down to lower space in a deeply-supervised setup.

As shown in Fig. 6(b), a set of 1 × 1 × 1 3D convolutional layers are used to collapse the feature map after each convolutional block into a logit image, i.e., $\tilde{f}_j^{(\ell)}$. This is then combined with the previous higher level segmentation logit image to create an aggregated segmentation map, i.e., $f_j^{(\ell)}$, for the ℓth feature block by element-wise summation:

$$f_j^{(m)} = \tilde{f}_j^{(m)}, \tag{4}$$

$$f_j^{(\ell)} = \tilde{f}_j^{(\ell)} + g\left(f_j^{(\ell+1)}\right), \forall \ell \in \{m-1, \dots 1\}, \tag{5}$$

where m denotes the total number of predicted feature maps and $g(\cdot)$. denotes an upsampling, i.e., trilinear upsampling. Like PHNN, the logits of all m convolutional blocks are trained using deeply-supervised auxiliary losses (Lee et al., 2015). As our experiments will demonstrate, PSNN can provide significant performance gains for GTV segmentation over both a densely connected version of UNet and PHNN.

3.3. Esophageal CTV segmentation

CTV delineation in RT planning is essentially a margin expansion process, starting from visible tumorous regions (the GTV and involved LNs) and extending into the neighboring regions by considering the possible tumor spread range and the distances to nearby healthy OARs. Fig. 7 depicts an overview of our CTV segmentation method, which consists of four major modularized components: (1) segmentation of prerequisite regions (described in image preprocessing Section 3.1.2); (2) SDM computation; (3) domain-specific data augmentation; and (4) a 3D deep network to execute the CTV delineation.

3.3.1. SDM computation

To encode the spatial context with respect to the GTV, involved LNs, and OARs, we compute 3D (SDMs) for each. Here, we combine GTV and LN masks together. In an SDM the value of each voxel measures the distance to the closest object boundary. Voxels inside and outside the object boundary have positive and negative values, respectively. More formally, let $\mathcal{O}_i$ denote a binary mask, where $i \in$ {GTV+LNs, lung, heart, spinal canal} and let $\Gamma(\cdot)$ be a function that computes boundary voxels of a binary image. The SDM value at a voxel v with respect to $\mathcal{O}_i$ is computed as

$$\mathrm{SDM}_{\Gamma(\mathcal{O}_i)}(v) = \begin{cases} \min\limits_{u \in \Gamma(\mathcal{O}_i)} d(v,u) & \text{if} \quad v \notin \mathcal{O}_i \\ -\min\limits_{u \in \Gamma(\mathcal{O}_i)} d(v,u) & \text{if} \quad v \in \mathcal{O}_i \end{cases}, \tag{6}$$

where $d(v,u)$ is a distance measure from v to u. We choose to use Euclidean distance in our work and use Maurer et al. (2003)'s efficient algorithm to compute the SDMs. The bottom row in Fig. 7 depicts example SDMs for the combined GTV and LNs and the three OARs. Note that we compute SDMs separately for each of the three OARs, meaning we can capture each organ's influence on the CTV.

3.3.2. Domain-specific data augmentation

We adopt specialized data augmentations to increase the robustness of the training and harden our network to noise in the prerequisite segmentations. Specifically, two types of data augmentations are carried out. (1) We spatially jitter the LN and GTV SDMs, by random shifts within $4 \times 4 \times 4\,\mathrm{mm}^3$, mimicking that in practice 4 mm average distance error represents the SOTA performance in automated esophageal GTV segmentation (Jin et al., 2019a). (2) We calculate SDMs of the OARs using both the manual annotations and the automatic segmentations from Section 3.1.2. We randomly choose to use or forego each augmentation strategy, leading to four possible combinations. This increases model robustness.

3.3.3. Execution of CTV segmentation

The CTV segmentation network takes the RTCT and SDMs of the GTV and LNs, and three OARs as inputs, which allows it to more easily infer the mixture of appearance and distance-based margins. We denote the training data as $S^{\mathrm{CTV}} = \left\{\left(X_n^{\mathrm{CT}}, \mathrm{SDM}_{\Gamma(\mathcal{O})_n}, Y_n^{\mathrm{CTV}}\right)\right\}_{n=1}^{N}$, where X_n^{CT}, $\mathrm{SDM}_{\Gamma(\mathcal{O})_n}$, and Y_n^{CTV} represent the input RTCT, SDMs,

Table 1
Demographic, clinical and tumor characteristics of 148 esophageal cancer patients. Note that some patients have tumors located across different esophagus region, hence, the total number summed at various tumor locations is greater than 148. * indicates values are presented as median [interquartile range, 25th–75th percentile].

Characteristics	Entire cohort ($n = 148$)
Sex	
Male	135 (91%)
Female	13 (9%)
Age at diagnosis*	55 [50–61]
T stage	
T1	0 (0%)
T2	24 (16%)
T3	71 (48%)
T4	53 (36%)
Tumor location	
Cervical	11 (7%)
Upper third	26 (18%)
Middle third	84 (57%)
Lower third	69 (47%)

and binary ground truth CTV mask, respectively. The CTV network function can be represented as

$$\hat{y}_j^{\mathrm{CTV}} = p_j^{\mathrm{CTV}}\left(y_j^{\mathrm{CTV}} = 1 | X^{\mathrm{CT}}, \mathrm{SDM}_{\Gamma(\mathcal{O})}, \mathbf{W}^{\mathrm{CTV}}\right), \tag{7}$$

where $\mathbf{W}^{\mathrm{CTV}}$ represents the corresponding CTV network parameters. We use the DSC loss to train the CTV segmentation network. As shown in the results Section 5.2, directly providing the spatial distances from the GTV, LNs, and OARs is essential to improve the CTV segmentation performance. Instead, different networks exhibit similar performance. This indeed confirms our observation that difficulty of CTV segmentation lies in the high-level problem formulation (appearance-based or spatial context-based) rather than the choice of networks. This is in contrast to the conventional appearance based tumor or organ segmentation, where network architecture often have a non-trivial impact on the performance.

4. Experimental

4.1. Dataset and evaluation

Dataset: To evaluate performance, we collected a dataset containing 148 esophageal cancer patients from Chang Gung Memorial Hospital, whose demographic, clinical and tumor characteristics are shown in Table 1. Each patient has a diagnostic PET/CT pair and a treatment RTCT scan and underwent the concurrent chemoradiatioan therapy (CCRT). To the best of our knowledge, this is the largest dataset for esophageal GTV and CTV segmentation to date. All 3D GTV, CTV, and the involved LN ground truth masks were delineated and confirmed by two experienced radiation oncologists during routine clinical workflow. In addition, OAR masks, generated during RT planning, were frequently available as well. We used these latter OAR masks to train the prerequisite segmentation networks of Section 3.1.2.

Evaluation: Extensive 4-fold cross validation, split at the patient level, was conducted for both GTV and CTV segmentation. We report the segmentation performance using three quantitative metrics: Dice score (DSC) in percentage, Hausdorff distance (HD) and average surface distance (ASD) in millimeters. Together, these metrics provide a comprehensive evaluation.

4.2. Implementation details

Training data sampling: Both GTV and CTV segmentation are conducted using 3D networks. We first resample all input imaging volumes of registered PET and RTCT and label images to a fixed resolution of 1.0 × 1.0 × 2.5 mm. Then, we apply a windowing of [200, 300] HU to every RTCT volume to cover the intensity range of soft tissues. We use a 3D volume of interest (VOI) patch-based fashion to train both GTV and CTV networks. To generate the 3D training samples, we extract 96 × 96 × 64 sub-volumes in two manners: (1) To ensure enough VOIs with postive GTV and CTV content, we randomly extract VOIs centered within the ground truth GTV and CTV masks, respectively. (2) To obtain sufficient negative examples, we randomly sample 20 VOIs from the whole volume. This results, on average, in 80 training VOIs per patient for both GTV and CTV segmentation. We further apply random rotations in the x-y plane within ±10 degrees to augment the training data. During inference, 3D sliding windows with sub-volumes of 96 × 96 × 64 and strides of 64 × 64 × 32 voxels are used. The probability maps of sub-volumes are aggregated to obtain the whole volume prediction. Using a single Titan-V GPU, it takes on average 20s and 6s to segment one input volume for GTV and CTV, respectively.

Network optimization: For both GTV and CTV segmentation tasks, the Adam solver (Kingma and Ba, 2014) is used to optimize all the segmentation models with a momentum of 0.99 and a weight decay of 0.005. Batch size is set to 24 for all models. For our GTV segmentation model, we train the CT and EF stream sub-networks for 50 epochs to convergence, and the LF sub-network for 30 epochs. Then, we combine all three sub-networks and fine-tune them end-to-end for 10 epochs to generate our final GTV segmentations. For other CT/PET co-segmentation methods, we follow the implementation details as described in their papers. For all CTV segmentation models, we train them for 40 epochs to convergence.

4.3. Experimental design

Esophageal GTV segmentation: **First**, we conduct a detailed ablation study to evaluate the effectiveness of our two-stream chained deep fusion method, as well as the proposed PSNN model. To do that, we compare our PSNN model versus PHNN (Harrison et al., 2017) and DenseUNet (Yousefi et al., 2018) under all three imaging settings, i.e., the CT, (EF), and (LF) of Eqs. (1)–(3), respectively. Note that the DenseUNet arguably represents the current SOTA esophageal GTV segmentation approach using CT. **Second**, equipped with the PSNN model, we further compare our CT/PET fusion method with three other SOTA CT/PET co-segmentation methods (Zhong et al., 2019; Zhao et al., 2019; Kumar et al., 2020).

Esophageal CTV segmentation: **First**, we compare our CTV segmentation method against 3 other setups to validate the effectiveness of the proposed appearance + spatial context formulation: (1) using only the CT appearance (Men et al., 2017; 2018); (2) using the CT and binary GTV/LN masks (Cardenas et al., 2018a); (3) using the CT + GTV/LN SDMs, which does not consider the effects of OARs. We compare these CTV setups using the proposed PSNN backbone. We also test the performance of DenseUNet and PHNN using the CTV setup of CT appearance alone and our proposed method. **Second**, we extensively examine the impact of using different automatic GTV predictions as inputs for the CTV segmentation. This provides an evaluation of the complete RT target contouring pipeline and also measures the robustness of our CTV segmentation to inaccuracies in GTV predictions.

5. Results and discussion

5.1. Esophageal GTV segmentation

Effectiveness of PSNN: The quantitative results and comparisons are tabulated in Table 2 and Fig. 9. When all network models are trained and evaluated using only RTCT stream, our proposed PSNN evidently outperforms the previous best esophageal

Table 2
Mean DSC, HD, and ASD, and their standard deviations, of GTV segmentation performance using: (1) only RTCT images; (2) early fusion (EF) of RTCT and PET images; (3) the proposed two-stream chained fusion model, i.e., LF. The 3D DenseUNet model using RTCT is equivalent to the previous SOTA work (Yousefi et al., 2018), which is shown in the first row in blue color. The best performance scores are shown in **bold**.

3D Model	CT	EF	LF	DSC	HD (mm)	ASD (mm)
DenseUNet	✓			0.703 ± 0.178	72.2 ± 77.5	14.2 ± 28.0
		✓		0.731 ± 0.139	53.4 ± 68.0	10.2 ± 17.6
			✓	0.741 ± 0.137	61.4 ± 71.0	11.0 ± 17.7
PHNN	✓			0.743 ± 0.144	59.1 ± 69.3	10.1 ± 16.9
		✓		0.757 ± 0.132	46.9 ± 61.7	8.9 ± 16.6
			✓	0.766 ± 0.137	47.4 ± 59.3	8.6 ± 15.6
PSNN	✓			0.751 ± 0.147	43.7 ± 55.9	6.7 ± 9.6
		✓		0.778 ± 0.112	**35.4** ± **47.6**	5.7 ± 11.6
			✓	**0.790** ± **0.095**	39.3 ± 56.5	**5.7** ± **11.4**

Table 3
Mean DSC, HD, and ASD, and their standard deviations, of our CT/PET based GTV segmentation as compared to previous SOTA CT/PET co-segmentation approaches. PSNN - EF: PSNN model with early fusion of RTCT and PET images. PSNN - LF: PSNN model with our proposed two-stream chained fusion method. The best performance scores are shown in **bold**.

CT/PET co-segment	DSC	HD (mm)	ASD (mm)
Zhao et al. (2019)	0.676 ± 0.123	129.2 ± 77.4	37.1 ± 28.3
Zhong et al. (2019)	0.732 ± 0.128	87.8 ± 82.1	20.8 ± 29.4
Kumar et al. (2020)	0.742 ± 0.119	37.8 ± 41.7	7.7 ± 12.2
PSNN - EF (Ours)	0.778 ± 0.112	**35.4** ± **47.6**	5.7 ± 11.6
PSNN - LF (Ours)	**0.790** ± **0.09**5	39.3 ± 56.5	**5.7** ± **11.4**

GTV segmentation method, i.e., DenseUNet (Yousefi et al., 2018). As can be seen, PSNN consistently improves upon in all metrics: with an absolute increase of 4.8% in DSC (from 0.703 to 0.751) and significantly dropping in distance metrics of HD (from 72.2 mm to 43.7 mm) and ASD (from 14.2 mm to 6.7 mm), despite it being a simpler architecture. PSNN also outperforms the 3D version of PHNN (Harrison et al., 2017), e.g., with 3.4 mm ASD reduction, which indicates that the semantically-nested high- to low-level information flow provides key performance increases for locating esophageal tumors. For the setup of EF and full two-stream chained pipeline, i.e., LF, PSNN also consistently outperforms DenseUNet and PHNN.

Although UNet and its variations often achieve SOTA segmentation performance, they may not always be ideal and have limitations. Their symmetric encoder-decoder setup results in doubled network parameters and tripled memory-consumption as compared to simpler aggregation methods like PHNN and PSNN. This symmetric setup is computationally heavy and subject to limited batch-sizes, which inevitably requires more training time and is more difficult to optimize. In contrast, networks equipped with high abstraction capabilities and light-weighted decoding pathway, like PSNN, are able to achieve similar or better performance with less than 1/3 of the training time, which can be favored in many tasks.

Two-stream chained deep fusion: Table 2 also outlines the performance of three network models under different imaging configurations. Several conclusions can be drawn. First, all three networks trained using the EF stream (RTCT + PET) consistently produce more accurate segmentation results than those trained with only RTCT. This validates the effectiveness of utilizing PET to complement RTCT for GTV segmentation. For instance, the EF stream equipped with DenseUNet outperforms the CT stream by ~ 3% DSC improvement, and 18.8 mm HD and 4 mm ASD reduction. Second, the full two-stream chained pipeline, i.e., LF, provides further performance improvements to EF in terms of DSC while preserving similar distance errors. Importantly, the performance boosts can be observed across all three deep CNNs, validating that the two-stream combination of CT and EF can universally improve upon different segmentation backbones. Last, the best performing results are the PSNN model using LF, demonstrating that each component of the system contributes to the final performance. When compared to the previous SOTA work of esophageal GTV segmentation, which uses DenseUNet applied to RTCT images (Yousefi et al., 2018), our best performing model exceeds in all metrics of DSC, HD, and ASD by 8.7%, 32.9 mm and 8.5 mm remarkable margins (refer to the 1st and last row in Table 2). Fig. 8 shows several qualitative examples visually underscoring the improvements that our GTV segmentation approach (two-stream chained fusion + PSNN) provides.

Comparison with SOTA CT/PET co-segmentation: An alternative to our approach are the CT/PET co-segmentation strategies. Comparisons between our proposed two-stream chained CT/PET fusion and other co-segmentation methods are summarized in Table 3. Out of other co-segmentation approaches, Zhao et al. (2019) exhibits the worst results with 67.6% DSC and 37.1 mm ASD. A separate VNet for PET alone seems overfitting and simply fusing the two probability maps without considering any anatomical information is just ineffective. Zhong et al. (2019) designs the coupled UNet, each of which takes a single RTCT or PET as input. Additional skip connections from the encoder of one UNet are added to the other's decoder to encourage cross-modality feature sharing. This fusion approach performs much better compared to Zhao et al. (2019) (73.2% vs. 67.6% in DSC). Kumar et al. (2020) achieves the best performance, particularly in distance metrics, by using a carefully designed co-learning fusion approach to integrate features from CT- and PET-specific encoders into a shared decoding path. The importance of properly fusing features from CT and PET can be seen through these corresponding performance improvements.

Yet, equipped with the our PSNN network, our two-stream chained approach achieves the best performance among all CT/PET fusion methods, e.g., outperforming Kumar et al. (2020) by a large margin of almost 5% in DSC and 2 mm in ASD. In contrast to these SOTA methods, we utilize the PET modality together with CT as an early fusion rather than training separate network or encoder for PET alone. The intuition is while PET is helpful, considered in isolation this modality is noisy and does not contain any anatomical structural information (see Fig. 2(c)). This makes PET ineffective to learn from alone. To further support our intuition, note that even the PHNN models with a simple EF stream have achieved at least comparable or better performance as compared to these SOTA CT/PET fusion methods. This validates our CT/PET fusion approach.

Previously reported inter-observer variation: It worth notes that our esophageal GTV segmentation performance is indeed very much comparable to the inter-observer variations reported in previous studies. For instance, Nowee et al. (2019) studied the inter-observer performance of GTV delineation using 6 esophageal cancer patients by recruiting 19 radiation oncologists in 14 institutions. Using PET/CT images, oncologists reported a median of 0.69 Jaccard conformity index (equivalent to Jaccard index when 2 users are applied), and an average of 7 mm standard deviation in terms of surface distance error. Similar inter-observer GTV results were presented by Vesprini et al. (2008) using 10 esophageal cancer patients delineated from 6 radiation oncologists. In comparison, our best model achieves an average of 0.68 and a median of 0.70 Jaccard index and a 5.7 mm ASD, which are very similar to the inter-observer variations reported in Nowee et al. (2019); Vesprini et al. (2008). Nonetheless, a large cohort multi-center evaluation would be needed to fully validate the clinical applicability of our model.

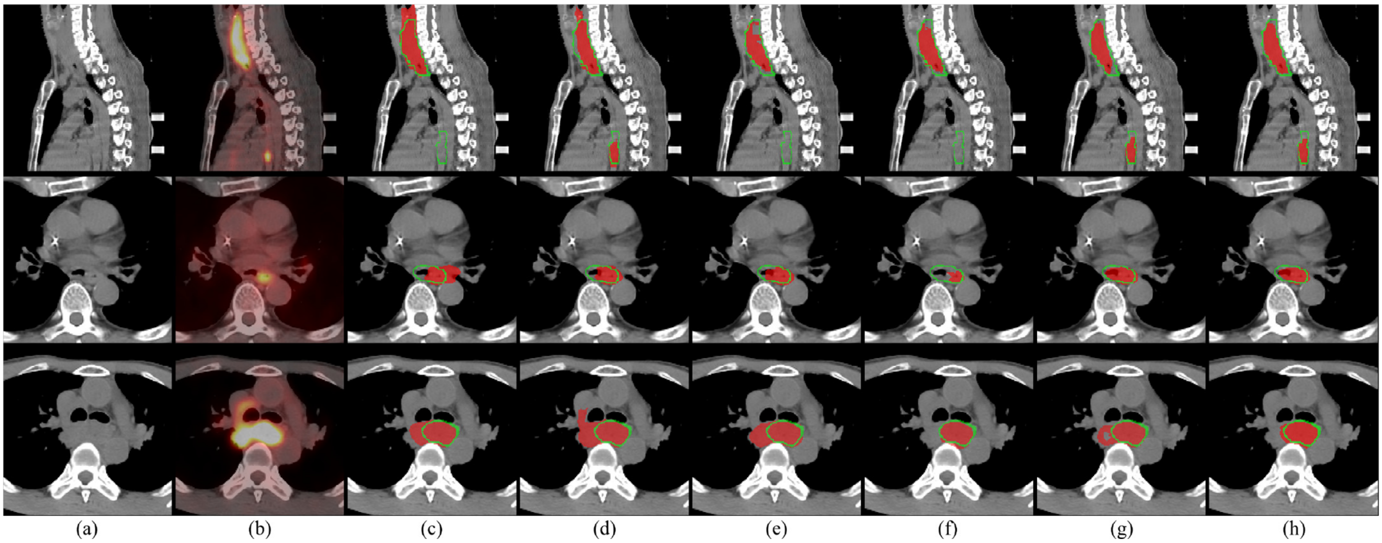

Fig. 8. Qualitative examples of esophageal GTV segmentation. (a) RTCT; (b) Registered PET overlaid on RTCT; (c) GTV segmentation results using CT stream with DenseUNet (Yousefi et al., 2018); (d) CT/PET co-segmentation method by Zhong et al. (2019); (e) CT/PET co-segmentation method by Kumar et al. (2020); (f) PSNN model using CT stream; (g) PSNN model using early fusion (EF) stream; (h) PSNN model using the full 2-stream chained fusion, i.e., LF (our final results). Red masks indicate automated segmentation results and green boundaries represent the ground truth. The first two rows demonstrate the importance of PET as using RTCT alone can cause over-, e.g., (c), or under-, e.g., (f), segmentation due to low contrast. Note that patient in first row has two separate tumors, one of which is relative small and missed completely by all CT-based methods ((c) and (f)) and even by a CT/PET co-segmentation method of Kumar et al. (2020), i.e., (e). The last row shows a case where over-segmentation can occur when the PET channel is spuriously noisy. In all cases, our final results (PSNN using LF) achieve good accuracy and robustness. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

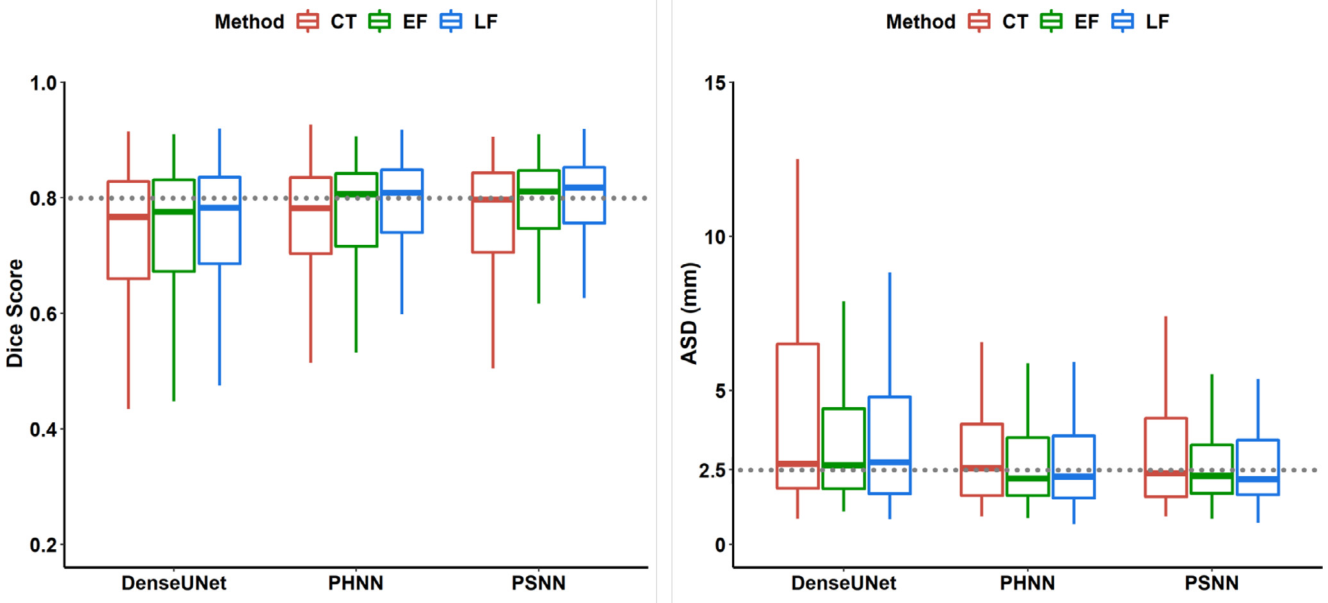

Fig. 9. Boxplots of the GTV segmentation performance using different fusion methods and deep networks on cross-validated 148 patients. From left to right depict the DSC score and ASD results, respectively.

5.2. Esophageal CTV segmentation

Effectiveness of spatial context encoding: Table 4 and Fig. 10 outline quantitative comparisons of the different CTV method setups and network choices. As can be seen, methods based on pure CT appearance (Men et al., 2017; 2018) exhibit the worst performance, regardless of the network choices. This is because inferring distance-based margins from appearance alone is too hard of a task for CNNs. Focusing on the PSNN performance, when adding the binary GTV and LN masks as contextual information (Cardenas et al., 2018a), the performance increases considerably from 74.7% to 79.2% in DSC scores. When using the SDM encoded spatial context of GTV/LN, the performance further improves by 1.3% in DSC and 1.8 mm in ASD. Finally, when the OAR SDM are also included, i.e., our proposed full workflow, it achieves the best performance reaching 82.6% in DSC, 39.1 mm in HD and 4.4 mm in ASD, with a remarked improvements of 3.4% in DSC, 29.4 mm in HD and 3.3 mm in HD as compared to the previous SOTA method (Cardenas et al., 2018a). These results confirm the importance of incorporating the distance-based context for esophageal

Table 4
Mean DSC, HD, and ASD, and their standard deviations, of different esophageal CTV segmentation methods and network models. Under each backbone, the best performance scores using various CTV segmentation methods are shown in **bold.**

3D models	CTV setups/methods	Dice	HD (mm)	ASD (mm)
DenseUNet	RTCT (Men et al., 2018)	0.734 ± 0.109	81.1 ± 45.7	10.8 ± 8.4
	RTCT + GTV/LN/OAR SDMs (Ours)	**0.820** ± **0.045**	**45.3** ± **18.9**	**4.8** ± **2.1**
PHNN	RTCT (Men et al., 2018)	0.735 ± 0.099	86.0 ± 48.1	10.7 ± 9.2
	RTCT + GTV/LN/OAR SDMs (Ours)	**0.828** ± **0.045**	**42.5** ± **21.8**	**4.2** ± **1.9**
PSNN	RTCT (Men et al., 2018)	0.747 ± 0.108	80.2 ± 45.8	10.0 ± 8.9
	RTCT + GTV/LN masks (Cardenas et al., 2018a)	0.792 ± 0.075	68.5 ± 47.5	7.7 ± 7.7
	RTCT + GTV/LN SDMs (Ours)	0.805 ± 0.056	53.1 ± 30.1	5.9 ± 3.8
	RTCT + GTV/LN/OAR SDMs (Ours)	**0.826** ± **0.050**	**39.1** ± **21.9**	**4.4** ± **2.1**

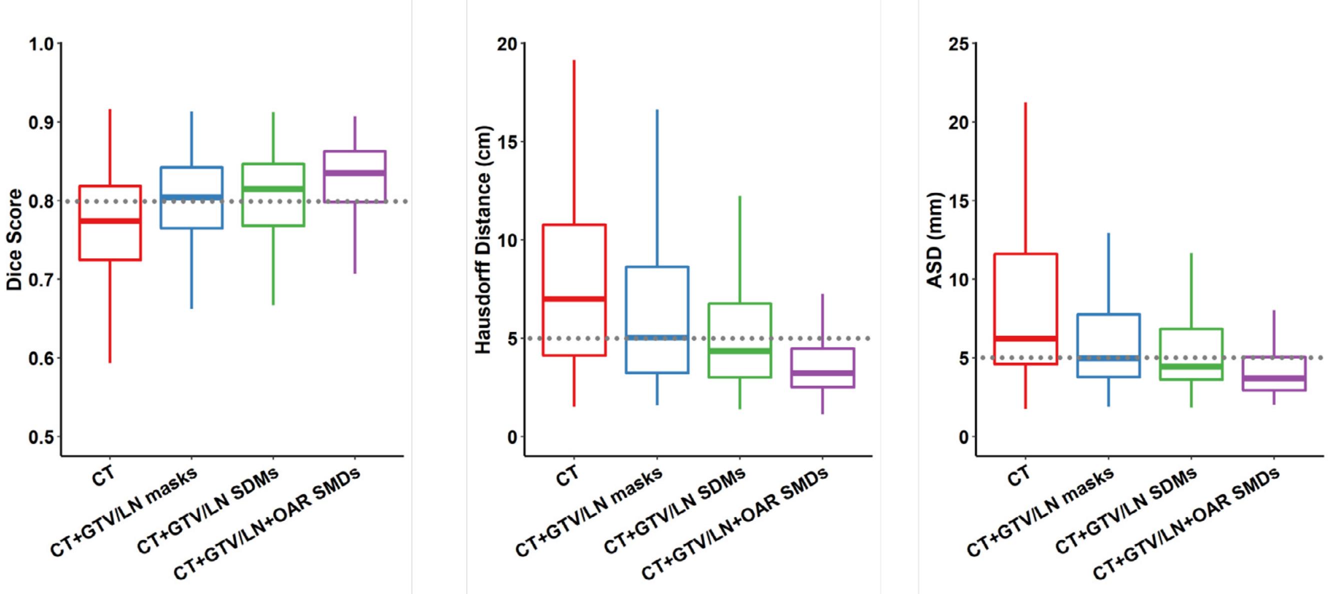

Fig. 10. Boxplots of the CTV segmentation performance under 4 setups equipped with the PSNN backbone on cross-validated 148 patients. From left to right depict the DSC score, HD and ASD results, respectively. CT means the pure appearance based CTV segmentation method (Men et al., 2018), while CT + GTV/LN masks represents the SOTA CTV segmentation method (Cardenas et al., 2018a).

CTV delineation, as well as the importance of OARs. In particular, as can be seen by looking at quartile results in Fig. 10, our proposed approach, "RTCT + GTV/LN/OAR SDMs", exhibits much greater robustness and reliability. Fig. 11 presents some qualitative examples further showing these performance improvements.

Effect of network choices: We also compare the performance of different networks when using the CT appearance setup and the proposed full CTV segmentation pipeline. As Table 4 demonstrates, when using the full pipeline, PSNN, PHNN and DenseUNet achieve very similar DSC scores ($<1\%$ difference), although PSNN and PHNN exhibit marginally reduced distance errors. When using only the CT appearance information, PSNN slightly improves the DSC and ASD by $\geq 1.2\%$ and $\geq 0.7\,\text{mm}$, respectively, as compared against both PHNN and DenseUNet. These results confirm our observation that for esophageal CTV delineation, the appropriate problem formulation is most important: under an appropriate setup, networks play a less crucial role to the performance. On the other hand, when the problem is not very well-defined, i.e., using RTCT appearance alone, a network with strong abstraction capacity might produce slightly better results.

Previously reported inter-observer variation: Wong et al. (2020) studied inter-observer variation on the head and neck CTV delineation, and the DSC between 2 oncologists in 5 patients ranged from 0.7 to 0.87 with an average value of 0.80. Eminowicz and McCormack (2015) reported the inter-observer performance on cervical cancer CTV delineation by recruiting more than 20 physicians. The Jaccard conformity index ranged from 0.51 to 0.81 with a mean value of 0.655. In comparison, our best esophageal CTV segmentation model achieves an average DSC of 0.83 (0.70 in terms of Jaccard index). More specifically, $\geq 75\%$ patients have DSC score ≥ 0.80, and $\geq 40\%$ patients have DSC score ≥ 0.85. These comparisons demonstrate that our automated method is at least comparable to those inter-observer variations reported in previous CTV delineation studies, which indicates our potential clinical applicability.

5.3. CTV performance using automated GTV predictions

Recall that we train our CTV segmentation model using the ground truth GTV mask for stability concerns. However, we do not make this assumption for inference. Once the CTV models are well-trained using GTV ground truth, we test the CTV performance using different GTV predictions. These results are summarized in Table 5. Several observations can be made. (1) For CTV delineation, the setup/method plays a more important role than the absolute accuracy of the GTV contouring. As can be seen, using our proposed "RTCT + GTV/LN/OAR SDM" CTV setup, even the worst results by using the GTV predictions of Zhong et al. (2019) still perform better than those under the "RTCT + GTV/LN masks (Cardenas et al., 2018a)" CTV setup, regardless of the origin of the GTV predictions (even using GTV ground truth). (2) Given a CTV setup, more accurate GTV pre-

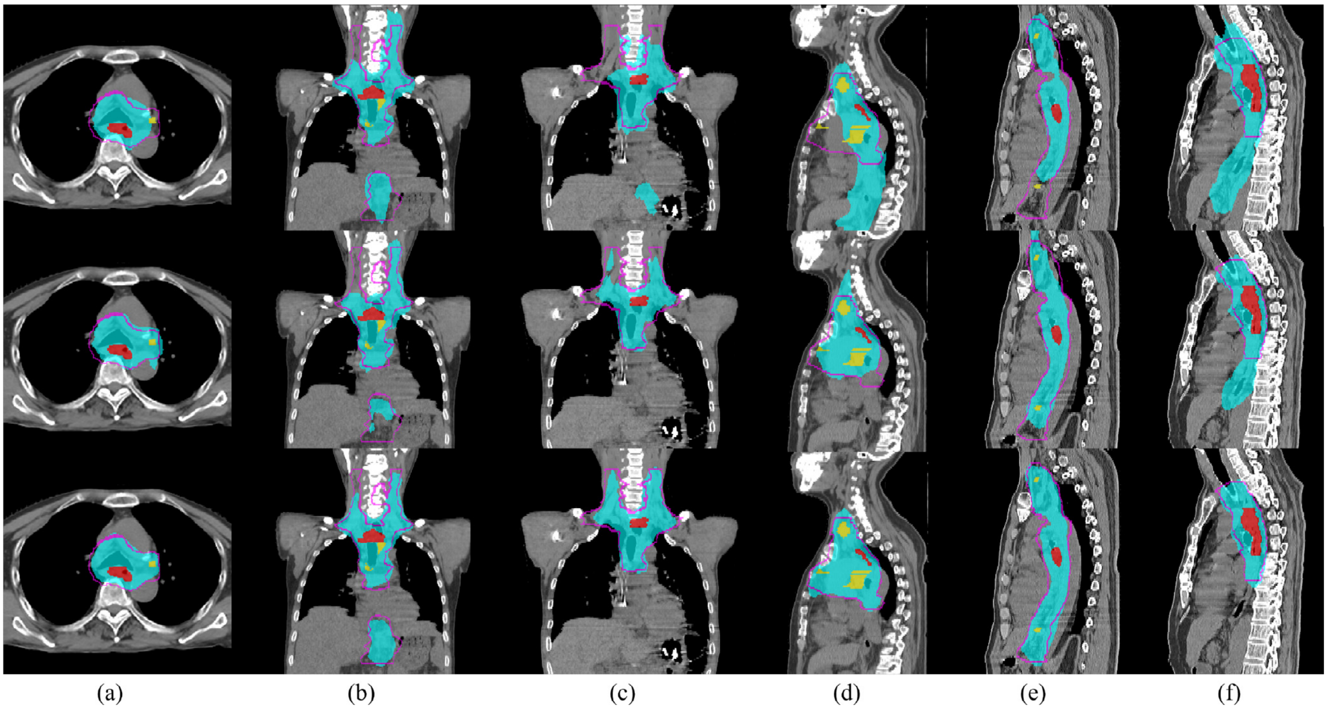

Fig. 11. Qualitative illustrations of esophageal CTV delineation using different setups. Red, yellow and cyan represent the GTV, LN and predicted CTV regions, respectively. The purple line indicates the ground truth CTV boundary. The 1st and 2nd rows show examples from setups using pure RTCT (Men et al., 2018) and when adding GTV/LN binary masks (Cardenas et al., 2018a), respectively. The 3rd row shows examples using our proposed GTV/LN/OAR SDMs setup. All results use the proposed PSNN backbone. (a), (d) and (e) demonstrate that the pure RTCT setups fail to include the involved LNs, while (c), (d) and (f) depict severe over-segmentations. While these errors are partially addressed using the GTV/LN mask setup, it still suffers from inaccurate CTV boundaries (c-e) or under/over-coverage of normal regions (b,f). These issues are much better addressed by our proposed method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Quantitative results of esophageal CTV segmentation using predicted GTV masks of different GTV segmentation methods under various CTV setups/methods. All CTV setups use the PSNN backbone. PSNN - LF is our proposed GTV segmentation approach using two-stream chained deep fusion. Under each CTV setup, the best performance scores using different automated GTV predictions are shown in **bold**. Meanwhile, CTV performance using the ground truth GTV label is presented in *italics* under each CTV setup as reference of best achieved results.

CTV setups/methods	GTV origins	DSC	HD (mm)	ASD (mm)
RTCT + GTV/LN masks (Cardenas et al., 2018a)	Yousefi et al., 2018	0.762 ± 0.112	73.7 ± 48.4	9.2 ± 9.0
	Zhong et al. (2019)	0.759 ± 0.116	76.2 ± 49.1	9.8 ± 9.8
	Kumar et al. (2020)	0.772 ± 0.105	71.5 ± 48.8	8.9 ± 9.3
	PSNN - LF (Ours)	**0.778** ± **0.090**	**70.4** ± **45.6**	**8.2** ± **6.9**
	GTV ground truth	*0.792* ± *0.075*	*68.5* ± *47.5*	*7.7* ± *7.7*
RTCT + GTV/LN SDMs (Ours)	Yousefi et al., 2018	0.778 ± 0.098	60.5 ± 37.8	7.5 ± 6.4
	Zhong et al. (2019)	0.778 ± 0.086	63.4 ± 40.2	7.7 ± 6.5
	Kumar et al. (2020)	0.789 ± 0.078	**57.4** ± **35.0**	**6.8** ± **5.0**
	PSNN - LF (Ours)	**0.790** ± **0.077**	57.7 ± 35.3	6.9 ± 5.4
	GTV ground truth	*0.805* ± *0.056*	*51.1* ± *30.1*	*5.9* ± *3.8*
RTCT + GTV/LN/OAR SDMs (Ours)	Yousefi et al., 2018	0.800 ± 0.082	47.0 ± 30.7	5.6 ± 4.3
	Zhong et al. (2019)	0.800 ± 0.077	51.1 ± 35.1	6.1 ± 5.0
	Kumar et al. (2020)	0.811 ± 0.069	43.1 ± 28.0	5.1 ± 3.8
	PSNN - LF (Ours)	**0.813** ± **0.068**	**42.6** ± **27.6**	**5.0** ± **3.7**
	GTV ground truth	*0.826* ± *0.050*	*39.1* ± *21.9*	*4.4* ± *2.1*

dictions generally produce better CTV results. However, different GTV results can exhibit similar CTV performance. For instance, CTV performance using PSNN-LF GTV predictions is similar to that using GTV predictions of Kumar et al. (2020), although Kumar et al. (2020) has a markedly lower performance in terms of DSC and ASD (74.2% vs. 79.0% in DSC and 7.7 mm vs. 5.7 mm in ASD as shown in Table 3). Similar phenomenon is observed when considering the GTV predictions of Zhong et al. (2019) and Yousefi et al., 2018. A close look at the GTV results show that Kumar et al. (2020) and PSNN-LF have similar HD errors (similar HD errors also exist between Zhong et al. (2019) and Yousefi et al., 2018). This finding indicates that our CTV segmentation method could tolerate the regional contour inaccuracy, and is only affected by distant false-positive or false-negative errors (characterized by the HD metric), which confirms the robustness of our CTV segmentation method.

Limitation of the CTV segmentation: One factor that limits the automated workflow of the proposed CTV segmentation is the utilization of manually identified LNs. Detecting the enlarged LNs is an difficult task and has long been studied in the literature (Barbu et al., 2011; Roth et al., 2016; Liu et al., 2016). The identification of the small and scatteredly-distributed metastasis LNs, refer to GTV$_{LN}$ in RT, is more challenging, especially in non-contrast RTCT. Considering that even the most recent GTV$_{LN}$ detection performance is prone to large false-positive rates (Chao et al., 2020; Zhu et al., 2020a; 2020b), we leave the metastasis LN identification as a separate and further research topic.

6. Conclusions

This work presented a complete workflow for esophageal GTV and CTV segmentation. First, we proposed a two-stream chained 3D deep network fusion method to segment esophageal GTVs using PET and RTCT imaging modalities. This two-stream fusion outperforms prior art, including leading co-segmentation alternatives. We also introduced the PSNN model as a new 3D segmentation architecture that uses a simple, parameter-less, and deeply-supervised CNN decoding path, and demonstrated its superior in the tumor segmentation task as compared against other SOTA networks, such as PHNN and DenseUNet. Second, we introduced a spatial context encoded deep esophageal CTV segmentation framework designed produce superior margin-based CTV boundaries. Our method encodes spatial context by computing the SDMs of the GTV, LNs and OARs and feeds them together with the RTCT image into a 3D deep CNN. Analogous to clinical practice, this allows the network to consider both appearance and distance-based information for segmentation. Experiments demonstrate that our spatial context-aware CTV segmentation approach significantly outperforms prior approaches to this task. Put together, our work represents a complete workflow for the target delineation in esophageal cancer radiotherapy and pushes forward the state of automated esophageal GTV and CTV segmentation towards a clinically applicable solution.

Declaration of Competing Interest

We have no conflicts of interest to disclose.

CRediT authorship contribution statement

Dakai Jin: Conceptualization, Data curation, Formal analysis, Methodology, Software, Writing - original draft. **Dazhou Guo:** Conceptualization, Formal analysis, Methodology, Software, Writing - original draft. **Tsung-Ying Ho:** Conceptualization, Data curation, Formal analysis, Writing - review & editing, Supervision. **Adam P. Harrison:** Methodology, Formal analysis, Writing - original draft. **Jing Xiao:** Formal analysis, Writing - review & editing. **Chen-kan Tseng:** Conceptualization, Writing - review & editing. **Le Lu:** Conceptualization, Formal analysis, Methodology, Writing - original draft, Supervision.

Acknowledgements

This work was partially supported by the Maintenance Project of the Center for Artificial Intelligence in Medicine (Grant CLRPG3H0012, SMRPG3I0011) at Chang Gung Memorial Hospital.

References

Barbu, A., Suehling, M., Xu, X., et al., 2011. Automatic detection and segmentation of lymph nodes from ct data. IEEE Trans. Med. Imaging 31 (2), 240–250.
Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23 (11), 1222–1239.
Bray, F., Ferlay, J., et al., 2018. Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal Clinicians 68 (6), 394–424.
Burnet, N.G., Thomas, S.J., Burton, K.E., Jefferies, S.J., 2004. Defining the tumour and target volumes for radiotherapy. Cancer Imaging 4 (2), 153.
Cardenas, C.E., Anderson, B.M., et al., 2018. Auto-delineation of oropharyngeal clinical target volumes using 3D convolutional neural networks. Phys. Med. Biol. 63 (21), 215026.
Cardenas, C.E., McCarroll, R.E., Court, L.E., et al., 2018. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. Int. J. Radiat. Oncol.* Biol.* Phys. 101 (2), 468–478.
Chao, C.-H., Zhu, Z., Guo, D., et al., 2020. Lymph node gross tumor volume detection in oncology imaging via relationship learning using graph neural network. In: MICCAI. Springer, pp. 772–782.
Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., et al., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: MICCAI, pp. 424–432.
Eminowicz, G., McCormack, M., 2015. Variability of clinical target volume delineation for definitive radiotherapy in cervix cancer. Radiother. Oncol. 117 (3), 542–547.
George, K., Harrison, A.P., Jin, D., et al., 2017. Pathological pulmonary lobe segmentation from ct images using progressive holistically nested neural networks and random walker. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 195–203.
Guo, D., Jin, D., Zhu, Z., et al., 2020. Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4223–4232.
Guo, Z., Guo, N., Gong, K., Li, Q., et al., 2019. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. Phys. Med. Biol. 64 (20), 205015.
Hao, Z., Liu, J., Liu, J., 2017. Esophagus tumor segmentation using fully convolutional neural network and graph cut. In: Intelligent Systems. Springer, pp. 413–420.
Harrison, A.P., Xu, Z., George, K., et al., 2017. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images. In: MICCAI. Springer, pp. 621–629.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
Huang, B., Chen, Z., Wu, P.-M., et al., 2018. Fully automated delineation of gross tumor volume for head and neck cancer on PET-CT using deep learning: a dual-center study. Contrast Media Mol. Imaging 2018, 1–12.
Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: IEEE CVPR, pp. 2261–2269.
Jin, D., Guo, D., Ho, T.-Y., et al., 2019a. Accurate esophageal gross tumor volume segmentation in PET/CT using two-stream chained 3D deep network fusion. In: MICCAI. Springer, pp. 182–191.
Jin, D., Guo, D., Ho, T.-Y., et al., 2019b. Deep esophageal clinical target volume delineation using encoded 3D spatial context of tumor, lymph nodes, and organs at risk. In: MICCAI. Springer, pp. 603–612.
Jin, D., Xu, Z., Harrison, A.P., et al., 2017. 3D convolutional neural networks with graph refinement for airway segmentation using incomplete data labels. In: Machine Learning in Medical Imaging. Springer, pp. 141–149.
Zhu, Z., Jin, D., Yan, K., 2020. Lymph node gross tumor volume detection and segmentation via distance-based gating using 3D CT/PET imaging in radiotherapy. In: MICCAI. Springer, pp. 753–762.
Kingma, D. P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980.
Jin, D., Xu, Z., Tang, Y., et al. J, 2018. CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. In: MICCAI. Springer, pp. 732–740.
Klein, S., Staring, M., Murphy, K., et al., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29 (1), 196–205.
Kumar, A., Fulham, M., Feng, D., Kim, J., 2020. Co-learning feature fusion maps from PET-CT images of lung cancer. IEEE Trans. Med. Imaging 39 (1), 204–217.
Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z., 2015. Deeply-supervised nets. In: Artificial Intelligence and Statistics, pp. 562–570.
Leong, T., Everitt, C., et al., 2006. A prospective study to evaluate the impact of FDG-PET on CT-based radiotherapy treatment planning for oesophageal cancer. Radiother. Oncol. 78 (3), 254–261.
Lian, C., Ruan, S., Denœux, T., Li, H., Vera, P., 2019. Joint tumor segmentation in PET-CT images using co-clustering and fusion based on belief functions. IEEE Trans. Image Process. 28 (2), 755–766.
Lin, L., Dou, Q., Jin, Y.-M., et al., 2019. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. Radiology 291 (3), 677–686.
Liu, J., Hoffman, J., Zhao, J., et al., 2016. Mediastinal lymph node detection and station mapping on chest ct using spatial priors and random forest. Med. Phys. 43 (7), 4362–4374.
Ma, Z., Zhou, S., Wu, X., et al., 2019. Nasopharyngeal carcinoma segmentation based on enhanced convolutional neural networks using multi-modal metric learning. Phys. Med. Biol. 64 (2), 025005.
Mattes, D., Haynor, D.R., Vesselle, H., Lewellen, T.K., Eubank, W., 2003. Pet-ct image registration in the chest using free-form deformations. IEEE Trans. Med. Imaging 22 (1), 120–128.
Maurer Jr., C.R., Qi, R., Raghavan, V., 2003. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions. IEEE Trans. Pattern Anal. Mach. Intell. 25 (2), 265–270.
Men, K., Dai, J., Li, Y., 2017. Automatic segmentation of the clinical target volume and organs at risk in the planning ct for rectal cancer using deep dilated convolutional neural networks. Med. Phys. 44 (12), 6377–6389.

Men, K., Zhang, T., et al., 2018. Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. Phys. Med. 50, 13–19.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571.

Muijs, C.T., Beukema, J.C., et al., 2010. A systematic review on the role of FDG-PET/CT in tumour delineation and radiotherapy planning in patients with esophageal cancer. Radiother. Oncol. 97 (2), 165–171.

Murphy, K., Van Ginneken, B., Reinhardt, J.M., et al., 2011. Evaluation of registration methods on thoracic CT: the empire10 challenge. IEEE Trans. Med. Imaging 30 (11), 1901–1920.

Nowee, M.E., Voncken, F.E., Kotte, A., et al., 2019. Gross tumour delineation on computed tomography and positron emission tomography-computed tomography in oesophageal cancer: anationwide study. Clin. Transl. Radiat. Oncol. 14, 33–39.

Pennathur, A., Gibson, M.K., Jobe, B.A., Luketich, J.D., 2013. Oesophageal carcinoma. Lancet 381 (9864), 400–412.

Pennec, X., Cachier, P., Ayache, N., 1999. Understanding the 'demon's algorithm': 3Dnon-rigid registration by gradient descent. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 597–605.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: MICCAI. Springer, pp. 234–241.

Roth, H.R., Lu, L., Liu, J., et al., 2016. Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE Trans. Med. Imaging 35 (5), 1170–1181.

Rueckert, D., Sonoda, L.I., Hayes, C., et al. J, 1999. Nonrigid registration using free–form deformations: application to breast MR images. IEEE TMI 18 (8), 712–721.

Tai, P., Van Dyk, J., Yu, E., et al., 1998. Variability of target volume delineation in cervical esophageal cancer. Int. J. Radiat. Oncol.* Biol.* Phys. 42 (2), 277–288.

Tan, S., Li, L., Choi, W., et al., 2017. Adaptive region-growing with maximum curvature strategy for tumor segmentation in 18F-FDG PET. Phys. Med. Biol. 62 (13), 5383.

Vesprini, D., Ung, Y., Dinniwell, R., et al., 2008. Improving observer variability in target delineation for gastro-oesophageal cancerthe role of 18ffluoro-2-deoxy-d-glucose positron emission tomography/computed tomography. Clin. Oncol. 20 (8), 631–638.

Wong, J., Fong, A., McVicar, N., et al., 2020. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiother. Oncol. 144, 152–158.

Xu, L., Tetteh, G., et al., 2018. Automated whole-body bone lesion detection for multiple myeloma on 68ga-pentixafor pet/ct imaging using deep learning methods. Contrast Media Mol. Imaging 2018.

Yang, J., Beadle, B.M., Garden, A.S., et al., 2014. Auto-segmentation of low-risk clinical target volume for head and neck radiation therapy. Pract. Radiat. Oncol. 4 (1), e31–e37.

Yousefi, S., Sokooti, H., Elmahdy, M., et al., 2018. Esophageal gross tumor volume segmentation using a 3D convolutional neural network. In: MICCAI. Springer, pp. 343–351.

Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations.

Zhao, X., Li, L., et al., 2019. Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. Phys. Med. Biol. 64 (1), 015011.

Zhong, Z., Kim, Y., et al., 2019. Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks. Med. Phys. 46 (2), 619–633.

Yan, K., Cai, J., Harrison, A. P., et al., 2020. Universal lesion detection by learning from multiple heterogeneously labeled datasets. arXiv:2005.13753.

Zhu, Z., Yan, K., Jin, D. et al., 2020b. Detecting scatteredly-distributed, small, andcritically important objects in 3D oncologyimaging via decision stratification. arXiv:2005.13705.