Accurate Esophageal Gross Tumor Volume Segmentation in PET/CT using Two-Stream Chained 3D Deep Network Fusion

MICCAI 2019
13-17 OCTOBER 2019
SHENZHEN, CHINA

Dakai Jin[1], Dazhou Guo[1], Tsung-Ying Ho[2], Adam P. Harrison[1], Jing Xiao[3], Chen-Kan Tseng[2], Le Lu[1]

1. PAII Inc., Bethesda, MD, USA 2. Chang Gung Memorial Hospital, Linkou, Taiwan, ROC 3. Ping An Technology, Shenzhen, China

Motivation and Objective

Motivation

- **Esophageal gross tumor volume (GTV) segmentation**
  - One of the most critical tasks in radiotherapy treatment planning
  - Time consuming and inconsistency in manual contouring
- **Segmentation challenges in RTCT**
  - Non-contrast imaging
  - Poor contrast for esophagus
  - Poor contrast for esophageal tumors
  - Large range from superior to inferior
  - Large shape/appearance variations
- **Prior art on GTV segmentation**
  - 3D DenseUnet in MICCAI2018[1]
  - Trained and tested on 49 distinct patients
  - *Performance: low Dice score (<70%) and large Hausdorff distance errors (>100 mm)*

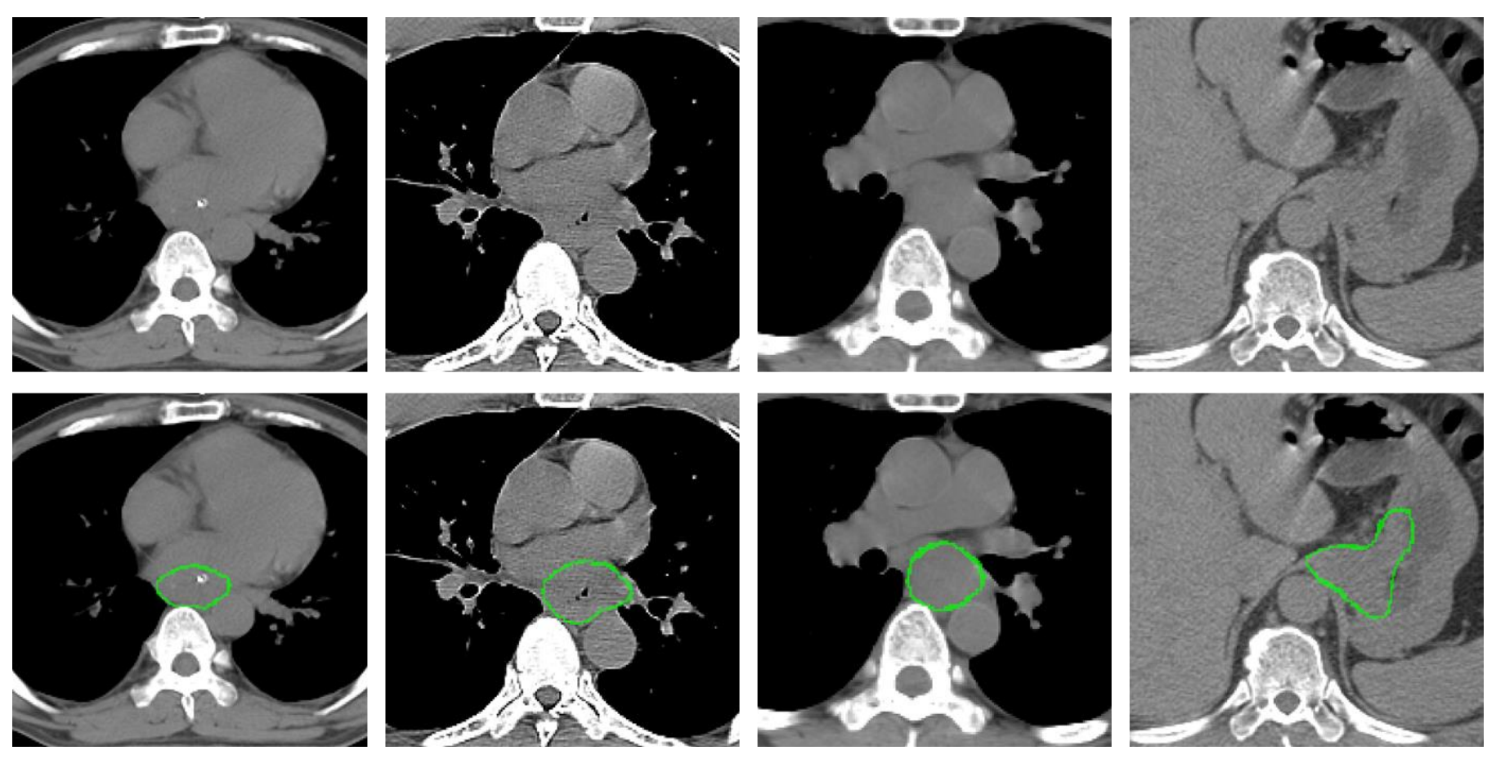


Fig. 1 Esophageal GTV examples

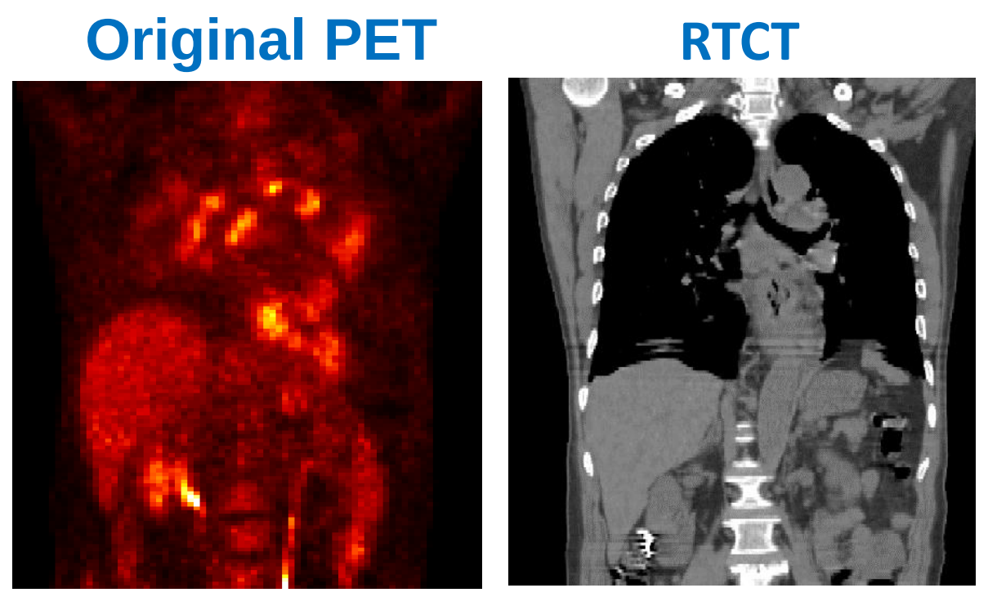
Aim

- ***Develop an accurate and robust 3D esophageal GTV segmentation method:***
  - Design a 2-stream chained pipeline incorporate the joint RTCT and PET information
  - Introduce a simple yet surprisingly powerful progressive semantically nested network (PSNN) model, which incorporates strengths of both UNet[2] and P-HNN[3]
  - 5-fold cross-evaluate the proposed method on 110 patients

Methods

PET and RTCT

- ***Adding PET modality is helpful yet not trivial:***
  - ✓ high sensitivity for tumors
  - low specificity for tumors
  - very coarse spatial resolution (3~4mm)
  - PET and RTCT is not aligned



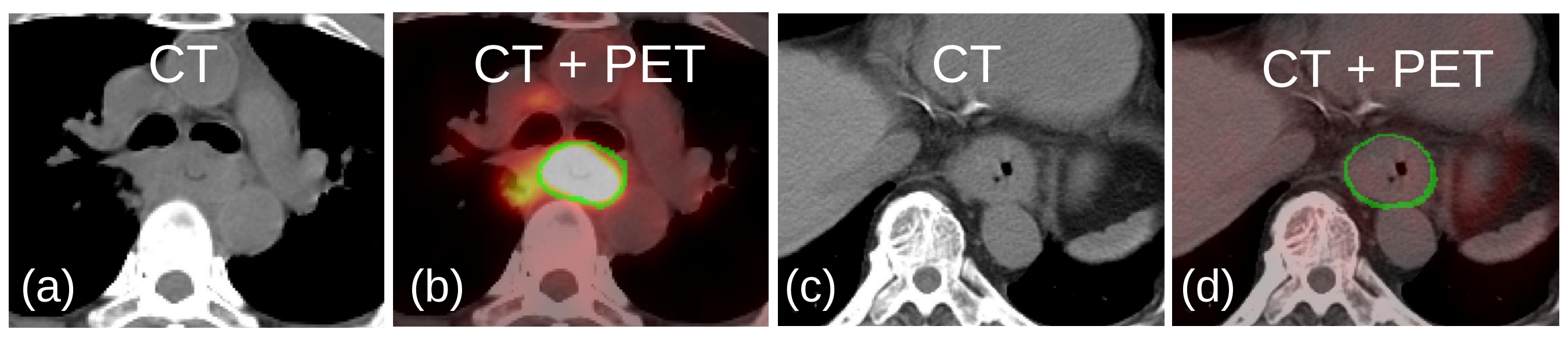


Fig. 2 The GTV boundaries are hardly distinguishable in CT (a), but it can be reasonably inferred from PET (b). No high uptake regions appear in PET (c), but the esophagus wall enlargement appears in CT (d).

Overall Framework

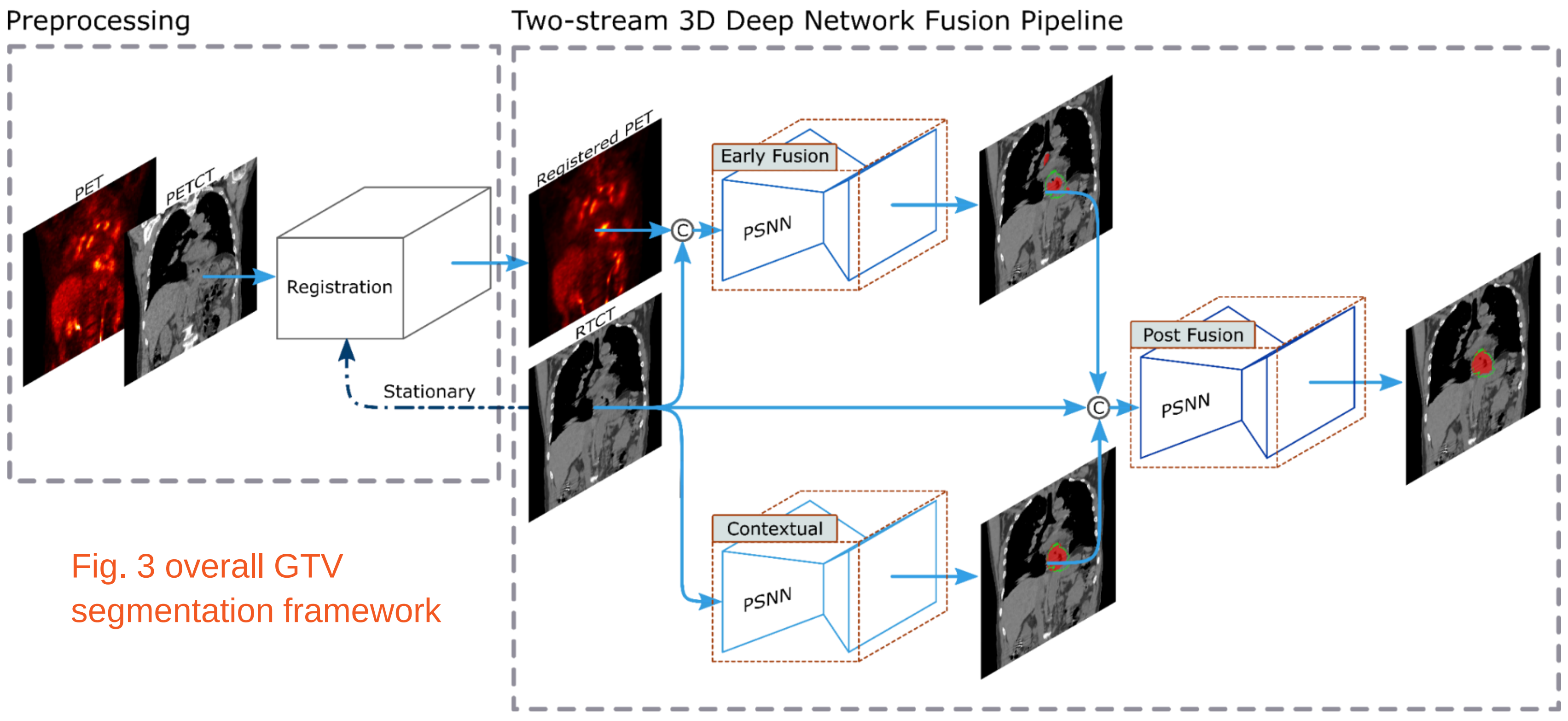


Fig. 3 overall GTV segmentation framework

- A 2-stream chained approach effectively fuses RTCT and PET modalities via early and late 3D deep-network-based fusion
  - One stream trained using only RTCT (contextual): $\hat{y}_j^{\mathrm{CT}} = p_j^{\mathrm{CT}}\left(y_j = 1 \middle| X^{\mathrm{CT}}; \mathbf{W}^{\mathrm{CT}}\right)$
  - One stream trained using both RTCT and aligned PET (early fusion):

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

  - Late fusion of the two streams from CT and early fusion models:

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

PET to RTCT Registration

- register PET to RTCT is difficult (different modality)
- Challenges to register diagnostic CT to RTCT:
  - large differences in body ranges
  - different poses for head & arms
  - soft and hard scanner boards

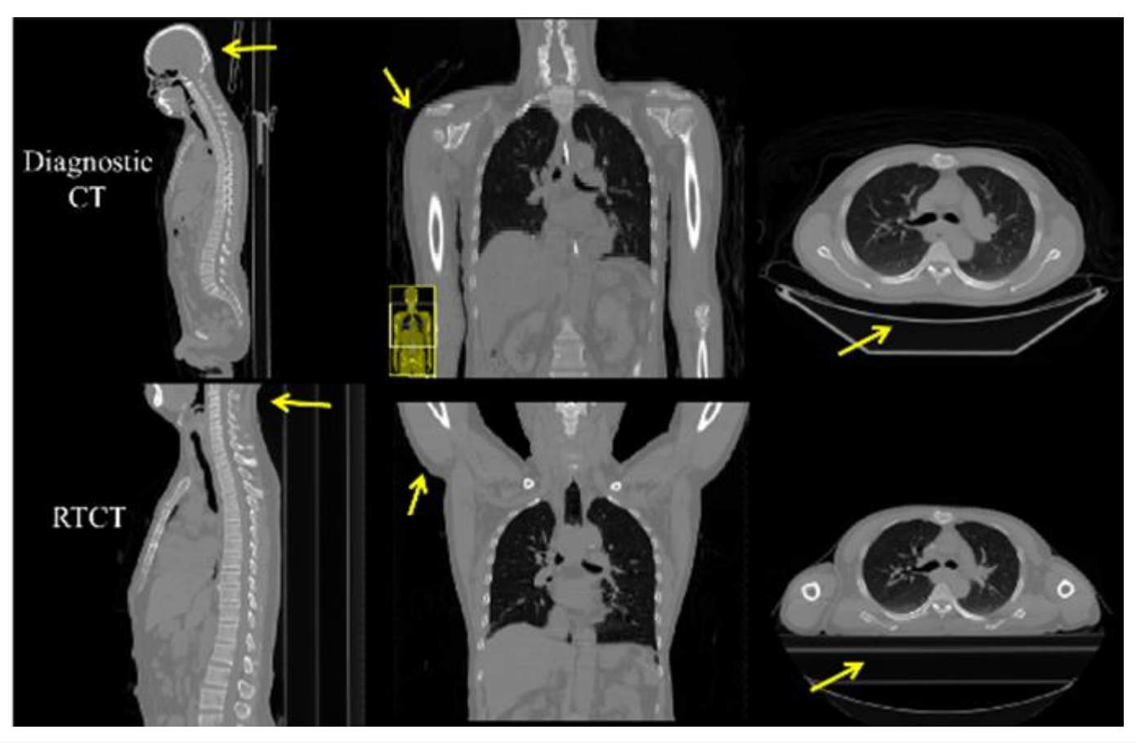
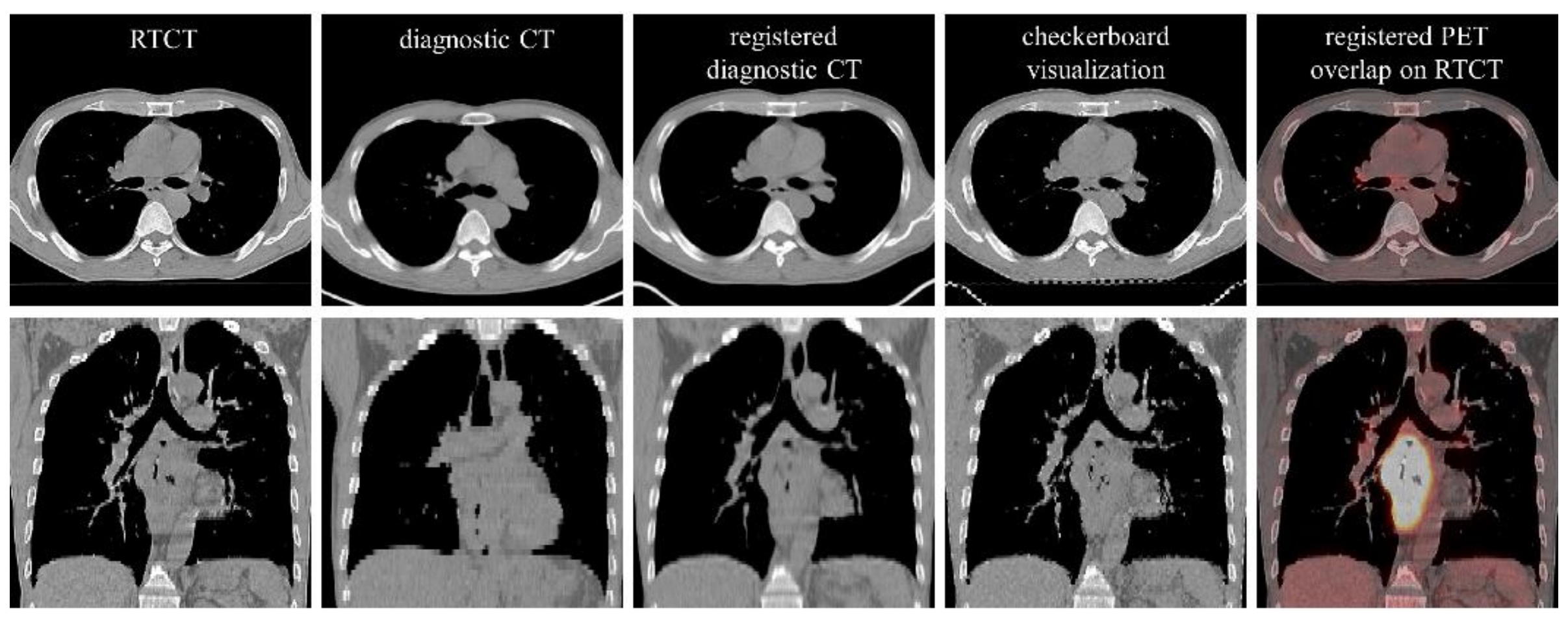
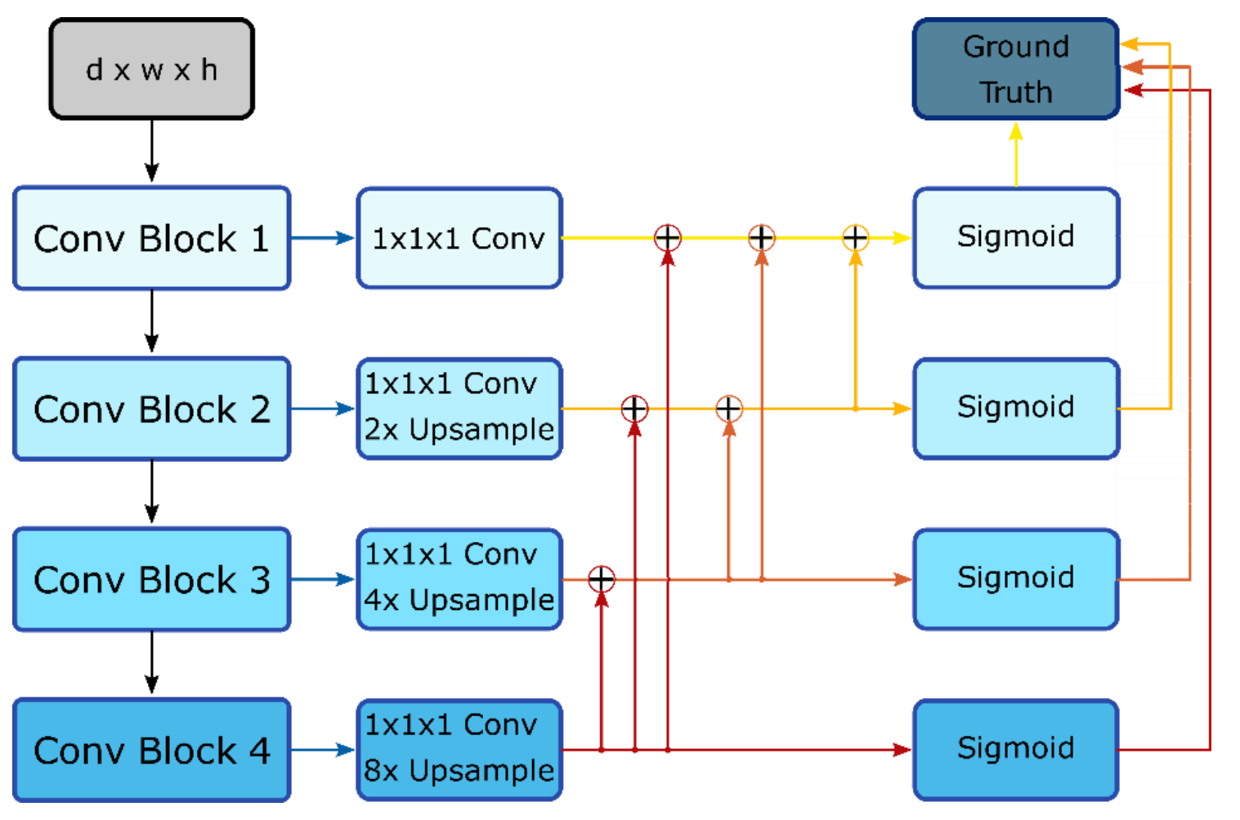


Fig. 4 Difference in diagnostic & RTCT

- Anatomy-based initialization: centers of 3D lung segmentation
- A multi-scale coarse to fine B-spline based deformable registration



PSNN model



- Progressive semantically nested network (PSNN)
  - Deeper layers: strong semantics but low spatial resolution
  - Shallower features: vice versa
  - Progressively propagate high-level semantics to guide low-level learning
  - Suitable for applications where objects have reasonable size, but exhibit poor contrast

Experiments and Results

Datasets & Evaluation Metrics

- 110 esophageal cancer patients diagnosed at stage II or later undergoing RT
- Each patient with a diagnostic PET/CT pair and a treatment RTCT scan
- Evaluation metrics: Dice score, Hausdorff distance (HD) in mm, and average surface distance with respect to the ground truth contour ($\mathrm{ASD}_{\mathrm{GT}}$) in mm

Training Data Generation and Training Parameters

- 80x80x64 training VOI near ground truth GTV or randomly sampled
- average ~80 training VOI per patient
- Adam solver with momentum 0.99 and a weight decay of 0.005, train for ~40 epochs

Qualitative compare:

- ❑ (a): RTCT overlayed with registered PET
- ❑ (b): DenseUNet trained by RTCT only
- ❑ (c): PSNN trained by RTCT only
- ❑ (d): PSNN trained by early fusion, e.g. RTCT+PET
- ❑ (e) PSNN trained by early + late fusion of RTCT and PET
- ✓ First two rows show importance of PET
- ✓ Last two rows shows importance of late fusion

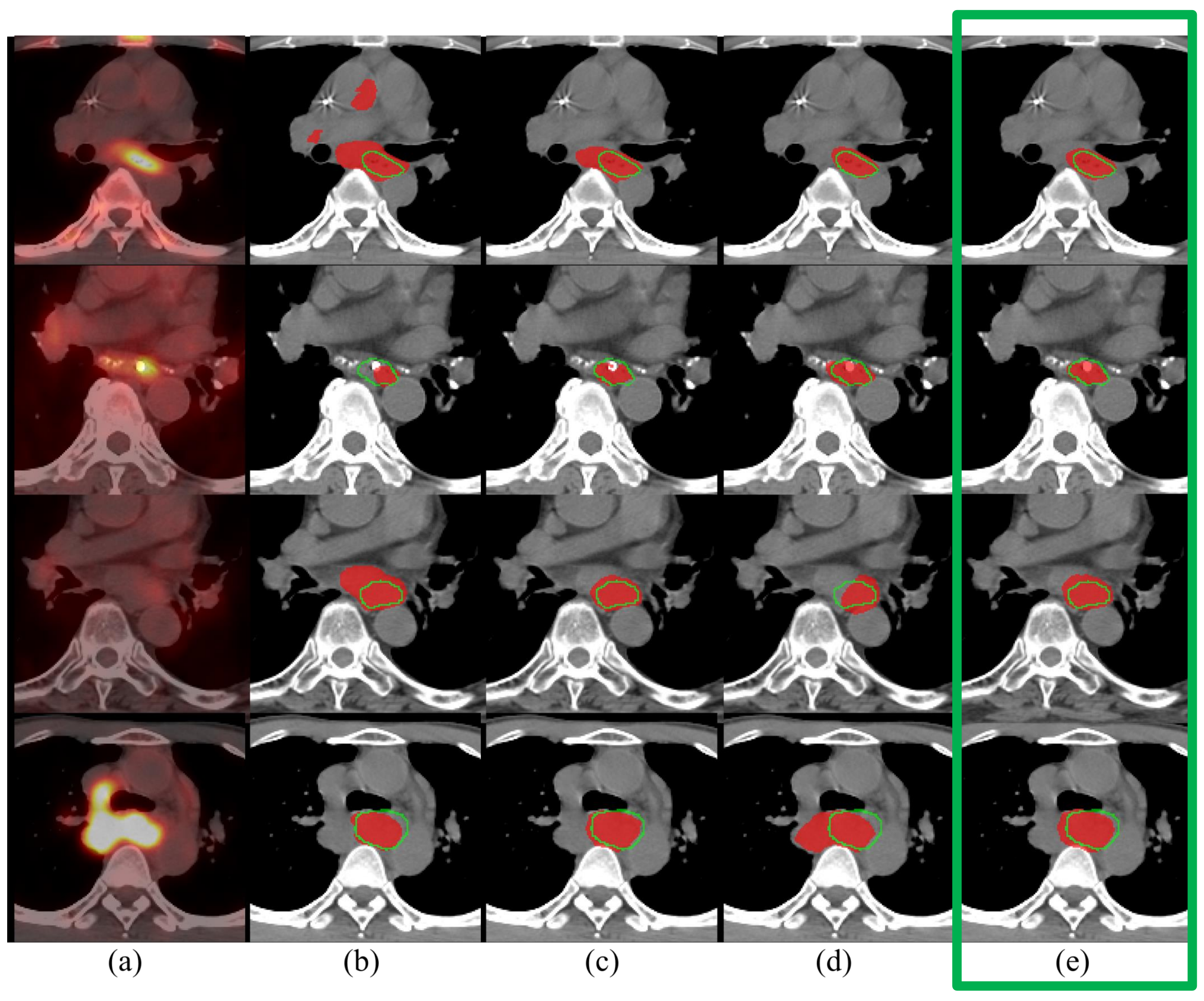


Table. 1 GTV segmentation performance using: (1) Contextual model (only RTCT); (2) Early fusion model (EF) using both RTCT and PET; (3) Proposed two-stream chained early and late fusion model (EF+LF)

	CT	EF	EF+LF	DSC	HD (mm)	ASD_{GT} (mm)
3D DenseUNet	✓			0.654±0.210	129.0±73.0	5.2±12.8
		✓		0.710±0.189	116.0±81.7	4.9±10.3
			✓	0.745±0.163	79.5±70.9	4.7±10.5
3D PSNN	✓			0.728±0.158	66.9±59.2	4.2±5.4
		✓		0.758±0.136	67.0±59.1	**3.2±3.1**
			✓	**0.764±0.134**	**47.1±56.0**	**3.2±3.3**

Conclusion

- Presented a two-stream chained 3D deep network fusion pipeline to segment esophageal GTVs using both RTCT and PET+RTCT imaging channels. And validate that it provides an effective means to exploit the complementary information seen within PET and CT
- Introduce a new 3D segmentation architecture, named PSNN, which uses a simple, parameter-less, and deeply-supervised CNN decoding stream.
- Demonstrate that our PSNN model outperform the state-of-the-art P-HNN and DenseUNet networks with remarked margins.

References:

[1] Yousefi, S, *et al.* "Esophageal Gross Tumor Volume Segmentation Using a 3D CNN." MICCAI, 2018.
[2] Cicek, O., *et al.*: "3d u-net: Learning dense volumetric segmentation from sparse annotation". In: MICCAI, 2016
[3] A. P. Harrison, *et al.*: Progressive and Multi-Path Holistically Nested Neural Networks for Pathological Lung Segmentation from CT Images, MICCAI, 2017