Thoracic Lymph Node Segmentation in CT imaging via Lymph Node Station Stratification and Size Encoding

Dazhou Guo[1], Jia Ge[2], Ke Yan[3], Puyang Wang[3], Zhuotun Zhu[4], Dandan Zheng[2], Xian-Sheng Hua[3], Le Lu[1], Tsung-Ying Ho[5], Xianghua Ye[2], and Dakai Jin[1]

[1]Alibaba DAMO Academy USA, NY, USA
[2]The First Affiliated Hospital Zhejiang University, Hangzhou, China
[3]Alibaba DAMO Academy, Hangzhou, China
[4]Johns Hopkins University, Baltimore, USA
[5] Chang Gung Memorial Hospital, Linkou, Taiwan
{guo2004131, dakai.jin}@gmail.com, hye1982@zju.edu.cn

Abstract. Visible lymph node (i.e., LN, short axis $\geq$ 5mm) assessment and delineation in thoracic computed tomography (CT) images is an indispensable step in radiology and oncology workflows. The high demanding of clinical expertise and prohibitive laboring cost motivate the automated approaches. Previous works focus on extracting effective LN imaging features and/or exploiting the anatomical priors to help LN segmentation. However, the performance in general is struggled with low recall/precision due to LN's low contrast in CT and tumor-induced shape and size variations. Given that LNs reside inside the lymph node station (LN-station), it is intuitive to directly utilize the LN-station maps to guide LN segmentation. We propose a stratified LN-station and LN size encoded segmentation framework by casting thoracic LN-stations into three super lymph node stations and subsequently learning the station-specific LN size variations. Four-fold cross-validation experiments on the public NIH 89-patient dataset are conducted. Compared to previous leading works, our framework produces significant performance improvements, with an average 74.2% (9.9% increases) in Dice score and 72.0% (15.6% increases) in detection recall at 4.0 (1.9 reduces) false positives per patient. When directly tested on an external dataset of 57 esophageal cancer patients, the proposed framework demonstrates good generalizability and achieves 70.4% in Dice score and 70.2% in detection Recall at 4.4 false positives per patient.

1 Introduction

Lymph node (LN) involvement assessment is an essential predictive or prognostic bio-marker in radiology and oncology. Precision quantitative LN analysis is an indispensable step in staging, treatment planning, and disease progression monitoring of cancers in thoracic region [10, 23]. Visual identification, measurement or

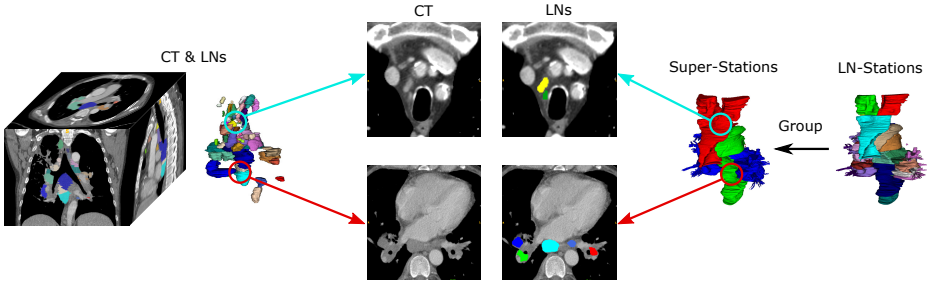

Fig. 1: An illustration of LN, LN-Station, and Super-Station. The top and bottom row show the LN contexts in LN-Station 2 (clear boundaries) and 8 (hard to discern).

delineation of thoracic LNs from CT scans are performed in the current clinical practice, which is tedious, time-consuming, expensive and expertise-demanding. Due to LN's low contrast to the surrounding anatomies and tumor-induced shape and size variations, it can be easily visually confused with vessels or muscles. Although enlarged LNs with short axis larger than 10mm are often considered as pathologically targeted lesions to be assessed [20], studies show that enlarged size alone is not the most reliable predictive factor for LN malignancy with only 60%-80% recall in lung cancer patients [6, 22]. Smaller LNs potentially involving metastasis should also be included for improved diagnostic accuracy [5, 21]. Thus an automated segmentation framework for both enlarged and smaller (short axis $\geq$5mm) throacic LNs is of high clinical importance.

Thoracic LN detection and segmentation has been exploited for more than a decade mainly focusing on extracting effective LN features, incorporating organ priors or utilizing advanced learning models [1–4, 7–9, 13–17, 25]. Early work often adopted the model-based or statistical learning-based methods. Feuerstein et al. built a statistical atlas for lymph node detection in chest CT [7]. Liu et al. detected and segmented LNs by first locating the approximate region of interest (mediastinum) and then integrating the spatial priors, intensity and shape features into a random forest classification model [15]. Recent deep learning based approaches have also been explored. Nogues et al. combined the holistically-nested neural networks and structured optimization to segment the LNs [16]. Bouget et al. segmented LNs in each cropped slabs and further ensembled with a full 3D UNet model that incorporated the organ priors [3]. Although being extensively studied, the general performance has been struggling with low recall/precision and is unable to deploy to clinical practice: e.g., for enlarged LNs (short axis $\geq$10mm), 70.4% recall at 4 false positives (FPs) per patient in [15]; for both enlarged and smaller LNs, 52.4% recall at 6 FPs per patient [3].

To tackle this tremendously challenging task, important anatomic knowledge and clinical reasoning insights from physicians should be leveraged. 1) almost all LNs reside in the lymph node station (LN-station) defined according to key anatomic organs or landmarks, i.e., thoracic LN-station recommended by International Association for the Study of Lung Cancer (IASLC) [19] and head &

neck LN-station outlined by American Academy of Otolaryngology–Head and Neck Surger (AAO-HN) [18]. Confining LN identification within the LN-station is beneficial, since the constrained searching and learning space can reduce FPs occurring at vessels, muscles, and soft-tissues in other body regions with similar local appearance. 2) LNs in different LN-stations often have distinct contexts and exhibit different levels of identification uncertainties. For example, thoracic LNs in stations 2 and 4 usually have clearer boundaries, while those in stations 7 and 8 can be more easily confused with adjacent vessels or esophagus (Fig. 1). Moreover, enlarged LNs often yield different texture patterns (e.g., calcified/necrosis) and shapes (e.g, tree/star-shape) to the smaller LNs. Further stratifying the LN segmentation task via different LN-stations and LN sizes could be beneficial.

Given the above key observations, we utilize the LN-station priors and propose a novel LN-station- and size-aware LN segmentation framework by explicitly incorporating the LN-station prior and learning the LN size variance. To achieve this, we first segment thoracic LN-stations 1 to 14 using a robust DeepStationing model in [11]. According to the LN-station context and physician's clinical experiences, we stratify and group the 14 LN-stations into 3 super lymph node stations (Super-Stations), i.e., LN-stations 1-4 (upper level), LN-stations 5-9 (lower level) and LN-stations 10-14 (lung regions). Then, a new deep segmentation network with multi-encoding paths are designed with each to focus on learning the LN features in a specific Super-Station (Fig. 2). Next, for explicitly learning LN's size variance, two decoding branches are adopted concentrating on the small and large LNs, respectively (Fig. 2). Results from the two decoding branches are then merged using a post-fusion module. For experimental evaluations, 1) we conduct 4-fold cross validations on a public LN dataset with 89 lung cancer patients and more than 2000 recently annotated LNs ($\geq$5mm) [3]. Our framework achieves an average 74.2% in Dice score and 72.0% in the detection recall at 4.0 false positives per patient (FP-PW), which significantly improves the segmentation and detection accuracy with at least 10% absolute Dice score improvement and 13% recall increase at 4.0 FP-PW, compared to previous leading LN segmentation/detection methods [3, 12, 16, 24]. 2) We further apply our pre-trained model to an external testing dataset of 57 esophageal caner patients with 360 thoracic LNs ($\geq$5mm labeled). We demonstrate good generalizability by obtaining 70.4% Dice score (Dice) (13.8% increases) and 70.2% Recall (17.0% increases) at 4.4 FP-PW as compared to the strong nnUNet baseline [12].

2 Method

Fig. 2 depicts the overview of our proposed LN-station-specific and size-aware LN segmentation framework. It contains three independent encoding paths based on the stratified Super-Stations and two decoding branches to learn size-specific LN's features. Together with the original CT image, post-fusion blocks leverage the predicted big and small LNs to generate the final prediction.

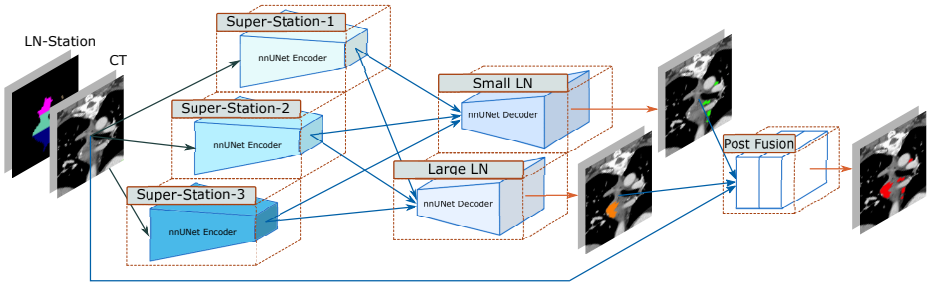

Fig. 2: Overall workflow of our proposed framework, which consists of Super-Station-based stratified encoders, size-aware decoder branches, and a post-fusion module.

2.1 LN-station Segmentation and Stratification

To utilize the LN-station priors, we first segment a set of 14 thoracic LN-stations. Motivated by [11], we adopt the key referencing organ guided LN-station segmentation model. As mentioned in [11], 6 key referencing organs are used: *esophagus, aortic arch, ascending aorta, heart, spine,* and *sternum*. Assuming N data instances, the training data is denoted as $\mathbf{D} = \{X_n, {Y_n}^K, {Y_n}^S, {Y_n}^L\}_{n=1}^N$, where X_n, Y_n^K, Y_n^S, Y_n^L denote the input CT image and ground-truth masks for the key referencing organs, LN-stations, and LNs, respectively. Let C_K and C_S denote the number of key organs and LN-stations, respectively. Dropping n for clarity, the key organ and LN-station segmentation models predict every voxel location j with a class c:

$$\hat{Y}^{\mathrm{K}}_{c_k}(j) = f^{\mathrm{K}}\left(Y^{\mathrm{K}}(j) = c_k \,|\, X; \mathbf{W}^{\mathrm{K}}\right), \quad \hat{\mathbf{Y}}^{\mathrm{K}} = \left[\hat{Y}^{\mathrm{K}}_1 \ldots \hat{Y}^{\mathrm{K}}_{C_{\mathrm{K}}}\right], \tag{1}$$

$$\hat{Y}^{\mathrm{S}}_{c_s}(j) = f^{\mathrm{S}}\left(Y^{\mathrm{S}}(j) = c_s \,|\, X, \hat{\mathbf{Y}}^{\mathrm{K}}; \mathbf{W}^{\mathrm{S}}\right), \quad \hat{\mathbf{Y}}^{\mathrm{S}} = \left[\hat{Y}^{\mathrm{S}}_1 \ldots \hat{Y}^{\mathrm{S}}_{C_{\mathrm{S}}}\right], \tag{2}$$

where $f^{(*)}(.)$ denotes the network functions, $\mathbf{W}^{(*)}$ represents the corresponding network parameters, and $\hat{Y}^{(*)}_{c_*}$ for the predicted segmentation maps.

According to LN-station context and physician's clinical experience, we combine the predicted LN-stations into three Super-Stations, i.e., LN-stations 1-4, 5-9 and 10-14. To avoid potential LN-station under-segmentation, we dilate each Super-Station with a diameter of 15 mm. We then use the dilated Super-Station binary maps to 'mask' Super-Station covered CT images (abbreviated as mCT), while setting other voxel values to a constant of -1024. We thereafter obtain three Super-Station-masked mCT images. Please note that the inputs of three encoders are independent. For instance, for Super-Station-1 encoder, its input is the mCT masked using dilated Super-Station-1 and the LN-stations 1-4 maps. Let $\hat{Y}_1^{\mathbb{S}+} = [\hat{Y}_c^{\mathrm{S}}]_{c=1}^{4}$, $\hat{Y}_2^{\mathbb{S}+} = [\hat{Y}_c^{\mathrm{S}}]_{c=5}^{9}$, and $\hat{Y}_3^{S+} = [\hat{Y}_c^{\mathrm{S}}]_{c=10}^{14}$ denote the grouped LN-stations maps. Each encoder targets Super-Station-specific LN features.

$$\hat{\mathbf{H}}_i^{\mathbb{S}+} = f_i^{\mathbb{S}+}\left(X_i^m, \hat{Y}_i^{\mathbb{S}+}; \mathbf{W}_i^{\mathbb{S}+}\right), \quad i \in \{1, 2, 3\}, \tag{3}$$

where X_i^m denotes the mCT and $\hat{\mathbf{H}}_{(*)}^{\mathbb{S}+}$ denotes the output feature maps of the stratified encoder.

2.2 LN Size Stratification and Post-fusion

Considering that enlarged LNs often yield different texture patterns (e.g., calcified/necrosis) or shapes (e.g., tree/star-shape) as compared to smaller ones, we further introduce two decoding branches to learn size-specific LN features. For the annotated LNs, we manually separate the enlarged LNs ($L+$) whose short-axes are greater than 10mm, and vice versa ($L-$). Each decoder is supervised using the respect enlarged/small LN labels. The input of each decoder branch is the channel-wise concatenation of the feature maps from three Super-Stations encoding paths: $\hat{\mathbf{H}}^{\mathbb{S}+} = \left[\hat{\mathbf{H}}_1^{\mathbb{S}+}, \hat{\mathbf{H}}_2^{\mathbb{S}+}, \hat{\mathbf{H}}_3^{\mathbb{S}+}\right]$.

$$\hat{Y}^{L+} = f^{L+}\left(\hat{\mathbf{H}}^{\mathbb{S}+}; \mathbf{W}^{L+}\right), \quad \hat{Y}^{L-} = f^{L-}\left(\hat{\mathbf{H}}^{\mathbb{S}+}; \mathbf{W}^{L-}\right), \tag{4}$$

where $\hat{Y}^{L+}$ and $\hat{Y}^{L-}$ denote the output prediction maps of the large- and small-LN decoders, respectively. The output feature maps of each decoder are additionally combined with the original CT image and input to a simple post fusion module. The post fusion module is created using the first two nnUNet convolutional blocks without the pooling function:

$$\hat{Y}^{L} = f^{P}\left(X, \hat{Y}^{L+}, \hat{Y}^{L-}; \mathbf{W}^{L-}\right). \tag{5}$$

The proposed segmentation framework explicitly encodes both the LN-station- and size-specific information in the model training. The final prediction is $\hat{Y}^{L}$.

3 Experimental Results

Dataset. *For the LN model development and validation*, we used 89 patients with contrast-enhanced venous-phase CT scans from the pubic NIH dataset[1], where more than 2000 thoracic LN instance labels were recently annotated by [3]. The average CT image size is $512 \times 512 \times 616$ voxels with the average voxel resolution of $0.8 \times 0.8 \times 1.2\text{mm}^3$. The average LN size is $7.7 \times 6.6 \times 9.4\text{mm}^3$. *For the external testing*, we collected additional 57 contrast-enhanced venous-phase CT scans of esophageal cancer patients underwent surgery and/or radiotherapy treatment. The average CT image size is $512 \times 512 \times 80$ voxels with the average voxel resolution of $0.7 \times 0.7 \times 5.0\text{mm}^3$. A board-certified radiation oncologist with more than 15 years of experience labeled each patient with visible LNs ($\geq$ 5 mm). The average LN size of external testing set is $9.5 \times 9.0 \times 9.1$mm. Moreover, to develop the LN-station segmentation model, 3D masks of thoracic LN-stations 1-14, and 6 key referencing organs were annotated in this esophageal dataset according to the IASLC guideline [19].

[1] `https://wiki.cancerimagingarchive.net/display/Public/CT+Lymph+Nodes`

Table 1: LN segmentation performance on the NIH dataset using 4-fold cross-validation. We abbreviate Super-Station-based stratified encoder as S.S.E. and size-aware decoder branches as S.A.D. We evaluate the segmentation performance for all annotated LNs (upper) and only enlarged LNs (lower). The best performance scores are highlighted in **bold**. Our *full version* is: 'mCT + S.S.E. + S.A.D. + post fusion module'. The 1^{st} row 'CT Only' is the default nnUNet [12] performance.

	Dice	Recall	Recall-PW	FP-PW
All LNs – Short Axis $\geq$ 5mm				
CT Only (Isensee et al. [12])	58.2±17.1	55.3	56.9±21.4	6.2±4.4
CT + LNS (Ours)	66.3±13.5	60.1	61.5±19.7	4.9±3.8
mCT + LNS (Ours)	70.4±12.6	66.2	67.4±20.1	**3.8±2.8**
mCT + S.S.E. (Ours)	72.1±13.8	69.4	70.8±19.1	4.1±2.6
mCT + S.S.E. + S.A.D. (Ours)	74.1±14.8	71.2	71.9±20.4	4.9±3.6
Nogues et al. 2016 [16]	57.7±19.4	57.2	62.2±20.7	4.9±4.1
Yan et al. 2020 [24]	60.6±15.1	59.3	64.8±18.3	6.1±2.5
Bouget et al. 2021 [3]	64.3±14.5	56.4	55.2±18.0	5.9±3.7
Full version	**74.2±14.7**	**72.0**	**72.4±19.0**	4.0±2.9
Enlarged LNs – Short Axis $\geq$ 10mm				
Nogues et al. 2016 [16]	58.6±14.2	79.2	81.4±18.5	3.1±3.5
Yan et al. 2020 [24]	65.6±12.1	83.3	89.8±15.3	4.0±3.5
Isensee et al. 2021 [12]	61.8±12.4	85.3	89.5±15.4	3.1±3.9
Bouget et al. 2021 [3]	-	82.7	88.6±15.6	-
Full version	**79.1±9.8**	**89.8**	**92.4±12.6**	**2.4±2.3**

Implementation details. We adopt '3d-fullres' version of nnU-Net [12] with Dice+CE losses as our backbone modules. Each encoder is the same as the default nnUNet encoder, and each block contains two "Conv+InstanceNorm+IReLU" layers. With additional two skip-connections, each decoder block receives 2x its original input channels. The default nnUNet's deep-supervision is not used in our experiment. Instead, we apply two side supervisions for the two decoding branches using the enlarged and small LN labels, respectively. We use the default nnUNet data augmentation settings for our model training, and set the patch size to $96 \times 128 \times 32$. We implemented our framework using PyTorch and trained on an NVIDIA Tesla V100. The total training epochs is 1000. The average training time is 5.5 GPU days.

Evaluation. For the NIH dataset, extensive four-fold cross-validation (CV), separated at the patient level, was conducted. The esophageal dataset was held out as an external testing dataset. We follow the LN evaluation metrics in [3] and calculate Dice, instance detection Recall, patient-wise detection recall (Recall-PW) and the patient-wise false positive numbers (FP-PW).

Results of LN-station segmentation. We first evaluate the performance of the LN-station segmentation model. The average LN-station segmentation performance is: Dice 81.2±5.8%, Hausdorff distance (HD) 9.6±4.2mm, and average surface distance (ASD) 0.9 ± 0.6mm. In our experiment, we select a diameter of 15mm to dilate the predicted and grouped Super-Stations to cover the thoracic LNs, which might be missed in the original LN-station prediction due to the under-segmentation. Meanwhile, the dilated Super-Stations should not include

Table 2: LN segmentation performance on the collected in-house 57 esophageal cancer patients testing set. The best performance scores are highlighted in **bold**.

	Dice	Recall	Recall-PW	FP-PW
All LNs – Short Axis $\geq$ 5mm				
Nogues et al. 2016 [16]	55.9±20.6	53.4	54.5±25.7	6.3±4.5
Yan et al. 2020 [24]	54.2±17.2	60.3	62.9±21.3	9.5±6.4
Isensee et al. 2021 [12]	56.6±18.4	53.2	55.3±22.4	5.8±3.9
Full version	**70.4±14.7**	**70.2**	**70.8±20.2**	**4.4±3.3**
Enlarged LNs – Short Axis $\geq$ 10mm				
Nogues et al. 2016 [16]	56.5±17.6	77.4	80.5±17.9	4.5±3.6
Yan et al. 2020 [24]	64.2±16.2	80.1	86.9±15.3	7.6±4.4
Isensee et al. 2021 [12]	59.2±15.2	79.6	83.2±16.7	3.7±4.1
Full version	**74.9±12.4**	**85.8**	**91.8±13.5**	**2.4±2.7**

too many similar tissues such as vessels. The quantitative instance/volume LN coverage using predicted LN-stations is reported in the supplementary materials.

Quantitative evaluation in NIH dataset. Table 1 outlines the quantitative comparisons of different input and model setups when evaluated in the NIH dataset: 1) only CT images, 2) early fusion of CT and LN-station (CT + LNS), 3) CT masked using the dilation of the whole LN-station region (mCT + LNS), 4) Super-Station-stratified encoders with the default single UNet decoder (mCT + S.S.E.), 5) Super-Station-stratified encoders + size-aware decoders ***without*** post fusion (mCT + S.S.E. + S.A.D.). Several observations can be drawn. First, LN segmentation exhibits the lowest performance with an average 58.2% Dice and 55.3% Recall at 6.2 FP-PW when using only CT. When using LN-station as an additional input channel, all metrics show remarked improvements: 8.1% and 4.8% increase in Dice and Recall, and 1.3 reduction in FP-PW. This demonstrates the importance and effectiveness of using LN-stations for LNs segmentation. Second, when adopting the 'mCT + LN-station' setup, the performance is markedly improved with another 6.1% boosted Recall and 1.1 reduced FP-PW. This indicates that constraining the learning space within the mCT region (hence, eliminating the confusing anatomy tissues in the irrelevant regions) make the LN identification task much easier. Third, the LN-station- and size-aware LN segmentation schemes are effective, since both S.S.E. and S.A.D. modules yield marked improvements boosting the Dice and Recall to 74.1% and 71.2, respectively. Finally, equipped with a simple post-fusion module, the final proposed model can further reduce the FP-PW from 4.9 to 4.0 while preserving the high Dice and Recall as compared to the mCT + S.S.E. + S.A.D. model.

Table 1 also shows the performance comparisons in NIH dataset between our proposed framework and four leading methods [3, 12, 16, 24]. Among the comparison methods, the best Recall of 59.3% at 6.1 FP-WP is achieved by [24], which is the leading approach for 3D universal lesion detection. It can be seen the proposed framework significantly outperforms [24] by 14.4% Dice, 12.7% Recall, and 2.1 in FP-PW. When analyzing the performance of enlarged LNs (short axis $>$ 10 mm) commonly studied in previous works, our framework achieves a high

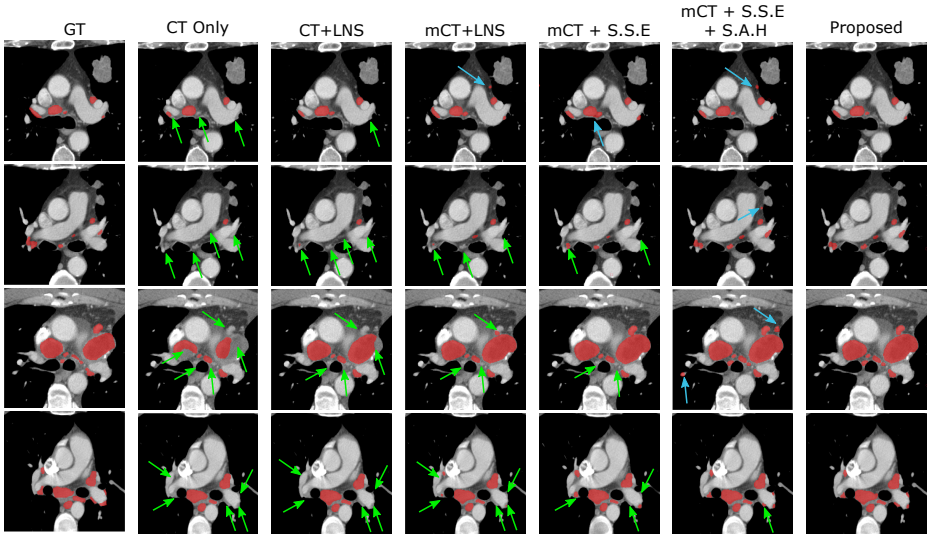

Fig. 3: Examples of LN segmentation results using different setups. *green* and *blue* arrows are used to depict under-segmentations and FPs. It can be observed that severe under-segmentation exists in the CT only method leading to low Dice and Recall (2nd column). In contrast, when LN-station information is explicitly incorporated into the model, more LNs can be correctly identified (from 3rd to 5th columns). Size-stratification (S.A.D) can further improve the LNs segmentation Recall, while the proposed final model suppresses some FPs while maintaining the high Recall.

Recall of 89.8% at 2.4 FP-PW. In comparison, previous leading methods exhibit inferior performance with the best 85.3% Recall at 3.1 FP-PW [12].

External testing on esophageal dataset. The independent external testing results on the esophageal dataset are illustrated in Table 2. The proposed framework demonstrates good generalizability by achieving 70.4% Dice and 70.2% Recall at 4.4 FP-PW, which are comparable to those in the NIH dataset. For the failure cases, under-segmentation along the z-direction for LNs in the inferior mediastinal region is observed. The assumed reasons might be: 1) unclear boundaries of the inferior mediastinal LNs, and 2) most LNs are relatively short in z-direction and the model might bias toward the majority average. For the enlarged LNs, our framework also shows robust performance of 74.9% Dice and 85.8 Recall at 2.4 FP-PW. The assumed reasons of achieving good generalizability might be that segmenting LNs in a much confined Super-Station region is comparably easy and robust. In contrast, the previous second best performing detection method [24] yields low generalizability as its FP-PW significantly increased from 6.1 (NIH) to 9.5 (external) for all LNs and from 4.0 (NIH) to 7.6 (external) for enlarged LNs.

4 Conclusion

In this paper, we propose a novel LN-station-specific and size-aware LN segmentation framework by explicitly utilizing the LN-station priors and learning the LN size variance. We first segment thoracic LN-stations and then group the LN-stations into 3 super lymph node stations, based on which a multi-encoder deep network is designed to learn LN-station-specific LN features. For learning LN's size variance, we further stratify decoding path into two decoding branches to concentrate on learning the small and large LNs, respectively. Validated on the public NIH dataset and further tested on the external esophageal dataset, the proposed framework demonstrates high LN segmentation performance while preserving good generalizability. Our work represents an important step towards the reliable and automated LN segmentation.

References

1. Barbu, A., Suehling, M., Xu, X., Liu, D., Zhou, S.K., Comaniciu, D.: Automatic detection and segmentation of lymph nodes from ct data. IEEE Transactions on Medical Imaging **31**(2), 240–250 (2011)
2. Bouget, D., Jørgensen, A., Kiss, G., Leira, H.O., Langø, T.: Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in ct data for lung cancer staging. International journal of computer assisted radiology and surgery **14**(6), 977–986 (2019)
3. Bouget, D., Pedersen, A., Vanel, J., Leira, H.O., Langø, T.: Mediastinal lymph nodes segmentation using 3d convolutional neural network ensembles and anatomical priors guiding. arXiv preprint arXiv:2102.06515 (2021)
4. Chao, C.H., Zhu, Z., Guo, D., Yan, K., Ho, T.Y., Cai, J., Harrison, A.P., Ye, X., Xiao, J., Yuille, A., et al.: Lymph node gross tumor volume detection in oncology imaging via relationship learning using graph neural network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 772–782. Springer, Cham (2020)
5. Choi, S.B., Han, H.J., Park, P., Kim, W.B., Song, T.J., Choi, S.Y.: Systematic review of the clinical significance of lymph node micrometastases of pancreatic adenocarcinoma following surgical resection. Pancreatology **17**(3), 342–349 (2017)
6. De Langen, A.J., Raijmakers, P., Riphagen, I., Paul, M.A., Hoekstra, O.S.: The size of mediastinal lymph nodes and its relation with metastatic involvement: a meta-analysis. European journal of cardio-thoracic surgery **29**(1), 26–29 (2006)
7. Feuerstein, M., Glocker, B., Kitasaka, T., Nakamura, Y., Iwano, S., Mori, K.: Mediastinal atlas creation from 3-d chest computed tomography images: application to automated detection and station mapping of lymph nodes. Medical image analysis **16**(1), 63–74 (2012)
8. Feulner, J., Zhou, S.K., Hammon, M., Hornegger, J., Comaniciu, D.: Lymph node detection and segmentation in chest ct data using discriminative learning and a spatial prior. Medical image analysis **17**(2), 254–270 (2013)
9. Feulner, J., Zhou, S.K., Huber, M., Hornegger, J., Comaniciu, D., Cavallaro, A.: Lymph node detection in 3-d chest ct using a spatial prior probability. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2926–2932. IEEE (2010)

10. Goldstraw, P., Crowley, J., Chansky, K., Giroux, D.J., Groome, P.A., Rami-Porta, R., Postmus, P.E., Rusch, V., Sobin, L., for the Study of Lung Cancer International Staging Committee, I.A., et al.: The iaslc lung cancer staging project: proposals for the revision of the tnm stage groupings in the forthcoming (seventh) edition of the tnm classification of malignant tumours. Journal of thoracic oncology **2**(8), 706–714 (2007)
11. Guo, D., Ye, X., Ge, J., Di, X., Lu, L., Huang, L., Xie, G., Xiao, J., Lu, Z., Peng, L., et al.: Deepstationing: thoracic lymph node station parsing in ct scans using anatomical context encoding and key organ auto-search. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–12. Springer (2021)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods pp. 1–9 (2021)
13. Iuga, A.I., Carolus, H., Höink, A.J., Brosch, T., Klinder, T., Maintz, D., Persigehl, T., Baeßler, B., Püsken, M.: Automated detection and segmentation of thoracic lymph nodes from ct using 3d foveal fully convolutional neural networks. BMC Medical Imaging **21**(1), 1–12 (2021)
14. Li, Z., Xia, Y.: Deep reinforcement learning for weakly-supervised lymph node segmentation in ct images. IEEE Journal of Biomedical and Health Informatics **25**(3), 774–783 (2020)
15. Liu, J., Hoffman, J., Zhao, J., Yao, J., Lu, L., Kim, L., Turkbey, E.B., Summers, R.M.: Mediastinal lymph node detection and station mapping on chest ct using spatial priors and random forest. Medical physics **43**(7), 4362–4374 (2016)
16. Nogues, I., Lu, L., Wang, X., Roth, H., Bertasius, G., Lay, N., Shi, J., Tsehay, Y., Summers, R.M.: Automatic lymph node cluster segmentation using holistically-nested neural networks and structured optimization in ct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 388–397. Springer (2016)
17. Roth, H.R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., Kim, L., Summers, R.M.: Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE transactions on medical imaging **35**(5), 1170–1181 (2016)
18. Roubbins, K.T., Clayman, G., Levine, P.A., Medina, J., Sessions, R., Shaha, A., Som, P., Wolf, G.T., et al.: Neck dissection classification update: revisions proposed by the american head and neck society and the american academy of otolaryngology–head and neck surgery. Archives of otolaryngology–head & neck surgery **128**(7), 751–758 (2002)
19. Rusch, V.W., Asamura, H., Watanabe, H., Giroux, D.J., Rami-Porta, R., Goldstraw, P.: The iaslc lung cancer staging project: a proposal for a new international lymph node map in the forthcoming seventh edition of the tnm classification for lung cancer. Journal of thoracic oncology **4**(5), 568–577 (2009)
20. Schwartz, L., Bogaerts, J., Ford, R., Shankar, L., Therasse, P., Gwyther, S., Eisenhauer, E.: Evaluation of lymph nodes with recist 1.1. European journal of cancer **45**(2), 261–267 (2009)
21. Stanley Leong, P., Tseng, W.W.: Micrometastatic cancer cells in lymph nodes, bone marrow, and blood: Clinical significance and biologic implications. CA: a cancer journal for clinicians **64**(3), 195–206 (2014)
22. T. McLoud, C., Bourgouin, P.M., Greenberg, R.W., Kosiuk, J.P., Templeton, P.A., Shepard, J.A., Moore, E.H., Wain, J.C., Mathisen, D.J., Grillo, H.C.: Bronchogenic

carcinoma: analysis of staging in the mediastinum with ct by correlative lymph node mapping and sampling. Radiology **182**(2), 319–323 (1992)
23. Terán, M.D., Brock, M.V.: Staging lymph node metastases from lung cancer in the mediastinum. Journal of Thoracic Disease **6**(3), 230 (2014)
24. Yan, K., Cai, J., Zheng, Y., Harrison, A.P., Jin, D., Tang, Y., Tang, Y., Huang, L., Xiao, J., Lu, L.: Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct. IEEE Transactions on Medical Imaging **40**(10), 2759–2770 (2020)
25. Zhu, Z., Jin, D., Yan, K., Ho, T.Y., Ye, X., Guo, D., Chao, C.H., Xiao, J., Yuille, A., Lu, L.: Lymph node gross tumor volume detection and segmentation via distance-based gating using 3d ct/pet imaging in radiotherapy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 753–762. Springer, Cham (2020)