Metastatic Lymph Node Station Classification in Esophageal Cancer via Prior-guided Supervision and Station-Aware Mixture-of-Experts

Haoshen Li[1,2⋆], Yirui Wang[2⋆], Qinji Yu[2,5⋆], Jie Zhu[3], Ke Yan[2,4], Dazhou Guo[2], Le Lu[2], Bin Dong[6,7], Li Zhang[1], Xianghua Ye[8], Qifeng Wang[3], and Dakai Jin[2]

[1]Center for Data Science, Peking University [2]DAMO Academy, Alibaba Group
[3]Sichuan Cancer Hospital [4]Hupan Lab, 310023 [5]Shanghai Jiao Tong University
[6]Beijing International Center for Mathematical Research and the New Cornerstone Science Laboratory, Peking University
[7]Center for Machine Learning Research, Peking University
[8]The First Affiliated Hospital, Zhejiang University
zhangli_pku@pku.edu.cn, littlecancer@163.com

Abstract. Assessing lymph node (LN) metastasis in CT is critical for esophageal cancer treatment planning. While clinical criteria are commonly used, the diagnostic accuracy is low with sensitivities ranging from 39.7% to 67.2% in previous studies. Deep learning would have the potential to improve it by learning from large-scale accurately labeled data. However, from the surgical procedure in LN dissection, pathological report only indicates the number of dissected LNs in each lymph node station (LN-station) with the number of metastatic ones found in the respective LN-station. So, it is difficult to establish one-to-one pairing between LN instances observed in CT and their metastasis status confirmed in the pathological report. In contrast, gold reference labels on LN-station metastasis can be readily retrieved from pathology reports at scale. Hence, instead of distinguishing LN instance metastasis, we directly classify LN-station metastasis using pathology-confirmed station labels. We first segment mediastinal LN-stations automatically to serve as input for classification. Then, to improve classification performance, we automatically segment all visible LN instances in CT and design a new LN prior-guided attention loss to explicitly regularize the network to focus on regions of suspicious LNs. Furthermore, considering the varying appearances and contexts of different LN-station, we propose a station-aware mixture-of-experts module, where the expert is trained to specialize in a group of LN-stations by learning to route each LN-station group tokens to the corresponding expert. We conduct five-fold cross-validation on 1,153 esophageal cancer patients with CT and pathology reports (the largest study to date), and our method significantly outperforms state-of-the-art approaches by 2.26% in AUROC.

Keywords: Lymph node station classification · Attention Aggregation · Mixture-of-Experts (MoE).

⋆ Equal contribution.

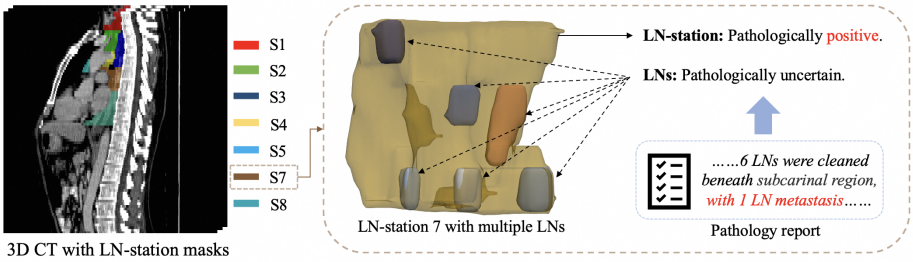

Fig. 1: An illustration of the difficulty to establish the one-to-one pairing between LN instances observed in preoperative CT and their metastatic status reported in post-surgery pathology report. From the report, LN-station-7 is known to be metastasis with one metastatic LN. Yet, there are six visible LNs at station-7 in CT, and it is difficult for radiologists to identify which LN instance is positive.

1 Introduction

Esophageal cancer (EC) is the sixth leading cause of cancer death worldwide [25]. LN metastasis is one of the most important prognostic factor in EC. Accurate preoperative identification of LN metastasis is essential to determine treatment decisions (surgery vs. neoadjuvant) and plans (surgical resection area or clinical target volume [CTV] in radiotherapy) [17]. Contrast-enhanced computed tomography (CT) is the standard imaging tool for assessing LN metastasis before treatment. Although criteria such as RECIST [23], morphology [22] and texture characteristics [1] are widely used in clinical practice, they often yield under-estimated diagnostic performance [2, 4, 12, 14, 27] (*e.g.*, sensitivity ranged from 39.7% to 67.2% with specificity around 80.0%). Thus, there is great need to develop an effective computer-aided diagnosis (CAD) solutions for this task.

With the success of deep learning in medical imaging CAD tasks [24, 30], preliminary attempts have been made for its application on LN abnormality diagnosis [19,26]. A major limitation of previous work is that the RECIST criterion (the short axis $\geq$ 10mm) is used as the LN metastasis reference label, which is not accurate. As mentioned, the RECIST criterion has a low sensitivity (39.7% to 67.2%) in identifying pathologically-confirmed LN metastasis, which is not suitable to serve as the gold reference label for training. To learn from "true" metastasis LN labels (pathologically confirmed), few works use LN dissection results of pathology reports to generate the gold reference label of LN instances in CT for patients with head and neck cancer [9–11]. However, as shown in Fig. 1, due to the practical surgery procedure of LN dissection in esophageal cancer, the pathology report only indicates the number of dissected LNs in each LN-station and the number of pathologically metastatic ones within this station. This makes it *extremely difficult and unscalable (without mentioning the time cost) for human experts to establish the one-to-one pairing between LN instances observed in CT and their metastatic status reported by pathological examination.*

Although individual LN metastasis status is difficult to confirm from the pathology report, the gold reference label for LN-station metastasis can be easily recovered from the pathology report at scale, i.e., a LN-station is labeled as metastasis if there exists at least one metastatic LNs reported at this station. Importantly, LN-station metastasis is sufficient in clinical usage since LN surgical dissection is indeed performed station-wise [7] and the GTV region in radiotherapy [13] is delineated based on the range of LN-station instead of individual LNs. Motivated by this observation, one can directly predict the LN-station metastasis status using the pathology-confirmed LN-station label, which eliminates the need for the labor-intensive yet ambiguous LN instance-wise label pairing process, allowing for the use of large-scale datasets.

In this work, we tackle the task of LN-station metastasis classification of EC patients using a large-scale dataset (> 1000 patients) with pathology-confirmed labels. We first segment the mediastinal LN-stations (stations 1 to 8) automatically using a robust DeepStationing model [5]. Then, the ROI for each LN-station is cropped in CT and used as input for classification. Since metastasis status for LN-station is determined by the metastasis status of LNs within it, we also auto-detect all visible LN instances ($\geq$ 5mm) in CT and utilize these LN priors as attention to guide the LN-station classification, i.e., a new LN prior-guided attention loss is introduced to explicitly regularize the network to focus on suspicious LN regions. Moreover, considering that different LN-stations have distinct appearances and contexts, we divide LN-station into multiple groups based on their location, and propose a Station-Aware Mixture-of-Expert (SA-MoE) module to guide each expert focusing on a specific group of LN-stations by learning to route each LN-station group tokens to the corresponding expert. This allows the network to learn metastasis features from different LN-stations.

The main contributions of this work are as follows:

- We address clinically essential yet under-studied task of LN-station metastasis diagnosis, leveraging large-scale dataset and circumventing the challenges to establish one-to-one LN matching between CT and pathology report.
- To solve the LN-station classification task, we propose a new LN prior-guided attention loss and a station-aware mixture-of-experts module, allowing us to learn effective metastasis imaging characteristics from different LN-stations.
- Using extensive five-fold cross-validation of 1153 EC patients with preoperative CT and postoperative pathological reports (as the largest cohort to date), our method significantly outperforms previous state-of-the-art approaches by 2.26% absolute AUROC value on this formidable task.

2 Method

Prerequisite Prior Segmentation: An overview of the proposed framework is illustrated in Fig. 2(a). We first generate the input ROI by segmenting the mediastinal LN-stations (station 1 to 8 per IASLC [20] guideline) automatically using a robust DeepStationing model [5], with an average of 81.1% Dice score and

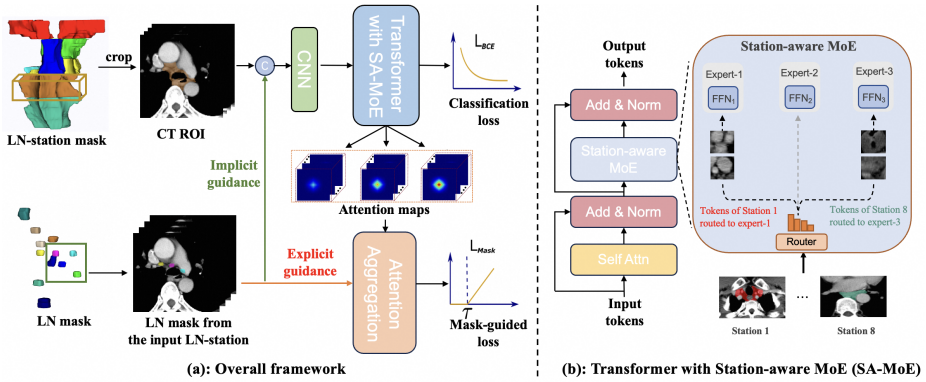

Fig. 2: An overview of the proposed LN-station metastasis classification method. **(a)** LNS masks are utilized to crop the LNS ROI from the CT scan, while the corresponding LN masks are not only concatenated with the image to serve as input for the model but also provide explicit supervision, guiding the network's focus on the LN region. **(b)** We replace original FFN layers with Station-aware MoE layers, which learns the mapping of tokens to their corresponding LN-station groups and routes the tokens to the corresponding experts.

0.9 mm average surface distance error. To provide direct supervision of LN priors for this task, LN instances are detected and segmented by an ensemble of recent transformer-based LN detectors [28, 29], which has $\geq$ 80% average detection recall at 4 false positives per patient on all visible LN instances (with a short axis $\geq$ 5 mm). We concatenate the region of interest (ROI) of the LN-station with its corresponding LN masks as input to the classification model.

2.1 Classification Network with LN Prior-guided Attention Loss

MobileViTv2 [16] of a hybrid design of ConvNets and transformers can achieve improved representation ability without compromising computational efficiency. Inspired by this design, we extend the 2D MobileViTv2 into 3D architecture and modify the last three transformer blocks to enable global semantic grouping and localized attention parsing. Specifically, we replace the original transformer block's feedforward network (FFN) layer with our proposed SA-MoE layer, which groups adjacent stations to better accommodate subtle inter-station LN feature variations. In addition, we introduce a LN prior-guided attention loss that explicitly regulates the network's attention to high-risk LN regions, to facilitate more accurate analysis of LN metastasis characteristics.

Due to anatomical complexity and subtle visual differences, a deep network can inadvertently associate irrelevant imaging features outside the LN region with metastasis labels and lead to degraded classification performance. We address this challenge through an implicit multi-channel embedding and an explicit intermediate LN location supervision. Specifically, we concatenate the CT RoI

with a binary LN instance mask. This auxiliary mask input implicitly emphasizes and restricts the search space of the network, improving both the efficiency and accuracy of feature learning. We propose a mask-guided margin loss that explicitly encourages the network to focus solely on regions where suspicious LNs are present. We utilize the attention score maps produced by the transformer self-attention mechanism and enforce higher attention values within the predicted LN mask regions. To achieve this, we extract three attention maps from the last layer of each SA-MoE block, interpolate them to the same resolution, and average them into a single attention map (see Fig. 2 for an illustration). At last, we use this aggregated attention map to estimate the total attention allocated to suspicious LN regions and compare it with a predefined threshold τ. Formally, the LN mask-guided margin loss can be written as:

$$\mathcal{L}_{\text{mask}} = \max\left(\tau - \sum\left(\mathcal{A} \times \mathcal{M}\right), 0\right). \tag{1}$$

where $\mathcal{A}$ is the aggregated multi-scale attention map, $\mathcal{M}$ is the LN instance mask, and $\times$ denotes the pixel-wise multiplication. This mask-guided margin loss can explicitly incorporate LN location priors, and meanwhile it tolerates the minor uncertainties/errors in automatic LN segmentation.

2.2 Station-aware Mixture-of-Experts Model

Given the varying appearances and contexts among different LN-stations, utilizing LN-station class priors is beneficial for the classification. Thus, we propose a station-aware mixture-of-experts (SA-MoE) module to learn the LN-station specific characteristics. Depending on the location of the LN stations, we stratify the LN-stations into three groups: the upper station (S1 and S2), the middle station (S3 and S4), and the lower station (S5−S8). Based on this stratification, we train the router in SA-MoE block, to distribute tokens to the specific expert that corresponds to the LN-station. Given the station groups $\{G_i\}_{i=1}^{3}$, and the input token $x \in G_i$, we assign the expert E_i to G_i, and use the route layer (a linear layer) to get the logits of different experts and assign the token to the expert corresponding to the maximum value of the logits. For training the route layer, we calculate the cross-entropy loss between logits p_i (probability of the selection expert E_i) and token labels $L(d_i)$ (target expert for token x from the station group G_i).

$$\mathcal{L}_{\text{SA-MoE}} = -\sum_{i=1}^{3} 1(L(d_i) = i) log(p_i(x)) \tag{2}$$

The final loss for the classification can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda_1 \mathcal{L}_{\text{mask}} + \lambda_2 \mathcal{L}_{\text{SA-MoE}} \tag{3}$$

where we empirically set $\lambda_1 = 0.5$ and $\lambda_2 = 1$ according to the results in the validation set. Note that we may also choose other expert numbers, such as two experts (S1−S4 in one group and S5−S8 in another group), or fine-grained experts where each station is assigned to a dedicated expert. We show these comparative results in the ablation study.

Table 1: Quantitative LN-station metastasis classification performance. Best results are shown in **bold**. ATTN-Loss: LN prior-guided attention loss.

Method	Params	AUC	S@R75	S@R80	R@S75	R@S80
ResNet-18 [6]	126.50M	83.92	78.48	72.97	78.58	72.10
ResNet-50 [6]	176.08M	83.81	75.48	71.27	75.72	70.59
MobileNetv2 [21]	10.58M	84.76	78.68	73.57	78.88	73.91
ViT-T [3]	153.36M	82.12	73.97	69.87	73.76	67.42
Swin-T [15]	105.19M	84.59	79.08	75.68	80.54	72.85
MobileViTv2 [16]	16.99M	84.77	79.18	73.97	79.03	73.76
MobileViTv2+M4oE [8]	18.43M	85.62	80.18	77.38	82.20	75.26
MobileViTv2+V-MoE [18]	17.00M	86.06	80.68	76.68	82.05	76.32
MobileViTv2+ATTN-Loss (ours)	16.99M	85.96	79.28	76.48	82.65	73.91
MobileViTv2+SA-MoE (ours)	17.00M	87.51	83.88	79.88	85.37	79.79
MobileViTv2+ATTN-Loss+SA-MoE (ours)	17.00M	**88.32** (+2.26%)	**84.28** (+3.60%)	**80.68** (+3.30%)	**85.67** (+3.47%)	**80.84** (+4.52%)

3 Experiments and Results

We collected a dataset consisting of 1153 patients with esophageal cancer who underwent esophagectomy treatment at a high-volume cancer institution. Each patient has a preoperative contrast-enhanced CT scan and a detailed pathological report after surgery. LN-station labels are determined by LN dissection results indicated in the pathology report. A LN-station is labeled as benign if all dissected LNs at this station are benign, while labeled as metastasis if there exists at least one metastatic LNs at this station. The median CT image size is $512 \times 512 \times 91$ voxels and the median resolution is $0.795 \times 0.795 \times 5.0$mm. For evaluation, we conducted a five-fold cross-validation, split at the patient level.

Implementation Details: 3D training image patches are generated by cropping $96 \times 96 \times 32$ ROIs on the CT image and the LN mask, respectively, centered at each LN-station. We further use each LN-station binary map to 'mask' the CT image by setting voxels outside the LN-station to a constant value of -1024, which is based on clinical guidance and serves to mask out regions outside the lymph node stations. We adopt the same number of transformer layers as in [16]. In addition, we empirically set the threshold τ in the margin loss at 0.8 because it produces the best validation performance. For training, AdamW optimizer with a learning rate of 9.6e-4 and weight decay of 5e-4 is adopted. We employ a mini-batch size of 32. The network is trained for 250 epochs for convergence, and we select the model with the best performance on validation set for testing.

Comparison Setup: Since there is no previous LN-station classification work, we use our preprocessing workflow and compare against six widely used classification networks, including three CNN networks: ResNet18 [6], ResNet50 [6] and MobileNetv2 [21], two Transformer networks: ViT [3], Swin-T [15], and one CNN+Transformer network MobileViTv2 [16]. We also compare the proposed

Table 2: Quantitative results (in term of AUC) in LN-station subgroups. Mixing refers to whether or not the model is trained using data of all LN-stations. First row shows the results where three MobibleViTv2 are separately trained using LN-station Group1, Group2 and Group3, respectively.

Methods	Mixing	Params	Group1	Group2	Group3	Mean
MobileViTv2	×	16.99M × 3	83.97	71.28	83.03	79.43
MobileViTv2	✓	16.99M	84.95	72.49	83.98	80.47
+Multi-branch [31]	✓	35.15M	<u>86.95</u>	71.06	<u>86.03</u>	81.35
+M4oE [8]	✓	18.43M	85.72	72.73	84.71	81.05
+V-MoE [18]	✓	17.00M	84.70	<u>76.26</u>	84.98	<u>81.98</u>
+SA-MoE	✓	17.00M	**88.24**	**77.99**	**86.79**	**84.34**

MoE method with two popular MoE approaches: V-MoE [18] and M4oE [8]. All methods use CT ROI + LN mask as input to the model for fair comparison.

Evaluation Metrics: Five metrics are calculated, including the area under the receiver operating characteristic curve (AUC, or AUROC), specificity at 75%, 80% recall (S@R75, S@R80) and recall at 75%, 80% specificity (R@S75, R@S80).

Results of LN-station Classification: Table 1 outlines the quantitative comparisons of our method with other network backbones and the state-of-the-art MoE approaches. Several conclusions can be drawn. First, larger model capacities may not yield the performance gain. The lightweight MobileNetv2 and MobileViTv2 achieve the overall top performance. Second, the proposed LN prior-guided attention loss can further improve the classification accuracy to 85.96% AUC (with an increase of 1.2%, row 9 vs. row 6). This illustrates the usefulness of explicit LN prior guiding (via the designed loss) in the network training. Third, incorporating the MoE module into LN-station classification noticeably improves the performance by at least 0.86% AUC (row 7, 8, 10), with very few parameter increases. This shows that the MoE mechanism can adaptively and effectively learn the distinct but related features between different LN-stations. Compared with other leading MoE methods (same backbone and same number of experts), our explicit station-aware MoE (SA-MoE) achieves the best performance (AUC: 87.51% vs. 85.62% of M4oE [8] and 86.06% of V-MoE [18]). Last, our proposed method (SA-MoE + LN prior attention loss) achieves the highest performance across all metrics compared to other models, exhibiting +2.26% AUC, +4.0% S@R80 and +4.52% R@S80 improvements when compared to the second best performing method. Several qualitative examples are shown in Fig. 4.

Ablation results of SA-MoE: (1) *Classification performance on LN-Station subgroups (Table. 2).* Training three separate networks [16] on each LN-station subgroup produces the lowest performance (row 1); and training a single network in all LN stations (mixed together, row 2) leads to better results (from 79.43% to 80.47% in mean AUC). Although a multi-branch approach [31] achieves good performance with 81.35% mean AUC, a significant increase in model parameters

Table 3: Effect of the LN-prior guided attention loss on different backbones.

Methods	AUC	S@R80	R@S80
resnet18 [6]	83.92	72.97	72.10
resnet18 [6]+ATTN	**86.19**	**76.18**	**75.57**
ViT-T [3]	82.12	69.87	67.42
ViT-T [3]+ATTN	**83.81**	**73.27**	**73.00**

Table 4: Effect of the number of experts in SA-MoE.

Experts	AUC	S@R80	R@S80
1	84.76	73.97	73.76
2	86.93	78.88	78.13
3	**87.51**	**79.88**	**79.39**
6	87.33	79.58	79.03

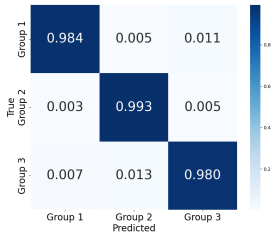

Fig. 3: Confusion matrix of route module in SA-MoE.

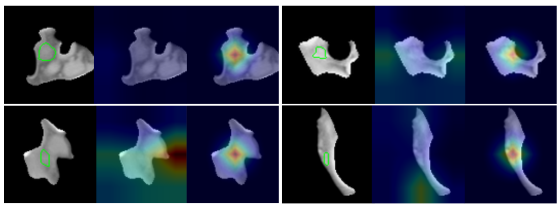

Fig. 4: Four qualitative examples of the attention map trained with (right) and without (middle) LN prior-guided attention loss. Green color indicates the segmented LN instances.

is required. In comparison, V-MoE [18] achieves better performance (81.98% mean AUC) with fewer parameters. Our proposed SA-MoE achieves the best performance in all three LN-station subgroups, surpassing the V-MoE [18] (the second best) by 2.36% in the mean AUC. (2) *Classification performance of the routing network in SA-MoE.* From the confusion matrix in Fig. 3, it is observed that the route module in SA-MoE accurately classifies the LN-stations into their respective groups, achieving >98% accuracy in all three groups. (3) *Effect of number of experts in SA-MoE.* We perform experiments using 2, 3 and 6 experts in SA-MoE and the results are shown in Table 4. The different numbers of experts all lead to significant performance improvements, with +2.17% in AUC, +4.91% in S@R80 and +4.37% in R@S80. Among them, three experts setting performs the best. We hypothesize the reason is that it balances the general LN-station features with station-specific features.

Ablation results of LN prior attention loss: We deploy the proposed LN prior-guided attention loss in both CNN (Resnet18 [6]) and transformer (ViT [3]) architectures. The results are shown in Table 3. For Resnet18, we use intermediate feature maps and apply softmax to serve as the basis for loss calculation with the LN mask. For ViT, we reshape the row (tied to class-tokens from the attention map) to the shape of raw feature map and then apply softmax for the loss calculation. Our attention loss consistently increases the classification performance in both CNN (+2.27% AUC) and transformer (+1.69% AUC) models.

4 Conclusion

This work addresses the important problem of classifying LN-station metastasis status in esophageal cancer by taking advantage of readily available LN station labels from large-scale pathology reports. A new LN prior-guided attention loss is proposed to explicitly regularize the network to focus on suspicious LN regions; a station-aware mixture-of-experts module is presented to effectively learn metastasis features of different LN-stations. Extensive 5-fold cross-validation conducted in 1,153 EC patients demonstrates the superiority of the proposed method by improving >2.26% AUC over other leading methods on this task.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgement This work was supported by Alibaba Group through Alibaba Research Intern Program, and Zhejiang Provincial Natural Science Foundation of China under Grant No. 2024-KYI-00I-I05. Li Zhang was partly supported by NSFC 81801778, NSFC 12090022 and Clinical Medicine Plus X-Young Scholars Project of Peking University PKU2023LCXQ041. Bin Dong was partly supported by NSFC 12090022, and the New Cornerstone Investigator Program.

References

1. Bayanati, H., E Thornhill, R., Souza, C.A., et al.: Quantitative ct texture and shape analysis: can it differentiate benign and malignant mediastinal lymph nodes in patients with primary lung cancer? European radiology **25**, 480–487 (2015)
2. Boiselle, P.M., Patz Jr, E.F., Vining, D.J., Weissleder, R., Shepard, J., McLoud, T.C.: Imaging of mediastinal lymph nodes: Ct, mr, and fdg pet. Radiographics **18**(5), 1061–1069 (1998)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Foley, K., Christian, A., Fielding, P., Lewis, W., Roberts, S.: Accuracy of contemporary oesophageal cancer lymph node staging with radiological-pathological correlation. Clinical radiology **72**(8), 693–e1 (2017)
5. Guo, D., Ye, X., Ge, J., Di, X., Lu, L., et al.: Deepstationing: thoracic lymph node station parsing in ct scans using anatomical context encoding and key organ auto-search. In: MICCAI. pp. 3–12. Springer (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
7. Isono, K., Sato, H., Nakayama, K.: Results of a nationwide study on the three-field lymph node dissection of esophageal cancer. Oncology **48**(5), 411–420 (1991)
8. Jiang, Y., Shen, Y.: M4oe: A foundation model for medical multimodal image segmentation with mixture of experts. In: MICCAI. pp. 621–631. Springer (2024)
9. Kann, B.H., Aneja, S., Loganadane, G.V., et al.: Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks. Scientific reports **8**(1), 14036 (2018)
10. Kann, B.H., Hicks, D.F., Payabvash, S., et al.: Multi-institutional validation of deep learning for pretreatment identification of extranodal extension in head and neck squamous cell carcinoma. Journal of Clinical Oncology **38**(12), 1304–1311 (2020)

11. Lee, J.H., Ha, E.J., Kim, J.H.: Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with ct. European radiology **29**, 5452–5457 (2019)
12. Li, B., Li, B., Jiang, H., Yang, Y., et al.: The value of enhanced ct scanning for predicting lymph node metastasis along the right recurrent laryngeal nerve in esophageal squamous cell carcinoma. Annals of translational medicine **8**(24) (2020)
13. Lim, K., Small Jr, W., Portelance, L., Creutzberg, C., Jürgenliemk-Schulz, I.M., Mundt, A., Mell, L.K., Mayr, N., Viswanathan, A., Jhingran, A., et al.: Consensus guidelines for delineation of clinical target volume for intensity-modulated pelvic radiotherapy for the definitive treatment of cervix cancer. International Journal of Radiation Oncology* Biology* Physics **79**(2), 348–355 (2011)
14. Liu, J., Wang, Z., Shao, H., Qu, D., Liu, J., Yao, L.: Improving ct detection sensitivity for nodal metastases in oesophageal cancer with combination of smaller size and lymph node axial ratio. European Radiology **28**, 188–195 (2018)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
16. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. Transactions on Machine Learning Research (2023)
17. Rice, T.W., Ishwaran, H., Ferguson, M.K., Blackstone, E.H., Goldstraw, P.: Cancer of the esophagus and esophagogastric junction: an eighth edition staging primer. Journal of Thoracic Oncology **12**(1), 36–42 (2017)
18. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. NIPS **34**, 8583–8595 (2021)
19. 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. TMI **35**(5), 1170–1181 (2015)
20. 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)
21. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR. pp. 4510–4520 (2018)
22. Schreuder, A., Jacobs, C., Scholten, E.T., van Ginneken, B., Schaefer-Prokop, C.M., Prokop, M.: Typical ct features of intrapulmonary lymph nodes: a review. Radiology: Cardiothoracic Imaging **2**(4), e190159 (2020)
23. 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)
24. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19**, 221–248 (2017)
25. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer statistics, 2022. CA: a cancer journal for clinicians **72**(1), 7–33 (2022)
26. Wang, S., Zhu, Y., Lee, S., Elton, D.C., Shen, T.C., Tang, Y., Peng, Y., Lu, Z., Summers, R.M.: Global-local attention network with multi-task uncertainty loss for abnormal lymph node detection in mr images. Medical Image Analysis **77**, 102345 (2022)
27. Wu, L., Yang, X., Cao, W., Zhao, K., Li, W., Ye, W., Chen, X., Zhou, Z., Liu, Z., Liang, C.: Multiple level ct radiomics features preoperatively predict lymph

node metastasis in esophageal cancer: a multicentre retrospective study. Frontiers in oncology **9**, 1548 (2020)
28. Yu, Q., Wang, Y., Yan, K., Li, H., Guo, D., Zhang, L., Shen, N., Wang, Q., Ding, X., Lu, L., et al.: Effective lymph nodes detection in ct scans using location debiased query selection and contrastive query representation in transformer. In: ECCV. pp. 180–198. Springer (2025)
29. Yu, Q., Wang, Y., Yan, K., Lu, L., Shen, N., Ye, X., Ding, X., Jin, D.: Slice-consistent lymph nodes detection transformer in ct scans via cross-slice query contrastive learning. In: MICCAI. pp. 616–626. Springer (2024)
30. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE **109**(5), 820–838 (2021)
31. 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: MICCAI. pp. 753–762. Springer (2020)