Semi-supervised Lymph Node Metastasis Classification with Pathology-guided Label Sharpening and Two-streamed Multi-scale Fusion

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

[1]DAMO Academy, Alibaba Group. [2]Peking University, Beijing, China.
[3]Sichuan Cancer Hospital, Cheng Du, China. [4]Hupan Lab, 310023, Hangzhou, China.
[5]Shanghai Jiao Tong University, Shanghai, China.
[6]The First Affiliated Hospital, Zhejiang University, Hangzhou, China.
zhangli_pku@pku.edu.cn, littlecancer@163.com
{yirui.wang, dakai.jin}@alibaba-inc.com

Abstract. Diagnosis of lymph node (LN) metastasis in computed tomography (CT) scans is an essential yet challenging task for esophageal cancer staging and treatment planning. Deep learning methods can potentially address this issue by learning from large-scale, accurately labeled data. However, even for highly experienced physicians, only a portion of LN metastases can be accurately determined in CT. Previous work conducted supervised training with a relatively small number of annotated LNs and achieved limited performance. In our work, we leverage the teacher-student semi-supervised paradigm and explore the potential of using a large amount of unlabeled LNs in performance improvement. For unlabeled LNs, pathology reports can indicate the presence of LN metastases within the lymph node station (LN-station). Hence, we propose a pathology-guided label sharpening loss by combining the metastasis status of LN-station from pathology reports with predictions of the teacher model. This combination assigns pseudo labels for LNs with high confidence and then the student model is updated for better performance. Besides, to improve the initial performance of the teacher model, we propose a two-stream multi-scale feature fusion deep network that effectively fuses the local and global LN characteristics to learn from labeled LNs. Extensive four-fold cross-validation is conducted on a patient cohort of 1052 esophageal cancer patients with corresponding pathology reports and 9961 LNs (3635 labeled and 6326 unlabeled). The results demonstrate that our proposed method markedly outperforms previous state-of-the-art methods by 2.95% (from 90.23% to 93.18%) in terms of the area under the receiver operating characteristic curve (AUROC) metric on this challenging task.

Keywords: Lymph node metastasis · Two stream multi-scale network · Semi-supervised Learning · Pathology-guided label refinement.

[*] Work was done during an internship at DAMO Academy, Alibaba Group.

1 Introduction

Esophageal cancer (EC) is the sixth leading cause of cancer death worldwide, accounting for 1 in 20 cancer deaths [23]. Lymph node (LN) metastasis is one of the most important prognostic factors in EC [1]. Accurate identification of preoperative LN metastasis is essential for making treatment decisions (surgery vs neoadjuvant) and determining treatment plans (surgical resection area and radiotherapy clinical target volume [CTV]) [8–10]. Therefore, LN metastasis assessment is of high clinical importance in EC diagnosis and treatment.

Assessing LN metastasis status in CT is a challenging task even for experienced physicians. Although size is an important indicator in distinguishing metastatic involvement, with larger nodes exhibiting a higher propensity for malignancy and smaller nodes displaying a greater likelihood of benign status, size as a sole predictive factor is not reliable. For instance, LN size demonstrates a sensitivity ranging from 60%-80% in identifying metastatic LNs in lung cancer patients [17, 21]. Considering the subtle differences of texture and intensity between malignant and benign LNs, it is extremely difficult and sometimes infeasible for physicians to annotate LN metastasis.

With the remarkable success of deep learning in various medical imaging computer-aided diagnosis (CAD) tasks [22, 27], preliminary attempts have been made to use deep learning for LN abnormality diagnosis [3, 12, 15, 20, 28]. Roth *et al.* proposes a two-stage 2.5D convolutional neural network (CNN) to detect and classify enlarged LNs in mediastinal and abdominal CT scans [19, 20]. Lee *et al.* examines different CNNs capacity to diagnose cervical LN metastasis in CT using 202 thyroid cancer patients [15]. Kann *et al.* train a dual network, *i.e.*, one for size-invariant and one for size-preserving, to classify metastatic LNs and extranodal extension (ENE) using CT scans of 270 head and neck cancer patients [11]. These studies employ the supervised learning, which ideally requires a large number of labeled LN data. However, it is extremely difficult to acquire the LN metastasis annotation. Therefore, it is of great benefit to incorporate a large amount of unlabeled LN data to improve the performance.

Prior semi-supervised learning (SSL) methods such as the Π-model [14], Mean Teacher [24], and MixMatch [2] have investigated similar scenarios, where models are trained using both labeled and unlabeled data. Nevertheless, these general SSL methods assume that there is not any relative label information in the unlabeled set. In our setting, although LN instances are unlabeled, there exists weak label information from pathology reports, which describe whether there is LN metastasis in the surgery-resected LN-stations. If the pathology report identifies a malignant LN-station, it indicates that there is malignant LNs inside this station (note there are often multiple LNs in a LN-station). Therefore, it is possible to utilize the pathology priors to develop a more effective SSL method for LN metastasis classification.

In this work, we propose an effective SSL method to better handle unlabeled LNs by using the priors from pathology reports. Specifically, we adopt the teacher-student mechanism, where a teacher model is utilized to get prediction scores, which then serve to guide the training of the student model. We propose a

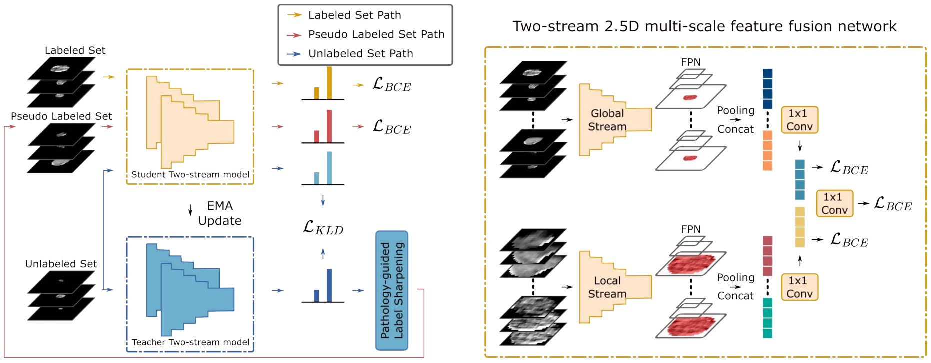

Fig. 1: Pipeline of the proposed framework. The left corresponds to the semi-supervised process, where the proposed pathology-guided label sharpening combines the teacher model predictions with information from pathology reports. This combination produced high-quality pseudo labels, which are used for the training of the student model subsequently. The right shows the details of network structure. For the input 3D CT patch, it is first transposed into multiple sets of 3-slice images with original and zoom-in size. Then, they are processed in parallel through a two-stream 2.5D multi-scale feature fusion network.

pathology-guided label sharpening loss to better integrate the LN-station metastasis information from pathology reports. For LN that the teacher model predicts with high confidence and the predicted label is consistent with its corresponding LN-station label confirmed in pathology report, we assign a pseudo label and introduce an additional classification loss to train the student model. Besides, to improve the initial performance of the teacher model, we first train a supervised model with labeled LNs. We propose a two-stream 2.5D multi-scale feature fusion deep network that effectively fuses the local and global LN characteristics to learn from labeled LNs. Using four-fold cross-validation, we evaluate our approach on CT scans of 1052 EC patients with surgery-pathology reports. Among them, the metastasis status of LNs in 310 patients is determined in their CT scans by radiologists by referring to pathology reports, while the rest of patients only have metastasis status of LN-stations from extracted from the pathology report. Experimental results show that our method significantly surpasses the state-of-the-art supervised and semi-supervised models and achieves the highest area under receiver operating characteristic curve (AUROC) score of 0.9318 on this challenging task.

2 Method

An overview of the proposed framework is illustrated in Fig. 1. A two-stream 2.5D multi-scale feature fusion network is proposed to simultaneously extract and merge global and local features of LNs. The training scheme consists of a supervised pre-training stage and a subsequent semi-supervised label mining and learning. Specifically, a supervised model is pre-trained using limited but accurately labeled LN data. Then, we employ a teacher-student SSL paradigm with the network's weights initialized by the supervised pre-trained model to iteratively mine large-scale but unlabeled LN data. We elaborate on the backbone design and the supervised training in Sec. 2.1 and describe the pathology prior-guided semi-supervised label mining in Sec. 2.2.

2.1 Two-stream 2.5D multi-scale feature fusion network

In clinical practice, metastasis LNs are identified in CT by both global characteristics (*e.g.*, size, shape) and localized features (*e.g.*, intensity inhomogeneity, textures, etc). With this observation, Kann *et al.* [11] proposes a 3D dual network architecture to simultaneously extract global and local features via a dimension-preserving network and a size-invariant network, respectively. Inspired but different from [11], we enhance the LN local-global metastasis identification in three folds. (1) 3D networks without proper pre-training weights are prone to overfitting. We propose a two-stream (local and global stream) 2.5D backbone to inherit the ability of joint feature extraction while allowing the leverage of large-scale pre-trained weights in 2D natural image dataset, e.g., ImageNet [4]. We group the CT slices into multiple sets of 3-channel images to adapt the 3D inputs for the 2.5D configuration. The adapted inputs are fed into the network to extract 2D features independently. We then concatenate and fuse these 2D features through a 1x1 convolution, producing the 3D context-aware fused feature. (2) We adopt the MobileNetv3 as our backbone and further add a Feature Pyramid Network (FPN) [16] to the last three feature layers to alleviate potential information loss during hierarchical downsampling. The features generated by different feature pyramids are then pooled and concatenated to produce a multi-scale representation. (3) Besides the supervision at the final fused head, we add side-supervision [6, 13] before fusing the two-stream outputs and conduct joint optimization to facilitate each stream learning more accurate features. Formally, this training objective can be formulated as:

$$\mathcal{L}_{\text{SUP}} = \mathcal{L}_{\text{BCE}}(f_l(X_l), Y) + \mathcal{L}_{\text{BCE}}(f_g(X_g), Y) + \mathcal{L}_{\text{BCE}}(f_{fuse}(X_{fuse}), Y) \tag{1}$$

where X_l, X_g, X_{fuse} denote the feature extracted from local stream, global stream, and fused output, respectively. f_l, f_g, f_{fuse} denote the corresponding classifiers, Y denotes the GT label and $\mathcal{L}_{\text{BCE}}$ denote the BCE loss. This is the supervised pre-training stage, which provides the initial model weight for the subsequent teacher-student SSL.

2.2 Prior-guided label sharpening for semi-supervised training

To effectively leverage unlabeled LNs, we adopt a teacher-student mechanism where the prediction scores produced by the teacher model on unlabeled LNs are then used to guide the student model training. The teacher model and student model share the same network described in Sec. 2.1, and are initialized with pre-trained weights from the supervised procedure. Following [24], the student model is trained via backpropagation, while the teacher model is updated iteratively by employing the exponential moving average (EMA) of the student model weights during training. Denoting the weights of the teacher model and student model at training step t as $\theta_t^{'}$ and θ_t, then $\theta_t^{'}$ can be updated as:

$$\theta_t^{'} \leftarrow \alpha\theta_{t-1}^{'} + (1-\alpha)\theta_t \tag{2}$$

where α serves as a smoothing coefficient to regulate the rate of knowledge update. In all our experiments, α is set to 0.999 following [24].

In previous SSL methods [2,14,24], since no prior knowledge is given for the unlabeled data, the pseudo labels produced by the teacher model are directly used to finetune the student model. In contrast, we have significant prior knowledge that can be leveraged. From the pathology report, we can know whether there is LN metastasis in corresponding LN-station. In other words, a malignant LN-station has malignant LN, while a benign one means all LNs inside are benign. After the supervised pre-training, the teacher model can generate predictions with considerable accuracy. Therefore, if the teacher model confidently predicts LN within a malignant LN-station to be malignant, we can assign a pseudo label to it and use these LNs to train the student model in a supervised manner. Specifically, for each malignant LN-station, we use a malignant threshold β of 0.7 as determined in the ablation study. The LNs with prediction scores above this threshold will be labeled as malignant (label=1). For benign LN-station, we have a lower benign threshold γ of 0.3 to account for noise labels. If an LN's prediction score is below this, we label it as benign (label=0). The pseudo labels for unlabeled LNs are updated every epoch. Before every epoch, we randomly select a fixed number of LNs with assigned pseudo labels to supervise the student model through an additional BCE loss. Combined with standard supervised loss and unsupervised consistency loss in SSL, the overall loss function is as:

$$\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{BCE}}(P_S^l, Y^l) + w(t)\mathcal{L}_{\text{KL}}(P_T^u, P_S^u) + \omega\mathcal{L}_{\text{BCE}}(P_S^a, \hat{Y}^a) \tag{3}$$

where P_S^l, P_S^u, P_S^a denote student model's prediction probability of labeled, unlabeled, and assigned pseudo label LNs, while P_T^u is the probability of unlabeled LNs predicted by the teacher model. Y^l and $\hat{Y}^a$ are the GT for labeled LNs and pseudo label for assigned LNs , respectively. $\mathcal{L}_{\text{KL}}$ means KL divergence, ω is the pathology-guided label sharpening loss weight and $w(t)$ denotes weight ramp-up function following [24].

3 Experiments

3.1 Experimental Settings

Dataset: We collected a dataset of 1052 esophageal cancer patients who underwent esophagectomy treatment. Each patient has a preoperative contrast-enhanced CT scan and a detailed pathology report after surgery indicating the metastasis status of resected LN-stations. The median CT scan size is 512 ×

512×91 voxels with the median resolution of $0.795 \times 0.795 \times 5.0$mm. Combining the automatic LN detection results [25] with a radiologist's examination and editing, we obtain 9961 LN masks (candidates without metastasis status). These LN masks, along with the CT scan, are then cropped using a 64×64×16 ROI centered on each 3D LN. Out of the 9961 LNs, 3635 LNs of 310 patients have labels of metastasis status (188 positive and 3447 negative), which is confirmed by the consensus of two radiologists according to the pathology report. The rest of the 6326 LNs of 742 patients are unlabeled, yet these patients have pathology reports indicating if an LN-station is metastatic or not. All experiments use four-fold cross-validation with a 60%/15%/25% training, validation, and testing split (at the patient level), respectively.

Implementation details: For supervised pre-training, MobileNetv3 [7] is used as a backbone for each of the two streams. We apply a single 1x1 convolutional layer for the multi-scale fusion, 3D slices fusion, and two-stream fusion. SGD optimizer with a learning rate of 3.2e-4 and cosine annealing decay is adopted, and the network is trained by 300 epochs with a mini-batch size of 32. For semi-supervised training, The threshold β and γ for producing pseudo labels are set to 0.7 and 0.3, respectively, and the pathology-guided label sharpening loss weight ω is set to 0.6. Detailed ablation results of these parameters are summarized in Fig. 2. SGD optimizer with a learning rate of 1.28e-4 and cosine annealing decay is adopted. We use a mini-batch of 128, with 16 labeled and 112 unlabeled samples. The network is trained by 300 epochs for convergence.

Comparison methods: We compare with other methods in the following two categories. 1) *Supervised category:* We evaluate widely-used CNN and Transformer classification networks: ResNet18 [5], MobileNetv3 [7] and MobileViTv2 [18] in 2.5D and 3D architecture. We also conduct the comparison with the 3D DualNet method of [11]. 2) *Semi-supervised category:* We compare to four popular SSL methods, Π-Model [14], Temporal Ensemble [14], Mean Teacher [24] and SimMatch [26]. For a fair comparison, MobileNetv3_2.5D [7] is used as the backbone in other SSL methods.

Evaluation metrics: To evaluate the performance comprehensively, we compute the area under the receiver operating characteristic curve (AUROC), specificity at a recall rate of 75% (S@R75), recall at a specificity rate of 75% (R@S75), accuracy at a recall rate of 75% (A@R75), and accuracy at a specificity rate of 75% (A@S75).

3.2 Comparison with Baseline Methods

Table 1 outlines the quantitative results of all compared methods and the proposed method. Regarding the supervised results, several observations can be drawn. (1) 2.5D methods outperform 3D methods with an improvement of 2-3% in AUROC. This shows that the 2.5D strategy equipped with the pre-trained 2D model weights results in markedly improved performance as compared to direct 3D classification. This may be due to the fact that LNs normally contain few slices in the z-axis, which is not sufficient to train the 3D convolutional kernels. (2) The 3D DualNet yields significantly lower performance as compared to

Table 1: Quantitative LN metastasis classification performance. Three groups correspond to 3D supervised models, 2.5D supervised models, and semi-supervised methods, respectively.

Method	AUROC	S@R75	R@S75	A@R75	A@S75
ResNet18_3D [5]	85.99	82.51	80.85	82.13	75.47
MobileViTv2_3D [18]	84.62	81.16	82.45	80.86	76.57
MobileNetv3_3D [7]	84.49	79.90	78.72	79.71	74.56
3D DualNet *et al.* [11]	86.48	82.96	81.91	82.54	74.89
ResNet18_2.5D [5]	87.94	84.28	84.04	83.87	74.94
MobileViTv2_2.5D [18]	87.34	85.14	83.51	84.61	76.60
MobileNetv3_2.5D [7]	87.88	84.26	83.51	83.78	75.41
Ours(Supervised)	90.68 (+2.74%)	88.26 (+3.12%)	88.30 (+4.26%)	87.58 (+2.97%)	77.04 (+0.44%)
Π-Model [14]	89.00	84.41	84.04	83.92	76.87
Temporal Ensemble [14]	89.31	83.33	82.45	82.93	74.94
Mean Teacher [24]	89.77	85.86	82.45	85.30	74.34
SimMatch [26]	90.23	86.56	87.77	86.01	75.91
Ours(Semi-supervised)	**93.18** (+2.95%)	**90.17** (+3.61%)	**90.96** (+3.19%)	**89.37** (+3.36%)	**80.45** (+3.58%)

our supervised model, although DualNet also considers both context and size information in LN classification. (3) Our supervised model improves the AUROC, S@R75, and R@S75 to 90.68%, 88.26%, and 88.30%, respectively, which are the highest among all supervised models and significantly ($p=0.006<0.01$) outperforms the second-best supervised model (ResNet18_2.5D [5]) by 2.74% in AUROC. This demonstrates the effectiveness of our two-stream multi-scale feature fusion design.

When further incorporating semi-supervised training, we can observe that all four comparing SSL methods improve the classification accuracy based on the supervised baseline MobileNetv3_2.5D. Among them, SimMatch [26] achieves the highest performance of 90.23% AUROC, 86.56%S@R75 and 87.77%R@S75. These results show that incorporating the unlabeled LNs in SSL is generally effective. Our final SSL model further significantly ($p=0.007<0.01$) boosts the AUROC to 93.18% with 2.95% improvement over SimMatch [26]. Other evaluation metrics show similar improvements.

3.3 Ablation Results

The effectiveness of each component in our method is demonstrated in ablation Table 2. First, the 2.5D feature fusion significantly increases the performance by 3.39% in AUROC. Based on the 2.5D setup, two-stream and multi-scale fusion can increase the AUROC by 1.74% and 1.49%, respectively. This indicates that both size and context information are helpful for LN metastasis classification, and aggregated multi-scale features also provide supportive information.

Table 2: Ablation studies of the effectiveness of the proposed 2.5D feature fusion, two-stream network (TS), multi-scale fusion (MS), and pathology-guided label sharpening loss (PGLS). MT refers to Mean Teacher mechanism [24].

2.5D	TS	MS	MT	PGLS	AUROC	S@R75	R@S75	A@R75	A@S75
					84.49	79.90	78.72	79.71	74.56
✓					87.88	84.26	83.51	83.78	75.41
✓	✓				89.61	87.36	87.77	86.76	75.91
✓		✓			89.37	83.96	85.64	83.56	75.03
✓	✓	✓			90.68	88.26	88.30	87.58	77.04
✓	✓	✓	✓		91.92	89.86	89.89	89.15	77.37
✓	✓	✓	✓	✓	**93.18**	**90.17**	**90.96**	**89.37**	**80.45**

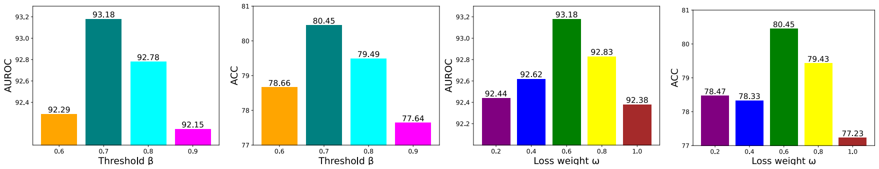

Fig. 2: Ablation study of label sharpening threshold β and loss weight ω

Then, combining these two modules, our fully supervised model achieves 90.68% in AUROC, boosting the performance by 2.80%. Finally, when conducting the original mean teacher SSL [24], the AUROC and other metrics further increase by ∼1%. In comparison, our proposed pathology-guided label sharpening loss further boosts the AUROC to 93.18%, outperforming the mean teacher SSL by 1.26% and demonstrating its effectiveness.

We further conduct an ablation study to investigate the impact of the malignant label threshold β and the pathology-guided label sharpening loss weight ω and results are shown in Fig .2. From left to right are AUROC and ACC (accuracy at 75% specificity) results across different thresholds and loss weights. Regarding the malignant label threshold β, it is observed that β=0.7 and 0.8 exhibit higher performance. When β possesses a low value of 0.6, more noisy labels may be included to reduce the performance. In contrast, when β is as high as 0.9, only a small portion of predictions become pseudo labels, which also hinders the performance. We then fix the β=0.7 and evaluate ω ranging from 0.2 to 1.0. We can see that ω=0.6 achieves the best performance. High ω might amplify the adverse impact of false assigned LNs, while low ω have limited influence on the model.

4 Conclusion

In this work, we introduce a specifically designed SSL method under limited labeled LNs and large-scale unlabeled LNs with pathology reports. We prove that the SSL paradigm is effective for challenges in labeling LNs, and the proposed pathology-guided label sharpening loss can further improve the performance with prior knowledge. Besides, for a better initial model for SSL, we introduce a two-stream 2.5D multi-scale feature fusion network. On a large-scale LN dataset, by combining supervised pre-training and semi-supervised training, our method achieves the top performance of 0.9318 AUROC, with about 3% improvement compared to previous methods.

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. Li Zhang was partly supported by NSFC 12090022 and 81801778, and Clinical Medicine Plus X-Young Scholars Project of Peking University PKU2023LCXQ041.

References

1. Ajani, J.A., D'Amico, T.A., Bentrem, D.J., Chao, J., Corvera, C., Das, P., Denlinger, C.S., Enzinger, P.C., Fanta, P., Farjah, F., et al.: Esophageal and esophagogastric junction cancers, version 2.2019, nccn clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network **17**(7), 855–883 (2019)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems **32** (2019)
3. 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 (2020)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Holste, G., van der Wal, D., Pinckaers, H., Yamashita, R., Mitani, A., Esteva, A.: Improved multimodal fusion for small datasets with auxiliary supervision. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
7. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
8. Jin, D., Guo, D., Ge, J., Ye, X., Lu, L.: Towards automated organs at risk and target volumes contouring: Defining precision radiation therapy in the modern era. Journal of the National Cancer Center **2**(4), 306–313 (2022)
9. Jin, D., Guo, D., Ho, T.Y., Harrison, A.P., Xiao, J., Tseng, C.k., Lu, L.: Deep esophageal clinical target volume delineation using encoded 3d spatial context

of tumors, lymph nodes, and organs at risk. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. pp. 603–612. Springer (2019)

10. Jin, D., Guo, D., Ho, T.Y., Harrison, A.P., Xiao, J., Tseng, C.K., Lu, L.: Deeptarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. Medical Image Analysis **68**, 101909 (2021)
11. Kann, B.H., Aneja, S., Loganadane, G.V., Kelly, J.R., Smith, S.M., Decker, R.H., Yu, J.B., Park, H.S., Yarbrough, W.G., Malhotra, A., 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)
12. Kann, B.H., Hicks, D.F., Payabvash, S., Mahajan, A., Du, J., Gupta, V., Park, H.S., Yu, J.B., Yarbrough, W.G., Burtness, B.A., 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)
13. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics **23**(2), 538–546 (2018)
14. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
15. 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)
16. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
17. McLoud, T., Bourgouin, P., Greenberg, R., Kosiuk, J., Templeton, P., Shepard, J.A., Moore, E., Wain, J., Mathisen, D., Grillo, H.: Bronchogenic carcinoma: analysis of staging in the mediastinum with ct by correlative lymph node mapping and sampling. Radiology **182**(2), 319–323 (1992)
18. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680 (2022)
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. IEEE transactions on medical imaging **35**(5), 1170–1181 (2015)
20. Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M.: A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part I 17. pp. 520–527. Springer (2014)
21. 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)
22. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19**, 221–248 (2017)
23. 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)

24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
25. Yan, K., Cai, J., Zheng, Y., Harrison, A.P., Jin, D., Tang, Y.B., Tang, Y.X., Huang, L., Xiao, J., Lu, L.: Learning from Multiple Datasets with Heterogeneous and Partial Labels for Universal Lesion Detection in CT. IEEE Trans. Med. Imaging **2020**, 1 (sep 2020)
26. Zheng, M., You, S., Huang, L., Wang, F., Qian, C., Xu, C.: Simmatch: Semi-supervised learning with similarity matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14471–14481 (2022)
27. 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)
28. Zhu, Z., Yan, K., Jin, D., Cai, J., Ho, T.Y., Harrison, A.P., Guo, D., Chao, C.H., Ye, X., Xiao, J., et al.: Detecting scatteredly-distributed, small, andcritically important objects in 3d oncologyimaging via decision stratification. arXiv preprint arXiv:2005.13705 (2020)