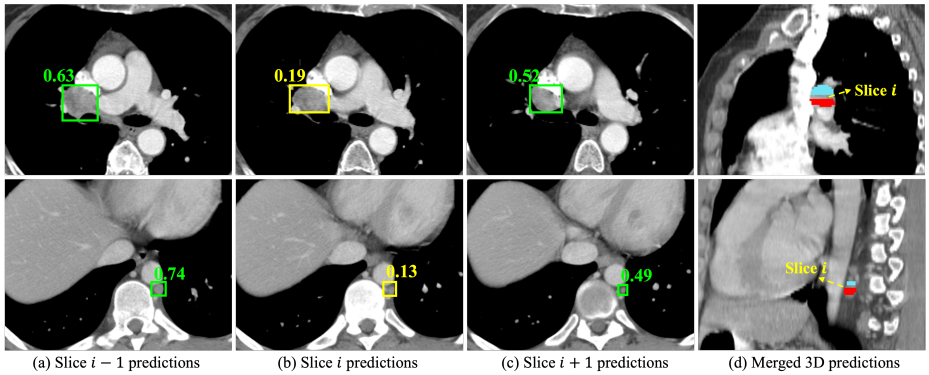
Slice-Consistent Lymph Nodes Detection Transformer in CT Scans via Cross-slice Query Contrastive Learning

Qinji Yu[1,2], Yirui Wang[2], Ke Yan[2,3], Le Lu[2], Na Shen[4], Xianghua Ye[5], Xiaowei Ding[1], and Dakai Jin[2]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] DAMO Academy, Alibaba Group
[3] Hupan Lab, 310023, Hangzhou, China
[4] Zhongshan Hospital Fudan University, Shanghai, China
[5] The First Affiliated Hospital Zhejiang University, Hangzhou, China
{yirui.wang, dakai.jin}@alibaba-inc.com
dingxiaowei@sjtu.edu.cn

Abstract. Lymph node (LN) assessment is an indispensable yet very challenging task in the daily clinical workload of radiology and oncology offering valuable insights for cancer staging and treatment planning. Finding scatteredly distributed, low-contrast clinically relevant LNs in 3D CT is difficult even for experienced physicians along with high inter-observer variations. Previous CNN-based lesion and LN detectors often take a 2.5D approach by using a 2D network architecture with multi-slice inputs, which utilizes the pretrained 2D model weights and shows better accuracy as compared to direct 3D detectors. However, slice-based 2.5D detectors fail to place explicit constraints on the inter-slice consistency, where a single 3D LN can be falsely predicted as two or more LN instances or multiple LNs are erroneously merged into one large LN. These will adversely affect the downstream LN metastasis diagnostic task as the 3D size information is one of the most important malignant indicators. In this work, we propose an effective and accurate 2.5D LN detection transformer that explicitly considers the inter-slice consistency within a LN. It first enhances a detection transformer by utilizing an efficient multi-scale 2.5D fusion scheme to leverage pre-trained 2D weights. Then, we introduce a novel cross-slice query contrastive learning module, which pulls the query embeddings of the same 3D LN instance closer and pushes the embeddings of adjacent similar anatomies (hard negatives) farther. Trained and tested on 3D CT scans of 670 patients (with 7252 labeled LN instances) of different body parts (neck, chest, and upper abdomen) and pathologies, our method significantly improves the performance of previous leading detection methods by at least 3% average recall at the same FP rates in both internal and external testing.

Keywords: Lymph node detection · Detection transformer · Slice-consistency · contrastive learning.



(a) Slice $i-1$ predictions (b) Slice i predictions (c) Slice $i+1$ predictions (d) Merged 3D predictions

Fig. 1: (a-c) 2D predictions with confidence of existing LN detection method [21] on consecutive slices, and yellow boxes will be filtered due to the low confidence. (d) Merged 3D predictions in the sagittal view. The inconsistent predictions across consecutive slices results in the entire lymph node instance being divided into two separate ones.

1 Introduction

Lymph node (LN) detection in computed tomography (CT) scans is a critical task in medical imaging, offering valuable insights for staging and treatment planning in various cancers [8,7]. However, the heterogeneity in size, shape, and density of LNs, along with their often ambiguous boundaries and low contrast against adjacent soft tissues (*e.g.*, vessels, muscles, and esophagus), poses significant challenges for both manual diagnosis and automated detection systems.

Due to its importance and difficulty, automatic LN detection and segmentation has been attracting increasing attentions [1,18,3,27,6,28,10]. Early work often relied on hand-crafted features and heuristic rules, which may not generalize well across the diverse manifestations of lymph nodes in medical imaging [1,9]. The advent of deep learning, particularly the application of convolutional neural networks (CNNs), has led to a paradigm shift in medical image analysis, as well as the LN detection [18,3,2,10,22]. Wang et al. [20] improved 2D Mask R-CNN [11] by proposing a global-local attention module and a multi-task uncertainty loss to detect LN in abdomen MR images. Yan et al. [22] adopted 2.5D backbone to extract 3D context information from multi-slice-input, and reported superior LN detection performance when compared to purely 3D detection methods [2]. While improved detection results are accomplished, these 2D or 2.5D CNN detectors have inherent limitations: (1) They rely on many hand-crafted components, *e.g.*, anchor matching strategy and non-maximum suppression (NMS) post-processing, which involves tuning a large number of hyper-parameters. (2) Although 2.5D approaches achieve the leading performance, simple multi-slice input or merging 2D slice features with a 3D convolution is not sufficient to chracterize the 3D continuity of LNs. This often leads to inconsistent predic-

tions between consecutive slices of the same LN. This will adversely affect the downstream LN metastasis diagnostic task as the 3D size information is one of the most important indicators. For example, as shown in Fig. 1, slice $i-1$ to slice $i+1$ belong to the same 3D LN. However, slice i has low prediction confidence, which divide the single LN instance into two separate ones after the 3D box merging.

Recently, a new detection paradigm, *i.e.*, DEtection TRansformer (DETR) [5], was proposed for natural images by reformulating the detection task as a set prediction problem. In contrast to CNN-based detectors, it directly predicts a set of objects without relying on manual heuristics such as anchors and post-processing procedures. Among the recently developed DETR-based detectors [26,17,14,25,15], DINO [25] and Mask DINO [15] have achieved the leading performance by introducing a denoising technique. However, transformer-based detectors also suffer from the aforementioned slice prediction inconsistency issue.

To conquer above issues, built upon the latest DETR-based detectors (*e.g.*, DINO, Mask DINO), we propose a novel cross-slice query contrastive learning module for the slice-consistent LN detection in CT. The key idea is to pull the similarity of the same 3D LN across slices closer in the embedding space and push the embeddings of adjacent similar anatomies (hard negatives) farther. This provides more discriminative and consistent LN instance features, which generate more accurate 3D LN instances after box merging. Moreover, we have replaced their original 2D backbone with a 2.5D backbone, allowing them to extract 3D context information from multi-slice inputs while leveraging pre-trained 2D weights. To provide a more comprehensive evaluation of our LN detection method, we collect and curate a large scale of LN CT scans containing 670 patients (with 7252 labeled LN instances of both enlarged and smaller sizes) from 5 institutional datasets across different body parts and diseases. Among them, three datasets are used as internal data to develop and internally test the LN detection performance while the rest two are used for independent external testing. In both internal and external testing, the proposed cross-slice contrastive learning module and 2.5D feature fusion bring significant improvement for DETR-based detectors and surpass the well-tuned CNN-based detectors by a large margin.

2 Methods

In this section, we first provide an overview of our LN detection framework in Sec. 2.1. Then, the details of the proposed cross-slice query contrastive learning module are introduced in Sec. 2.2.

2.1 LN Detection Framework

Fig. 2 depicts the proposed framework, which consists of a 2.5D CNN backbone, and a detection transformer (*i.e.*, transformer encoder and decoder) with multiple prediction heads (*i.e.*, mask head, box head, and class head). We elaborate on each component below.

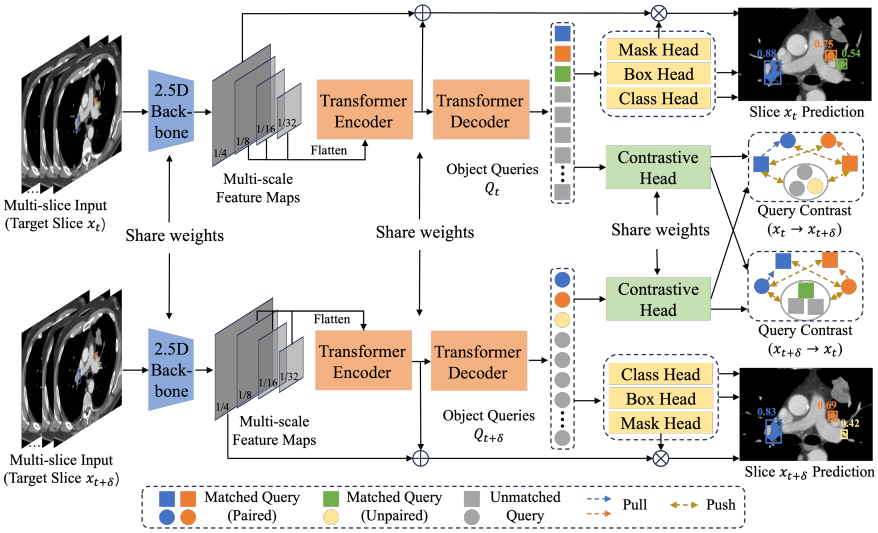


Fig. 2: The training framework of our methods. Given a target slice x_t and its neighbor slice $x_{t+\delta}$, the shared-weight 2.5D backbone and transformer generate the object queries on them, respectively. The queries are used to predict masks, boxes, and classes and do cross-slice contrastive learning. Queries of the same color belong to the same 3D LN instance. The denoising branch, position embeddings, and intermediate queries are omitted in the figure for clarification.

2.5D Backbone. As noted in [22], 3D context information is important for distinguishing LNs from other tube-shaped organs such as vessels and esophagus. However, directly applying 3D CNNs is memory-consuming and lacks pre-trained weights. To bridge the gap, we adopt the 2.5D feature fusion layer in [23] to leverage the rich 3D context information across slices. Specifically, we extract four upper and four lower slices from the CT scan to serve as the 3D context of the central target slice (see details in the supplementary). A feature map is then extracted for each slice with a shared 2D CNN, *e.g.*, ResNet50 [12] with ImageNet pre-trained weights in our methods. Finally, the 2.5D fusion layer is inserted after each res-block to produce the multi-scale 3D-context-enhanced feature maps for the target slice (*i.e.*, multi-scale 2.5D fusion scheme).

Detection Transformer. The overall architecture of our detection transformer is based on a unified DETR-like object detection and segmentation framework, Mask DINO [15]. It first takes the multi-scale feature maps from the 2.5D backbone with corresponding positional embeddings as input. After feature enhancement with the transformer encoder, top-scoring encoder features (used as region proposals) are selected to initialize the object queries $Q \in \mathbb{R}^{N \times C}$ for the transformer decoder. These queries will be further updated by a set of cross- and self-attention layers in the decoder, and then fed into the multiple prediction heads to produce the final N predictions for the input slice, including masks,

boxes, and classification results. Furthermore, an additional denoising branch is added to accelerate training convergence.

Then we calculate the pair-wise matching cost [15] between N predictions and total ground truth. For each ground truth, its optimal matched prediction can be assigned efficiently with the Hungarian algorithm [5]. Finally, the whole model is optimized with a combination loss,

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{\text{box}} + \lambda_3 \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{contra}} \tag{1}$$

where loss weights $\lambda_1, \lambda_2,$ and λ_3 are set to 1.0, 2.0 and 5.0 as [15]. In detail, $\mathcal{L}_{\text{cls}}$ is the focal loss. $\mathcal{L}_{\text{box}}$ is the combination of L1 loss and GIoU loss [19], while $\mathcal{L}_{\text{mask}}$ adopts cross-entropy and dice loss. $\mathcal{L}_{\text{contra}}$ is the cross-slice query contrast loss which will be described in the next section.

2.2 Cross-slice Query Contrastive Learning

More discriminative query embeddings can help distinguish LN instances on different slices, thereby improving the prediction consistency of the same instance across slices. Therefore, we introduce contrastive learning between slices to pull the query embedding of the same 3D LN instance closer in the embedding space and push the embedding of different 3D LN instances and similar adjacent anatomies far away.

Given a target slice x_t and its neighbor slice $x_{t+\delta}$, where δ is a random slice interval sampled from $[1, T]$, the corresponding object queries are $Q_t = \{q_t^1, q_t^2, \ldots, q_t^N\}$ and $Q_{t+\delta} = \{q_{t+\delta}^1, q_{t+\delta}^2, \ldots, q_{t+\delta}^N\}$, respectively. A simple two-layer MLP followed by an L2 normalization layer, named contrastive head ϕ, is then adopted to project queries into the embedding space.

For the m-th LN instance on the target slice x_t, we assume the previous Hungarian matching has assigned the i_m-th object query in Q_t as its matched query. If the same instance also appears on the slice $x_{t+\delta}$, we denote the j_m-th object query in $Q_{t+\delta}$ as the matched query. Thus, two indexes i_m and j_m will give us a pair of object queries $q_t^{i_m}$ and $q_{t+\delta}^{j_m}$ that are consistent in their instance identity. As shown in Fig. 2, the LN instances (boxes marked in blue and orange) appearing on the slice x_t may have different positions and appearances on the neighbor slice $x_{t+\delta}$, but their object queries should be as close as possible in embedding space. To achieve this, we take the query $q_{t+\delta}^{j_m}$ as positive sample for query $q_t^{i_m}$, and the top K queries in $Q_{t+\delta} \setminus q_{t+\delta}^{j_m}$ with the highest classification scores as negative samples. Then the contrastive loss for slice x_t to $x_{t+\delta}$ can be defined as follows:

$$\mathcal{L}_{contra}^{t \to t+\delta} = -\sum_{m=1}^{M} \log \left[\frac{\exp\left(\phi(q_t^{i_m})\phi(q_{t+\delta}^{j_m})\right)}{\exp\left(\phi(q_t^{i_m})\phi(q_{t+\delta}^{j_m})\right) + \sum_{k=1}^{K} \exp\left(\phi(q_t^{i_m})\phi(q_{t+\delta}^{k})\right)} \right] \tag{2}$$

where M is the number of shared LN instances between slice x_t and $x_{t+\delta}$, and $\phi(q)$ means the query projection in the embedding space. Similarly, we can derive the contrastive loss for slice $x_{t+\delta}$ to x_t and the total contrastive loss $\mathcal{L}_{contra} = \mathcal{L}_{contra}^{t \to t+\delta} + \mathcal{L}_{contra}^{t+\delta \to t}$.

3 Experiments and Results

3.1 Experimental Setup

Datasets. In this work, we collected and curated CT scans from 5 LN datasets of different body parts (neck, chest, and upper abdomen) and various diseases containing a total of 670 patients and 7252 instance-level LN annotations. Among them, NIH-LN [4] is a public LN dataset, while the rest are from four clinical centers (denoted as Center1-4 for simplification). Specifically, NIH-LN comprises of 89 lung cancer patients. Center1 and Center4 include 256 and 50 head & neck cancer patients, respectively. Center2 provides 91 esophageal cancer patients. Center3 contains 184 patients with different types of diseases (lung cancer, esophageal cancer and infectious lung disease). More descriptions including LN regions and characteristics of each dataset can been seen in the supplementary. We use NIH-LN and datasets of Center 1-2 to develop and **internally test** the LN detection performance (70% training, 10% validation, and 20% testing). Datasets of Center3-4 are used for independent **external testing**.

Implementation Details. The code will be available at Github URL. All CT scans are first resampled to a constant spacing of $0.8 \times 0.8 \times 2.0$mm, and image intensity was then clipped to $[-200, 300]$ HU. In training, each mini-batch consists of 8 samples, *i.e.*, 4 pairs of neighbor slices. We use RAdam optimizer with the initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-4} for 30 epochs training. Besides, cosine annealing scheduler is adopted to reduce the learning rate to 1×10^{-5} with a warm-up step of 500 iterations. For the cross-slice query contrastive learning module, we set the total object query number N, negative sample number K and neighborhood range T to 300, 100, and 3, respectively. Detailed ablation results of these parameters are summarized in Table 2. Data augmentation includes random scaling, cropping, rotation, intensity scaling, and gamma augmentation. In inference, we select the op 20 ranked query predictions as the 2D detection results in each slice, and then merge the 2D detection boxes to 3D ones following [21,22]. All experiments are conducted with PyTorch 1.12 on a Tesla A100 GPU. The average time for training and inference is 0.6 GPU days and 5s per CT scan, respectively.

Evaluation Metrics. Following previous lesion detection works [1,23,24,21,22], we use the free-response receiver operating characteristic (FROC) curve as the evaluation metric and report the recall at 0.5, 1, 2, 4 FPs per CT scan. When comparing each predicted 3D box with the GT 3D boxes, the predicted box is counted as true positive if the 3D intersection over detected bounding-box ratio (IoBB) is larger than 0.3 [23,22]. As the short axis of metastatic LNs are almost all larger than 5mm [10], we only detect LNs with short axis equal to or larger than 5mm during inference. If a GT LN smaller than 5mm is detected, it is neither counted as a TP nor an FP. In training, we still use LN annotations of all sizes.

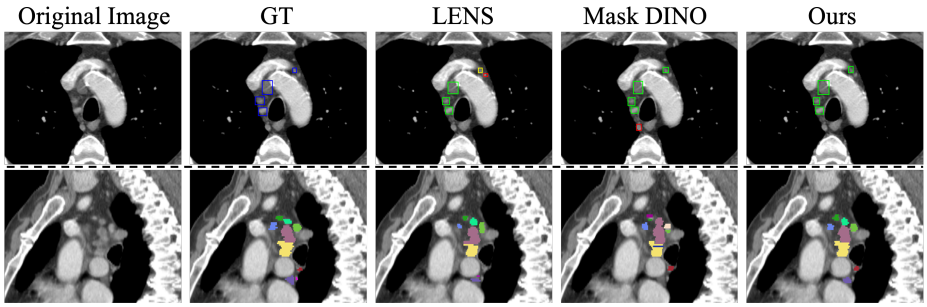


Fig. 3: Qualitative LN detection results. The first row shows slice-level prediction results. Blue, green, red, and yellow denote for GTs, TPs, FPs and FNs, respectively. The second row is the merged 3D predictions in the sagittal view.

3.2 LN Detection Results

Comparison to state-of-the-art methods. We conduct extensive comparison evaluation for LN detection, including the leading DETR-based general object detection methods (DINO [25] and Mask DINO [15]), medical lesion detection methods (MULAN [23], LENS [21], nnDetection [2], A3D [24] + SATr [16]), and the leading medical segmentation method nnUNet [13].

The quantitative evaluation results for internal and external testing are presented in Table 1. Several observations can be drawn. First, in internal testing, our 2.5D feature fusion and cross-slice contrastive learning both significantly improve the detection performance for DINO and Mask DINO. For example, the 2.5D fusion increase the Mask DINO average recall from 48.59% to 51.43%, where the contrastive learning further improves the average recall by 3.51%. Second, when compared to the lesion detection methods, our Mask DINO† achieves substantial improvement over nnDetection, LENS, and MULAN by 19.43%, 6.87%, and 6.97%, respectively. It is worth noting that LN detection by segmentation generally yields inferior performance as compared to direct detection methods, *e.g.*, nnUNet only obtains 51.00% recall while its FPs is as high as 4.2. Finally, in external testing, our method generalizes well, where Mask DINO† increases the original Mask DINO's average recall by 5.98% and outperforms the second best lesion detection model MULAN by 3.15%.

We also show some qualitative comparisons with other leading detection methods in Fig. 3. It can be observed that our method exhibits higher sensitivity and reduces FPs. Furthermore, the merged 3D predictions illustrate that our cross-slice contrastive learning module promotes the prediction consistency between slices avoiding one 3D LN being falsely divided as two or multiple LNs being erroneously merged into one large LN.

Parameter analysis. The influence of parameters in our cross-slice contrastive learning method is summarized in Table 2. Regarding the negative sample num-

Table 1: Results for our method and other detection methods averaged on the internal hold-out test sets and external datasets. "*" means w/ 2.5D backbone. "†" means w/ 2.5D backbone and cross-slice query contrastive learning. Best in **bold**, second underline.

Model	Internal test					External test				
	Recall(%)@FPs ↑					Recall(%)@FPs ↑				
	@0.5	@1	@2	@4	**Avg.**	@0.5	@1	@2	@4	**Avg.**
nnDetection [2]	19.60	27.70	40.55	54.18	35.51	23.24	33.04	43.07	51.09	37.61
A3D [24] + SATr [16]	31.07	40.93	51.44	61.47	46.23	30.83	38.29	46.33	55.17	42.66
LENS [21]	31.04	42.97	53.72	64.54	48.07	29.39	39.52	52.39	60.84	45.53
MULAN [23]	29.51	42.21	55.14	65.01	47.97	32.20	42.01	51.34	60.33	46.47
DINO [25]	28.43	39.10	48.32	59.38	43.81	27.95	35.55	44.78	54.06	40.58
DINO*	30.99	40.88	54.30	65.02	47.80	29.47	35.95	45.80	57.12	42.08
DINO†	35.52	45.27	55.80	66.65	50.81	31.40	38.49	49.15	57.97	44.25
Mask DINO [15]	34.29	42.26	52.15	65.64	48.59	29.76	37.86	49.85	57.11	43.64
Mask DINO*	31.26	47.94	58.69	67.83	51.43	28.56	39.62	51.98	61.37	45.38
Mask DINO†(Ours)	**39.70**	**49.94**	**60.75**	**69.38**	**54.94**	**34.74**	**44.50**	**55.09**	**64.17**	**49.62**
nnUNet [13]	51.00@4.2 FPs (vs. **71.11**)					42.30@3.7 FPs (vs. **62.08**)				

ber K (Table 2(a)), we can see that different K numbers all bring noticeable gains (from 1.27% to 3.51%) over the 2.5D Mask DINO*, which further demonstrates the effectiveness of our proposed cross-slice contrastive learning method. Among them, selecting the top-50 and top-100 queries achieve the highest performing results. In comparison, using all queries yields the least improvement. We hypothesize that queries with low classification scores may represent easy negative samples that contribute little to contrastive learning.

Regarding the slice interval range T (Table 2(b)), we conclude that: (1) overall, the proposed cross-slice query contrastive learning is effective under a wide range of slice interval, *i.e.*, from $T = 1$ to $T = 4$, and achieves the best average recall when $T = 3$. (2) When the slice interval becomes large, *e.g.*, $T = 4$, the performance drops markedly. This may be because that appearance and context of two far away slices of the same LN instance change dramatically, especially for CT scans with coarse slice thickness, which hampers the efficiency of contrastive learning.

4 Conclusion

Lymph node (LN) assessment is an indispensable yet very challenging task in the daily clinical workload of radiology and oncology. In this work, we propose an effective 2.5D LN detection transformer that explicitly considers the inter-slice consistency within a LN. It first enhances a detection transformer by utilizing an efficient multi-scale 2.5D fusion scheme to leverage pre-trained 2D weights. Then, we introduce a novel contrastive learning between slices to pull the query embedding of the same 3D LN instance closer and push the embedding of different 3D LN instances and other similar adjacent anatomies far away. Trained and

Table 2: Effect of the negative sample number K and slice interval range T for cross-slice query contrastive learning module.

K	Recall(%)@FPs ↑				
	@0.5	@1	@2	@4	Avg.
20	35.51	49.29	59.92	68.08	53.20
50	37.59	**50.70**	60.49	**69.66**	54.61
100	**39.70**	49.94	**60.75**	69.38	**54.94**
All	36.96	46.27	59.67	67.88	52.70

(a) Negative sample number K

T	Recall(%)@FPs ↑				
	@0.5	@1	@2	@4	Avg.
1	34.98	50.37	**61.27**	68.84	53.87
2	34.66	**50.73**	60.83	**70.25**	54.12
3	**39.70**	49.94	60.75	69.38	**54.94**
4	36.14	49.36	58.74	68.40	53.16

(b) Interval range T

tested on 3D CT scans of 670 patients of different body parts and pathologies, our method significantly improves the performance of previous leading detection methods in both internal and external testing.

Disclosure of Interests. The authors declare no competing interests.

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. Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H.: nndetection: a self-configuring method for medical object detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI. pp. 530–539. Springer (2021)
3. 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**, 977–986 (2019)
4. 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. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**(1), 44–58 (2023)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
6. 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)
7. Detterbeck, F.C., Boffa, D.J., Kim, A.W., Tanoue, L.T.: The eighth edition lung cancer stage classification. Chest **151**(1), 193–203 (2017)

8. El-Sherief, A.H., Lau, C.T., Wu, C.C., Drake, R.L., Abbott, G.F., Rice, T.W.: International association for the study of lung cancer (iaslc) lymph node map: radiologic review with ct illustration. Radiographics **34**(6), 1680–1691 (2014)
9. 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)
10. Guo, D., Ge, J., Yan, K., Wang, P., Zhu, Z., Zheng, D., Hua, X.S., Lu, L., Ho, T.Y., Ye, X., et al.: Thoracic lymph node segmentation in ct imaging via lymph node station stratification and size encoding. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 55–65. Springer (2022)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
12. 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)
13. 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 **18**(2), 203–211 (2021)
14. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13619–13627 (2022)
15. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3041–3050 (2023)
16. Li, H., Chen, L., Han, H., Kevin Zhou, S.: Satr: Slice attention with transformer for universal lesion detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 163–174. Springer (2022)
17. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022)
18. Oda, H., Roth, H.R., Bhatia, K.K., Oda, M., Kitasaka, T., Iwano, S., Homma, H., Takabatake, H., Mori, M., Natori, H., et al.: Dense volumetric detection and segmentation of mediastinal lymph nodes in chest ct images. In: Medical Imaging 2018: Computer-Aided Diagnosis. vol. 10575, p. 1057502. SPIE (2018)
19. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
20. 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)
21. 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 (2021)
22. Yan, K., Jin, D., Guo, D., Xu, M., Shen, N., Hua, X.S., Ye, X., Lu, L.: Anatomy-aware lymph node detection in chest ct using implicit station stratification. arXiv preprint arXiv:2307.15271 (2023)

23. Yan, K., Tang, Y., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M.: Mulan: multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI. pp. 194–202. Springer (2019)
24. Yang, J., He, Y., Kuang, K., Lin, Z., Pfister, H., Ni, B.: Asymmetric 3d context fusion for universal lesion detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI. pp. 571–580. Springer (2021)
25. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
26. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
27. 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 (2020)
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)

Supplementary Material

Qinji Yu[1,2], Yirui Wang[2], Ke Yan[2,3], Le Lu[2], Na Shen[4], Xianghua Ye[5], Xiaowei Ding[1], and Dakai Jin[2]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] DAMO Academy, Alibaba Group
[3] Hupan Lab, 310023, Hangzhou, China
[4] Zhongshan Hospital Fudan University, Shanghai, China
[5] The First Affiliated Hospital Zhejiang University, Hangzhou, China
dingxiaowei@sjtu.edu.cn, dakai.jin@alibaba-inc.com

Table 1: Statistics of 5 LN detection datasets, where 3 of them are used as the internal data for the model development and internal testing and the rest 2 are used as independent external testing set. NIH-LN is a public dataset, and other datasets are in-house datasets collected from 4 different clinical centers. HN, Eso and Mul. represent head & neck cancer, esophageal cancer and multiple types of diseases, respectively.

Dataset	#Patient	#LN	Avg. Res. (mm)	Body Parts	Setting
NIH-LN	89	1956	$(0.82, 0.82, 2.0)$	chest & abdomen	internal
Center1-HN	256	1890	$(0.46, 0.46, 4.0)$	head & neck	
Center2-Eso	91	857	$(0.70, 0.70, 4.9)$	chest	
Center3-Mul	184	2131	$(0.76, 0.76, 2.0)$	chest & abdomen	external
Center4-HN	50	418	$(0.48, 0.48, 1.2)$	head & neck	
Total	670	7252	-	-	-

Table 2: Detailed performance of our method across ***different size*** lymph nodes (Recall@FPs=[0.5, 1, 2, 4]).

Ln size	Recall@FPs=[0.5, 1, 2, 4]).				
	@0.5	@1	@2	@4	Avg.
All size	26.00	33.86	44.88	54.16	39.73
$\geq$ 5mm	39.70	49.94	60.75	69.38	54.94
$\geq$ 7mm	54.27	65.18	74.08	79.71	68.31
$\geq$ 10mm	66.39	76.12	78.94	83.24	76.17

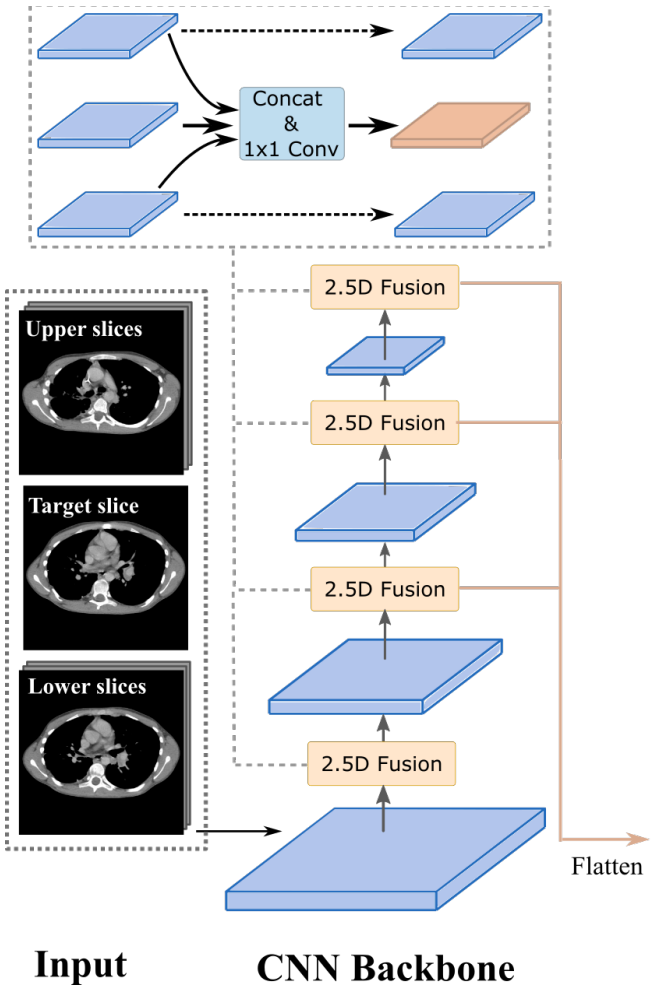


Fig. 1: Details of our 2.5D backbone and the multi-scale 2.5D feature fusion scheme. Specifically, to detect LN in a target CT slice, we extract four upper and four lower slices from the original CT scan to serve as the 3D context of the central target slice. Then, these total nine consecutive CT slices are grouped into three sets of 3-channel images. Each set is processed by the shared CNN backbone independently, and the three sets are then fused by concatenation and 1×1 conv. After that, the feature map of the original central target slice is replaced by the fused feature map, while the upper and lower sets' feature maps remain unchanged.