Holistic and Comprehensive Annotation of Clinically Significant Findings on Diverse CT Images: Learning from Radiology Reports and Label Ontology – Supplementary Material

Ke Yan[1], Yifan Peng[2], Veit Sandfort[1], Mohammadhadi Bagheri[1], Zhiyong Lu[2], Ronald M. Summers[1]
[1] Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Clinical Center
[2] National Center for Biotechnology Information, National Library of Medicine
[1,2] National Institutes of Health, Bethesda, MD 20892
{ke.yan, yifan.peng, veit.sandfort, mohammad.bagheri, zhiyong.lu, rms}@nih.gov

1. More Lesion Annotation Results

1.1. Examples

Fig. 1 shows more lesion annotation examples of LesaNet in various body parts. We found that:

- LesaNet is good at identifying fine-grained lymph nodes (subplots (c),(e),(g),(h)), which account for a major part of the DeepLesion dataset.
- In (d), LesaNet correctly recognized the coarse-scale body part (axilla), but it classified the lesion as a lymph node instead of a mass-like skin thickening (groundtruth). This is possibly because most axillary lesions in DeepLesion are lymph nodes, while axillary skin lesions are rare.

1.2. Quantitative Results

In order to observe the effect of the components in LesaNet more clearly, we randomly re-split the training and validation set in the patient level 10 times and rerun the ablation study. Mean and standard deviation accuracies are reported in Table 1. Similar conclusions can be drawn from the table compared to Sec. 5.5 of the main paper.

The batch size during training may affect results because of the triplet loss and RHEM strategies used in LesaNet. We tested various batch sizes from 16 to 200 with or without the two strategies. No significant correlation was observed between the settings of batch size and accuracy. Methods with triplet loss and RHEM were consistently better than those without them.

2. More Lesion Retrieval Examples

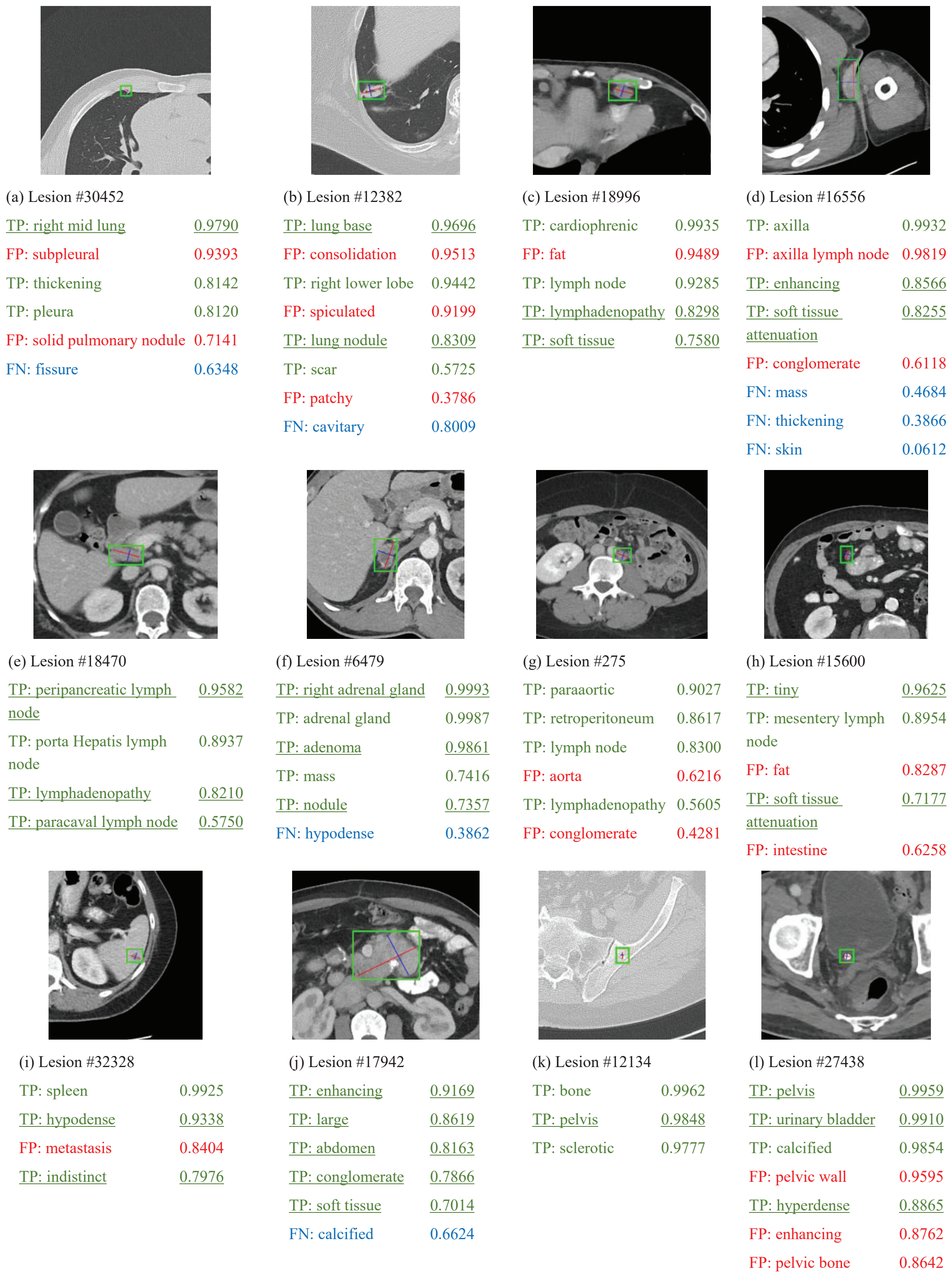
Fig. 2 demonstrates more lesion retrieval examples of LesaNet (please refer to Fig. 7 in the main paper). We constrain that the query and all retrieved lesions must come from different patients, so as to better exhibit the retrieval ability and avoid finding identical lesions of the same patient. For lesions that are common in DeepLesion, such as lung nodules and liver masses, it is easy for LesaNet to retrieve lesions that are very similar in both visual appearance and semantic labels, *e.g*. Fig. 2 (a) and (b). Moreover, LesaNet is also able to retrieve lesions that look different but share similar semantic labels, *e.g*. the rib/chest wall mass in subplot (c), the pancreatic tail mass in (d), and the left adrenal nodule in (e).

We have conducted another experiment to quantitatively compare the lesion retrieval accuracy of LesaNet and lesion embedding [1]. We used the lesions in the text-mined test set as queries to retrieve similar lesions from the training set, which has no patient-level overlap with the test set. The accuracy criterion is the average cumulative gain (ACG), which is defined as the average number of overlapping labels between the query and each of the top-K retrieved samples [2]. The ACG@top-5 of lesion embedding [1] is 2.25, meaning that a retrieved lesion shares an average of 2.25 common labels with the query lesion. The ACG@top-5 of LesaNet is 2.36. LesaNet learned from more fine-grained labels text-mined from radiology reports, which is the main reason of its improved accuracy, despite the fact that it uses a shorter embedding vector (256D vs. 1024D) and was not primarily trained for retrieval.

References

[1] Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam Harrison, Mohammadhadi Bagheri, and Ronald Summers. Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-scale Lesion Database. In *CVPR*, pages 9261–9270, 2018.

[2] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan. Deep semantic ranking based hashing for multi-label image retrieval. In *CVPR*, pages 1556–1564, 2015.



(a) Lesion #30452

TP: right mid lung	0.9790
FP: subpleural	0.9393
TP: thickening	0.8142
TP: pleura	0.8120
FP: solid pulmonary nodule	0.7141
FN: fissure	0.6348

(b) Lesion #12382

TP: lung base	0.9696
FP: consolidation	0.9513
TP: right lower lobe	0.9442
FP: spiculated	0.9199
TP: lung nodule	0.8309
TP: scar	0.5725
FP: patchy	0.3786
FN: cavitary	0.8009

(c) Lesion #18996

TP: cardiophrenic	0.9935
FP: fat	0.9489
TP: lymph node	0.9285
TP: lymphadenopathy	0.8298
TP: soft tissue	0.7580

(d) Lesion #16556

TP: axilla	0.9932
FP: axilla lymph node	0.9819
TP: enhancing	0.8566
TP: soft tissue attenuation	0.8255
FP: conglomerate	0.6118
FN: mass	0.4684
FN: thickening	0.3866
FN: skin	0.0612

(e) Lesion #18470

TP: peripancreatic lymph node	0.9582
TP: porta Hepatis lymph node	0.8937
TP: lymphadenopathy	0.8210
TP: paracaval lymph node	0.5750

(f) Lesion #6479

TP: right adrenal gland	0.9993
TP: adrenal gland	0.9987
TP: adenoma	0.9861
TP: mass	0.7416
TP: nodule	0.7357
FN: hypodense	0.3862

(g) Lesion #275

TP: paraaortic	0.9027
TP: retroperitoneum	0.8617
TP: lymph node	0.8300
FP: aorta	0.6216
TP: lymphadenopathy	0.5605
FP: conglomerate	0.4281

(h) Lesion #15600

TP: tiny	0.9625
TP: mesentery lymph node	0.8954
FP: fat	0.8287
TP: soft tissue attenuation	0.7177
FP: intestine	0.6258

(i) Lesion #32328

TP: spleen	0.9925
TP: hypodense	0.9338
FP: metastasis	0.8404
TP: indistinct	0.7976

(j) Lesion #17942

TP: enhancing	0.9169
TP: large	0.8619
TP: abdomen	0.8163
TP: conglomerate	0.7866
TP: soft tissue	0.7014
FN: calcified	0.6624

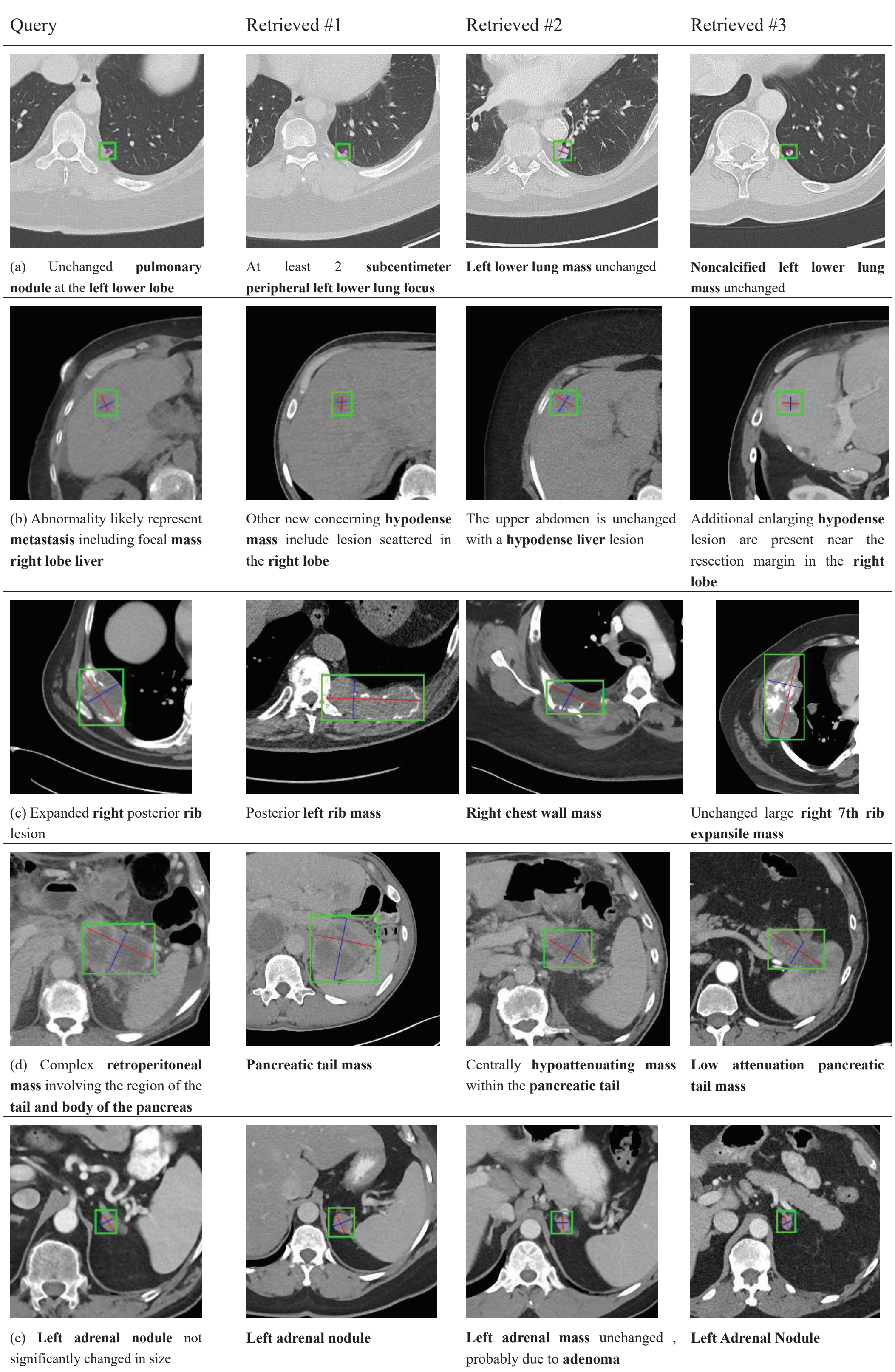
(k) Lesion #12134

TP: bone	0.9962
TP: pelvis	0.9848
TP: sclerotic	0.9777

(l) Lesion #27438

TP: pelvis	0.9959
TP: urinary bladder	0.9910
TP: calcified	0.9854
FP: pelvic wall	0.9595
TP: hyperdense	0.8865
FP: enhancing	0.8762
FP: pelvic bone	0.8642

Figure 1. Sample predicted labels with confidence scores on the text-mined test set. Green, red, and blue results correspond to TPs, FPs, and FNs, respectively. Underlined labels are TPs with missing annotations, thus were treated as FPs during evaluation. Only the most fine-grained predictions are shown with their parents omitted for clarity.



Query	Retrieved #1	Retrieved #2	Retrieved #3
(a) Unchanged **pulmonary nodule** at the **left lower lobe**	At least 2 **subcentimeter peripheral left lower lung focus**	**Left lower lung mass** unchanged	**Noncalcified left lower lung mass** unchanged
(b) Abnormality likely represent **metastasis** including focal **mass right lobe liver**	Other new concerning **hypodense mass** include lesion scattered in the **right lobe**	The upper abdomen is unchanged with a **hypodense liver** lesion	Additional enlarging **hypodense** lesion are present near the resection margin in the **right lobe**
(c) Expanded **right** posterior **rib** lesion	Posterior **left rib mass**	**Right chest wall mass**	Unchanged large **right 7th rib expansile mass**
(d) Complex **retroperitoneal mass** involving the region of the **tail and body of the pancreas**	**Pancreatic tail mass**	Centrally **hypoattenuating mass** within the **pancreatic tail**	**Low attenuation pancreatic tail mass**
(e) **Left adrenal nodule** not significantly changed in size	**Left adrenal nodule**	**Left adrenal mass** unchanged , probably due to **adenoma**	**Left Adrenal Nodule**

Figure 2. Sample lesion retrieval results of LesaNet. The input of LesaNet is the lesion image patch only, whereas the associated report sentence is shown for reference. The irrelevant words in the sentences describing other lesions have been removed for clarity.

Method	Text-mined test set				Hand-labeled test set			
	AUC	Precision	Recall	F1	AUC	Precision	Recall	F1
LesaNet	$93.24_{0.08}$	$30.89_{1.23}$	$53.74_{1.62}$	$31.76_{0.90}$	$\mathbf{93.83_{0.18}}$	$47.01_{2.09}$	$54.63_{1.41}$	$\mathbf{42.29_{1.08}}$
w/o score propagation layer	$92.42_{0.09}$	$\mathbf{34.25_{2.60}}$	$\underline{49.61_{1.55}}$	$30.89_{0.83}$	$\underline{93.28_{0.30}}$	$\mathbf{50.60_{2.06}}$	$\underline{51.74_{1.72}}$	$41.09_{1.09}$
w/o RHEM	$93.21_{0.10}$	$\underline{28.40_{1.49}}$	$\mathbf{56.05_{2.19}}$	$31.02_{0.93}$	$93.62_{0.22}$	$\underline{43.09_{1.49}}$	$\mathbf{57.65_{2.11}}$	$42.04_{1.06}$
w/o label expansion	$\underline{92.37_{0.12}}$	$30.16_{1.72}$	$55.68_{1.95}$	$\underline{30.73_{0.60}}$	$93.32_{0.30}$	$45.61_{2.09}$	$55.87_{3.14}$	$\underline{40.94_{1.24}}$
w/o text-mining module	$\mathbf{93.27_{0.09}}$	$30.79_{1.43}$	$53.77_{1.90}$	$\mathbf{31.94_{1.16}}$	$93.68_{0.23}$	$46.16_{2.05}$	$54.05_{2.68}$	$41.49_{0.65}$
w/o triplet loss	$93.03_{0.07}$	$30.65_{1.94}$	$53.91_{1.86}$	$31.60_{1.19}$	$93.56_{0.18}$	$46.29_{1.30}$	$54.73_{1.53}$	$41.84_{1.22}$

Table 1. Multilabel classification accuracy averaged across labels on two test sets. Bold results are the best ones. Red underlined results in the ablation studies are the worst ones, indicating the ablated strategy is the most important for the criterion. We report mean and standard deviation of accuracies calculated on 10 random data splits formatted as mean $_{\text{std.}}$.