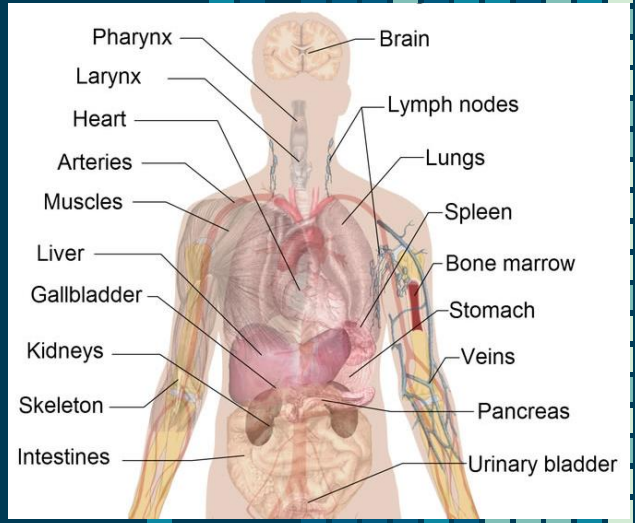
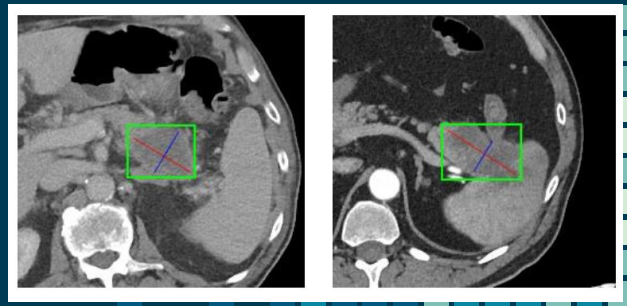
Large-scale, Universal Lesion Mining and Analysis in Medical Images





Ke Yan
Senior Researcher
PAII Bethesda Research Lab

5/20/2020

Motivation

- Lesion analysis
  - **Radiologists**: find, measure, describe, compare, …
  - **Algorithms**: detect, segment, classify, retrieve, …
- Existing studies
  - Focus on certain body parts
    - Lung, breast, liver, brain, etc.
  - Require large annotation effort to annotate a small set of images (~1K CT volumes)

Motivation

- Our goal
  - Mine large-scale lesion data from PACS, with minimum human efforts
  - Explore a variety of lesions (universal)
  - Perform multiple clinically important tasks
  - And eventually, help in radiologists' daily work and improve the efficiency and accuracy

Step 1:

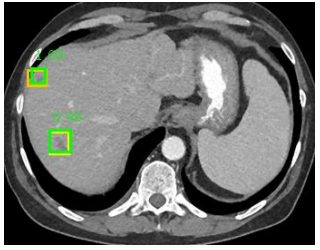
Data curation

Mining from PACS

Human annotation

Step 2:

Lesion detection

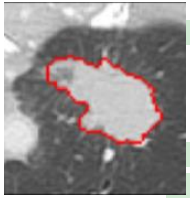


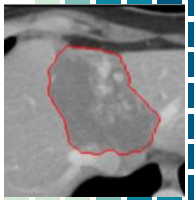
or

Human selection

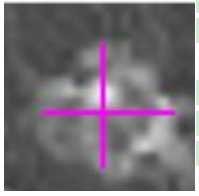
Step 3:

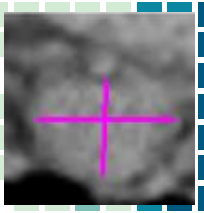
Segmentation



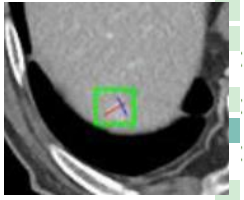


Measurement



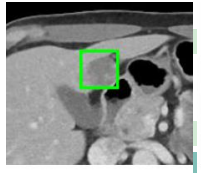


Classification

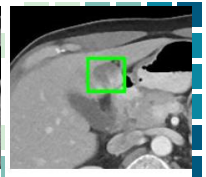


liver
hemangioma
enhancing

Matching

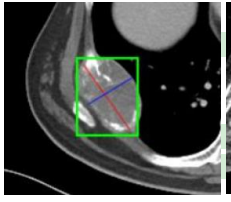


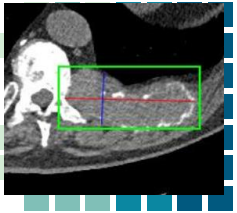
Study 1, series 1



Study 2, series 1

Retrieval





Step 1:

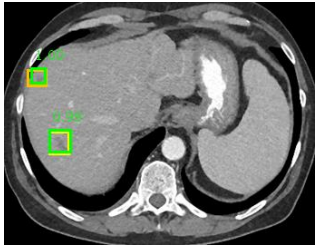
Data curation

Mining from PACS

Human annotation

Step 2:

Lesion detection

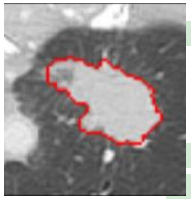


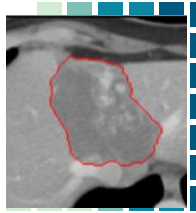
or

Human selection

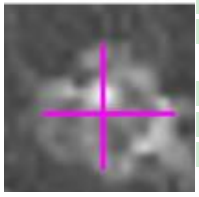
Step 3:

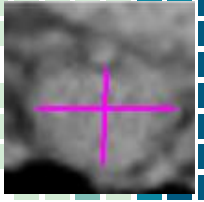
Segmentation



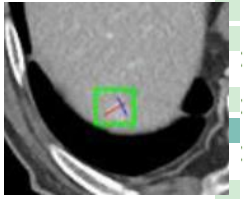


Measurement



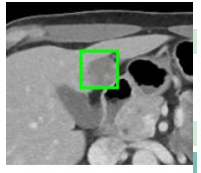


Classification

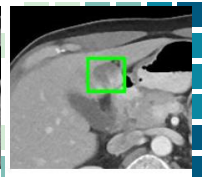


liver
hemangioma
enhancing

Matching

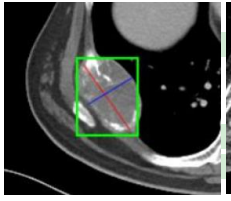


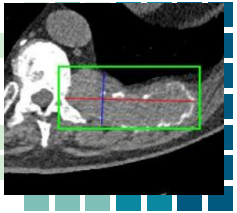
Study 1, series 1



Study 2, series 1

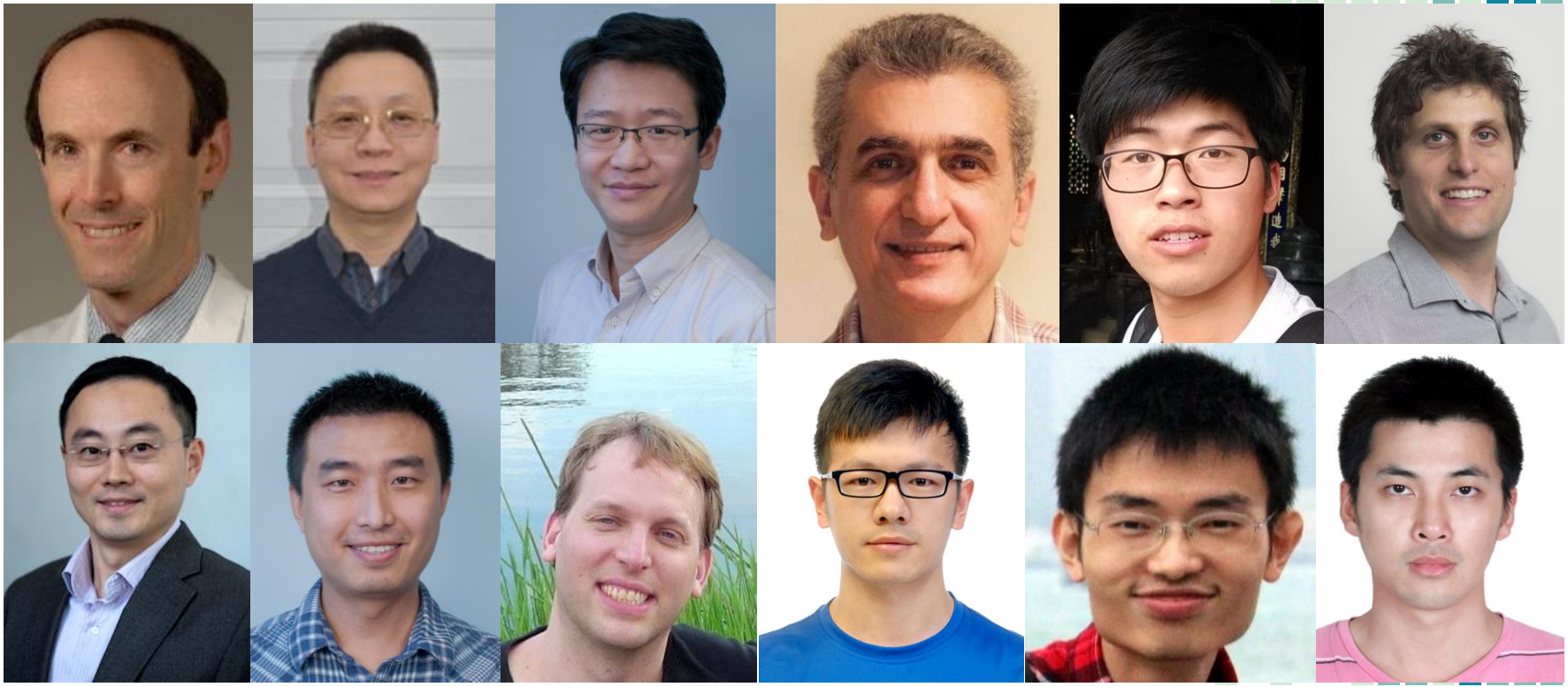
Retrieval





Imaging Biomarkers and Computer-Aided Diagnosis Laboratory,
National Institutes of Health + National Library of Medicine

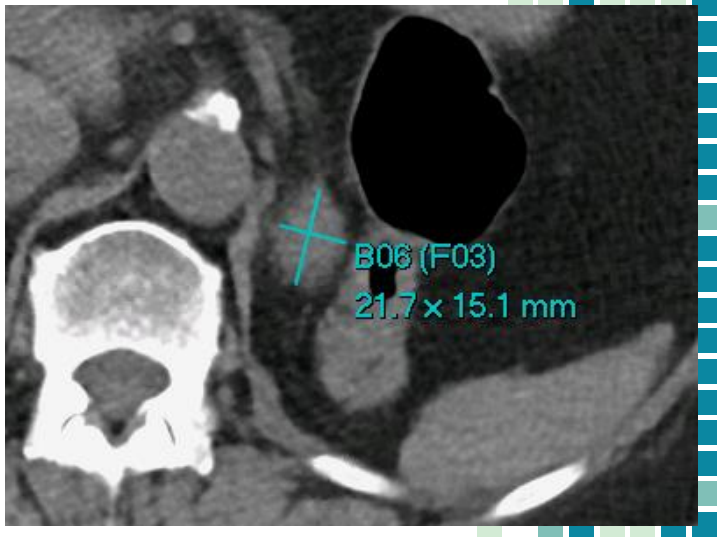
NIH National Institutes of Health



Data Curation

Ke Yan, Xiaosong Wang, Le Lu, Ronald M. Summers, "**DeepLesion: Automated Mining of Large-Scale Lesion Annotations and Universal Lesion Detection with Deep Learning**", Journal of Medical Imaging, 2018

The DeepLesion dataset

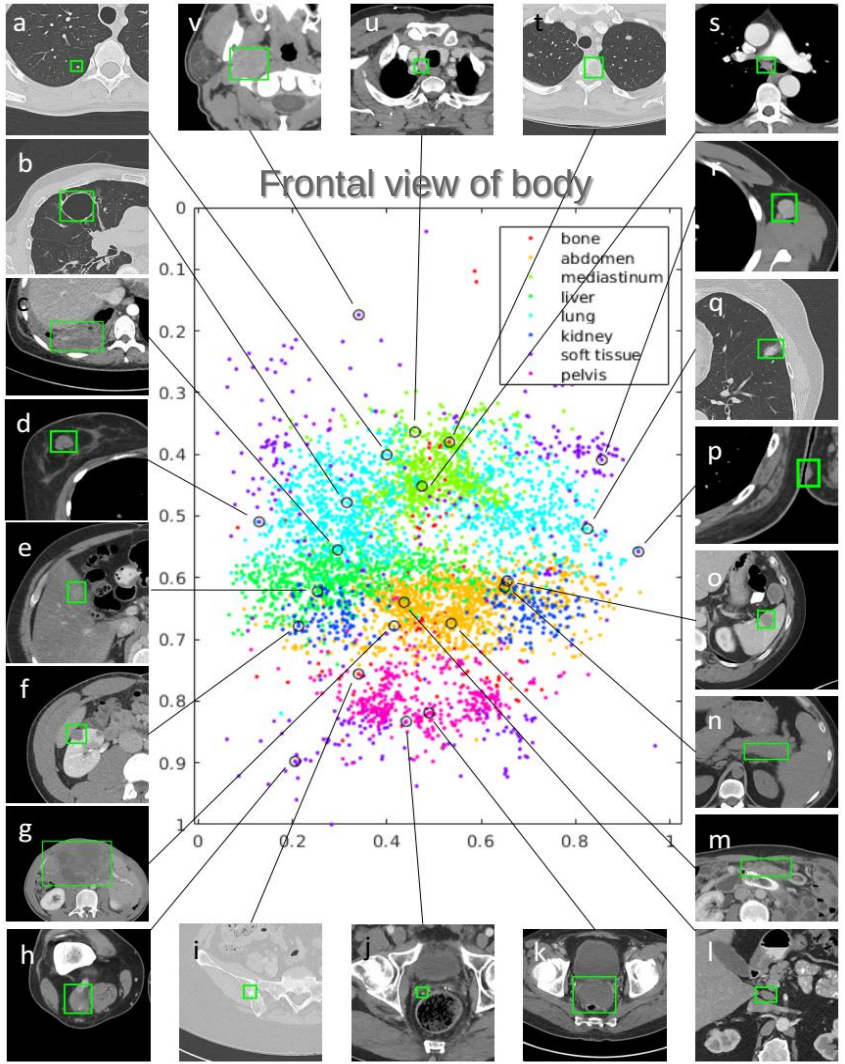


- Dataset collection by mining "bookmarks"
  - Marked by radiologists in their daily work
  - Measure significant abnormalities or "lesions" according to the RECIST (Response Evaluation Criteria in Solid Tumors) guidelines
  - Collected over years and stored in hospitals' PACS

The DeepLesion dataset

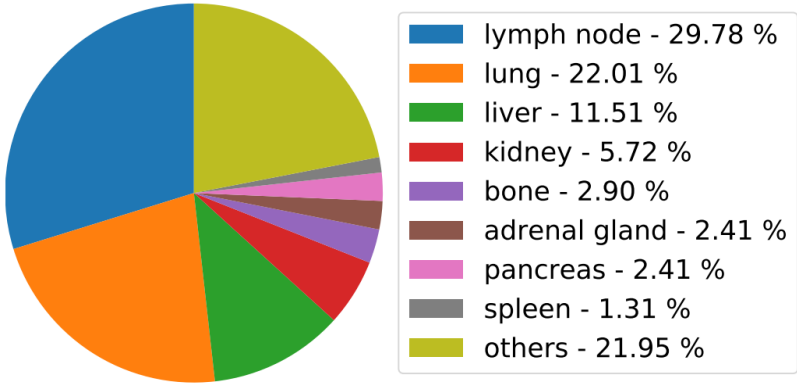
- 4,427 patients
- 10,594 CT studies
- 928K 2D images
- 32,735 lesions
- 0.2 ~ 343 mm in size

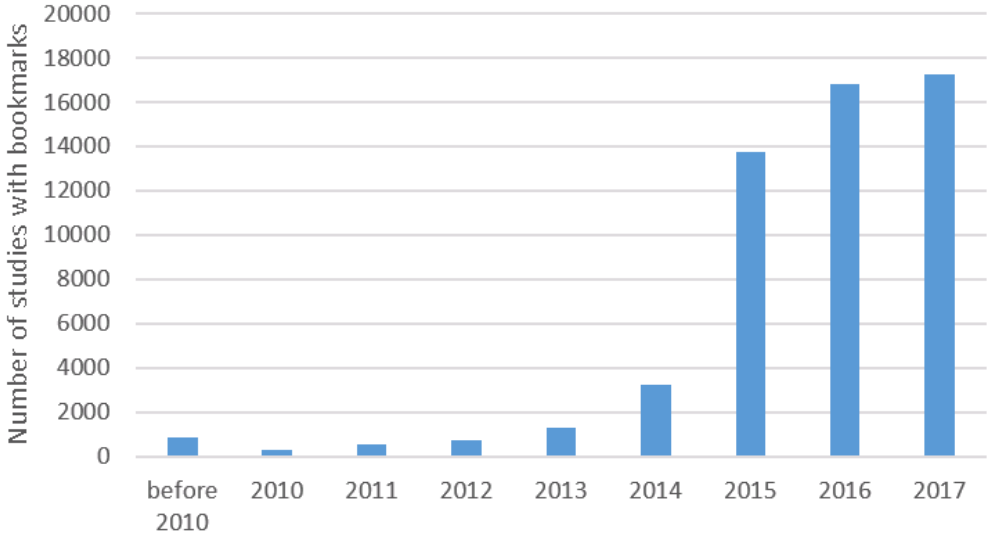
https://nihcc.app.box.com/v/DeepLesion



The DeepLesion project

- Economical
- Universal
- Systematic
- Challenging
  - Many lesion types
  - Relatively limited data
  - Subtle appearance
  - Imperfect labels





What is good in universality?

- Radiologists are responsible to find and report all possible abnormal findings
- Single-type models are unable to cover all
  - Single-type and universal models can be complementary
- More in-depth analysis possibilities
  - Retrieval, relation analysis, reasoning, …

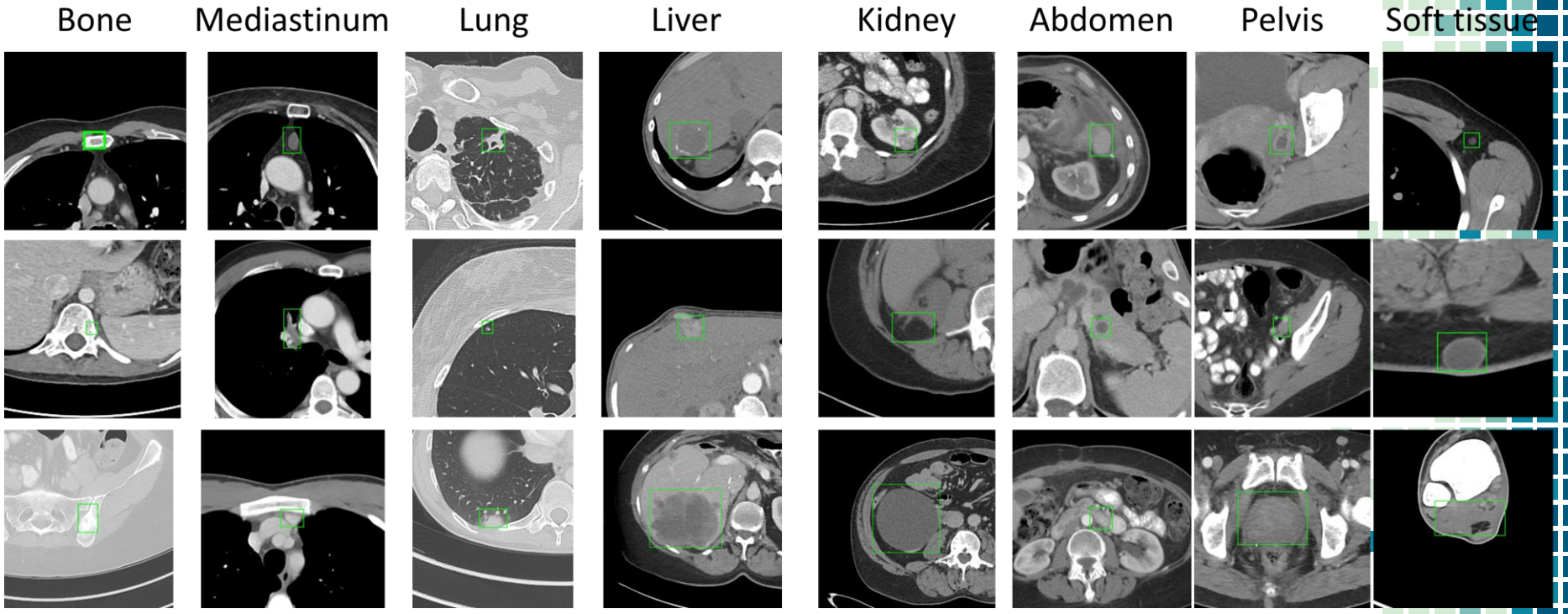
Retrieval and Matching

K. Yan, X. Wang, L. Lu, L. Zhang, A. P. Harrison, M. Bagheri, R. M. 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*, 2018.

Motivation

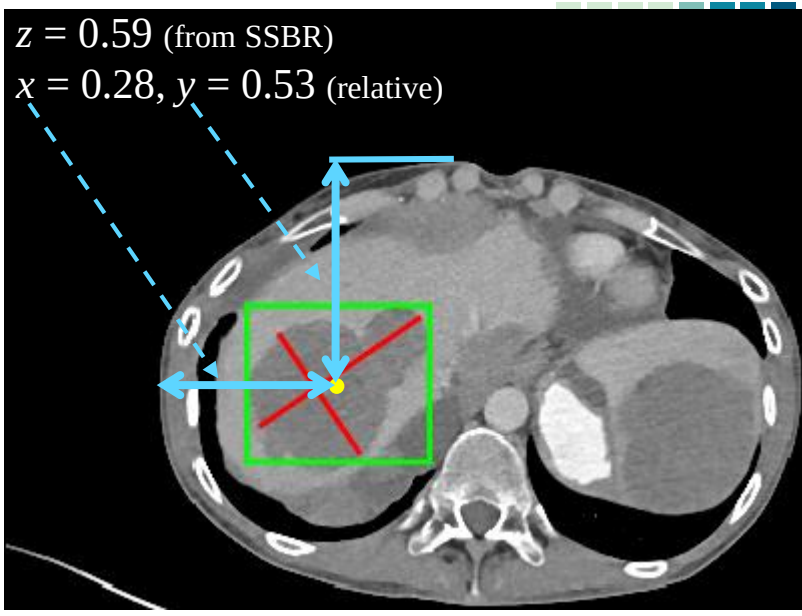
- Model the similarity between lesions
- **Retrieval**: find similar lesions from other patients
  - Usage: help understanding
- **Matching**: find identical lesion instance from the same patient
  - Usage: longitudinal comparison
- **Approach**: learn deep lesion embedding on a large diverse dataset with weak cues

Supervision Cue (I): Coarse Body Part



Supervision Cue (II): Relative Body Location

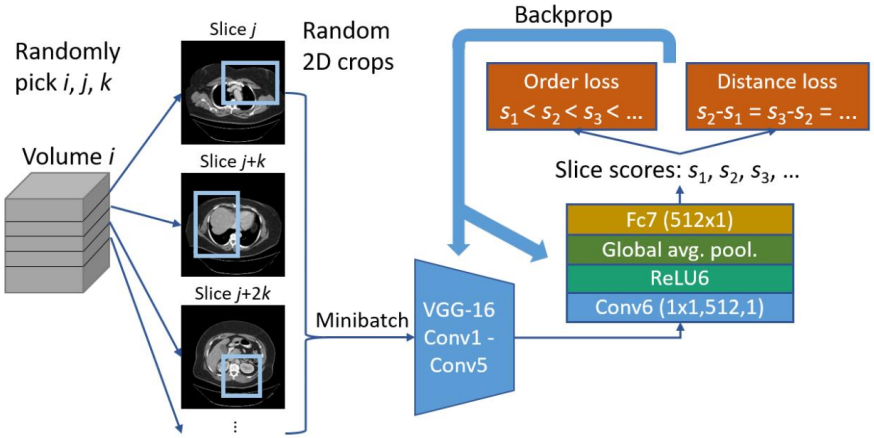
- X and Y: easy ☺
- Z: self-supervised body part regressor (SSBR)
- SSBR
  - Intuition: volumetric medical images are intrinsically structured!
  - The superior-inferior slice order information can be leveraged for self-supervision



Yan, Lu, Summers. **Unsupervised Body Part Regression via Spatially Self-ordering Convolutional Neural Networks**, ISBI 2018

Supervision Cue (II): Relative Body Location

- *h* is the sigmoid function, *g* is the smooth L1 loss
- The order loss and distance loss terms collaborate to push each slice score towards the correct direction relative to other slices

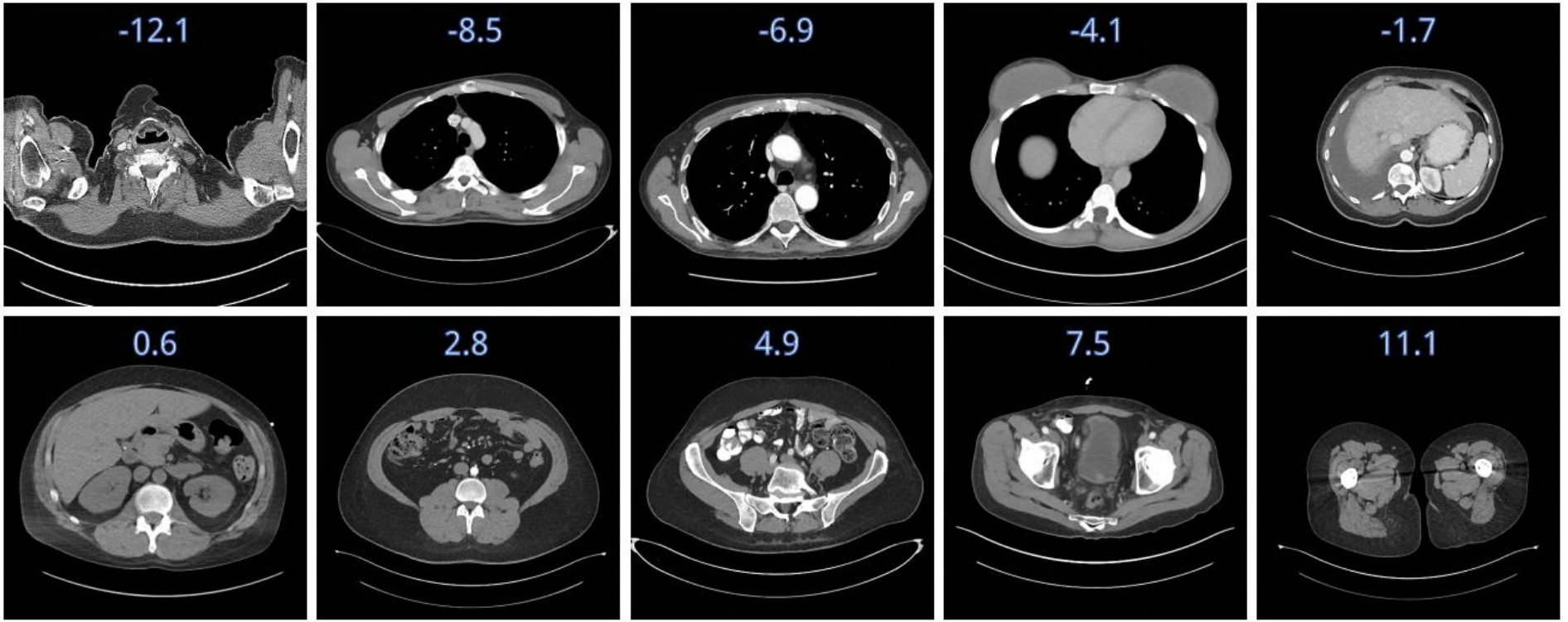


$$L_{\text{SSBR}} = L_{\text{order}} + L_{\text{dist}};$$

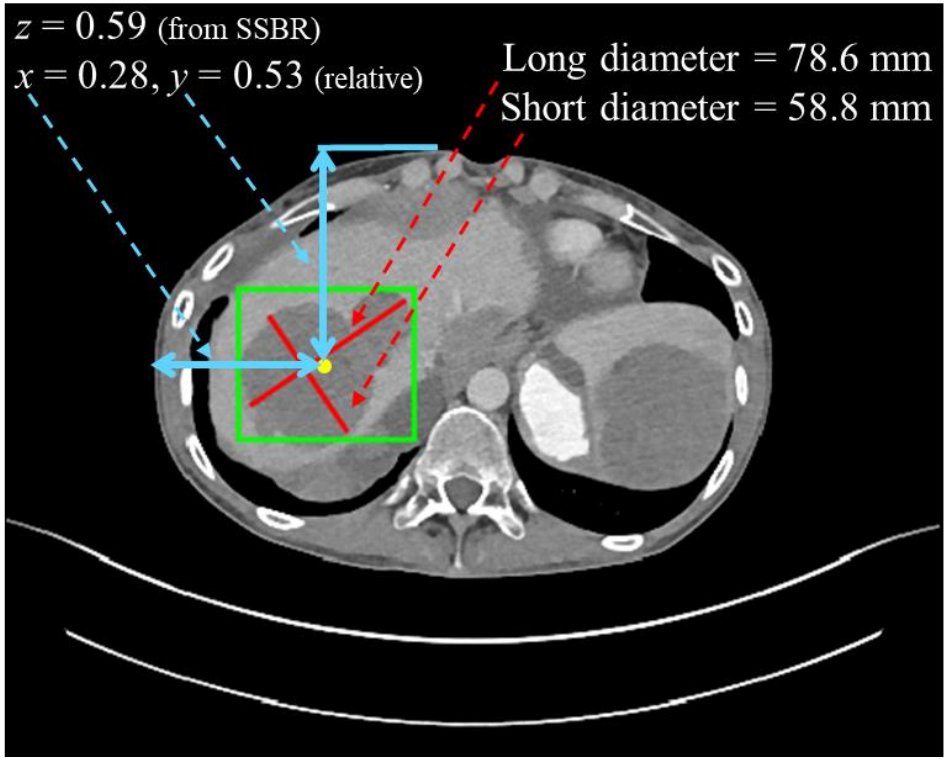
$$L_{\text{order}} = -\sum_{i=0}^{m-2} \log h\left(s_{j+k(i+1)} - s_{j+ki}\right);$$

$$L_{\text{dist}} = \sum_{i=0}^{m-3} g(\Delta_{i+1} - \Delta_i),$$

$$\Delta_i = s_{j+k(i+1)} - s_{j+ki},$$


-12.1
-8.5
-6.9
-4.1
-1.7
0.6
2.8
4.9
7.5
11.1

Supervision Cue (III): Lesion Size

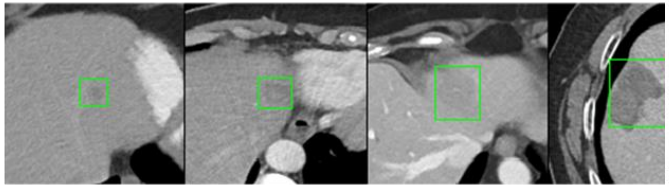


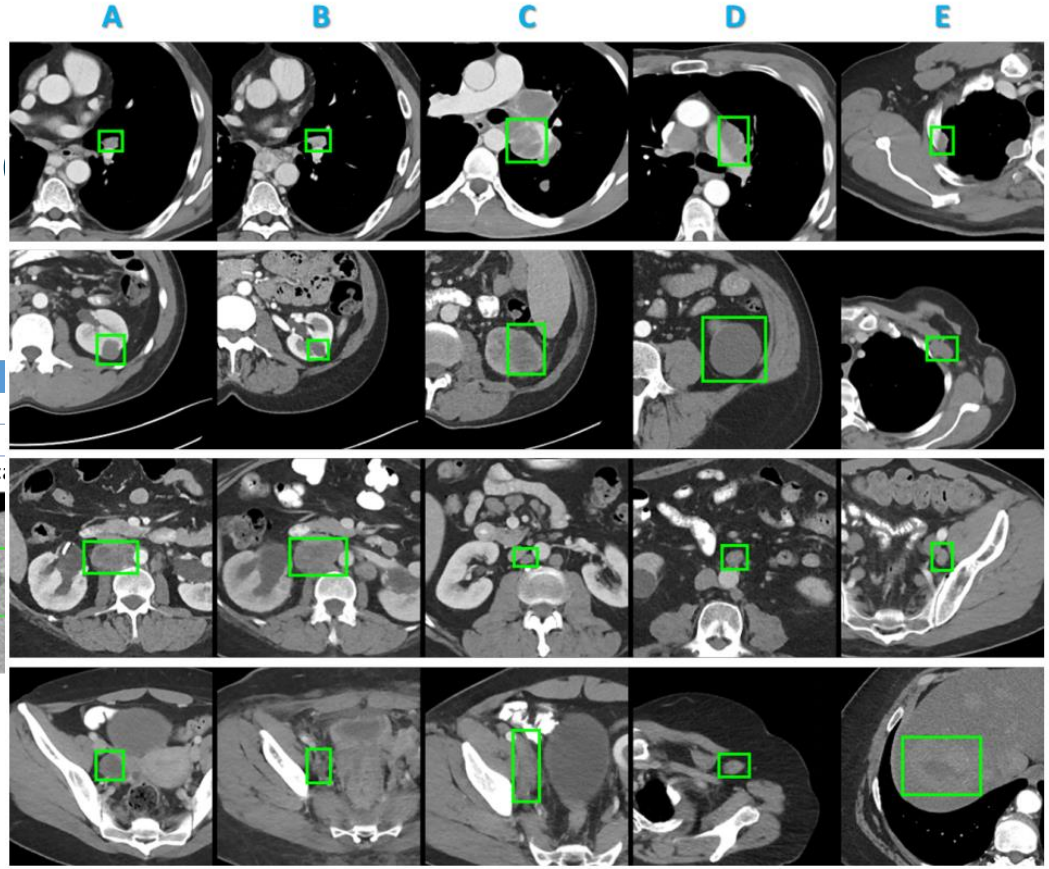
Algorithm

- Triplet network with s

Lesion	A	B	C	D
Same body part?		√	√	√
Similar location?	Anchor	√	√	x
Similar size?		√	x	Don't c

Sequential sampling





Algorithm

- Joint Loss function
  - A selected sequence of 5 instances can be decomposed into three triplets: {*ABC*, *ACD* and *ADE*}; Joint Loss →

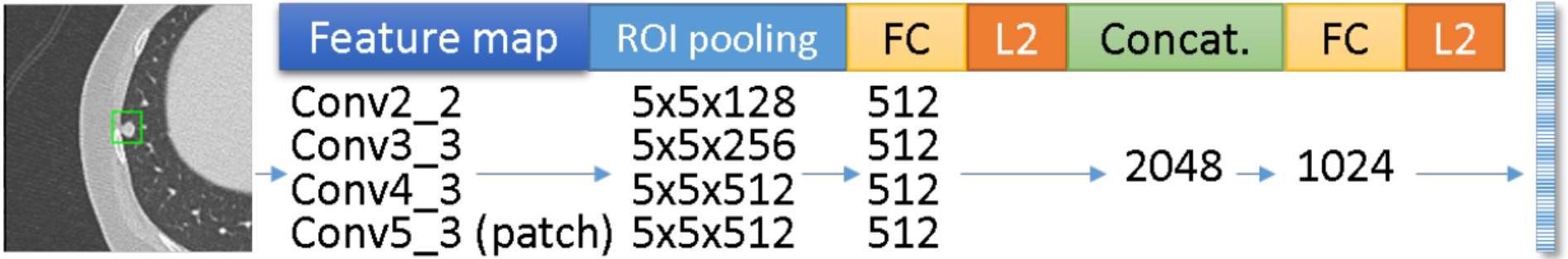
$$L = \frac{1}{2S}\sum_{i=1}^{S}\left[\max(0, d_{AB}^2 - d_{AC}^2 + m_1) + \max(0, d_{AC}^2 - d_{AD}^2 + m_2) + \max(0, d_{AD}^2 - d_{AE}^2 + m_3)\right]$$

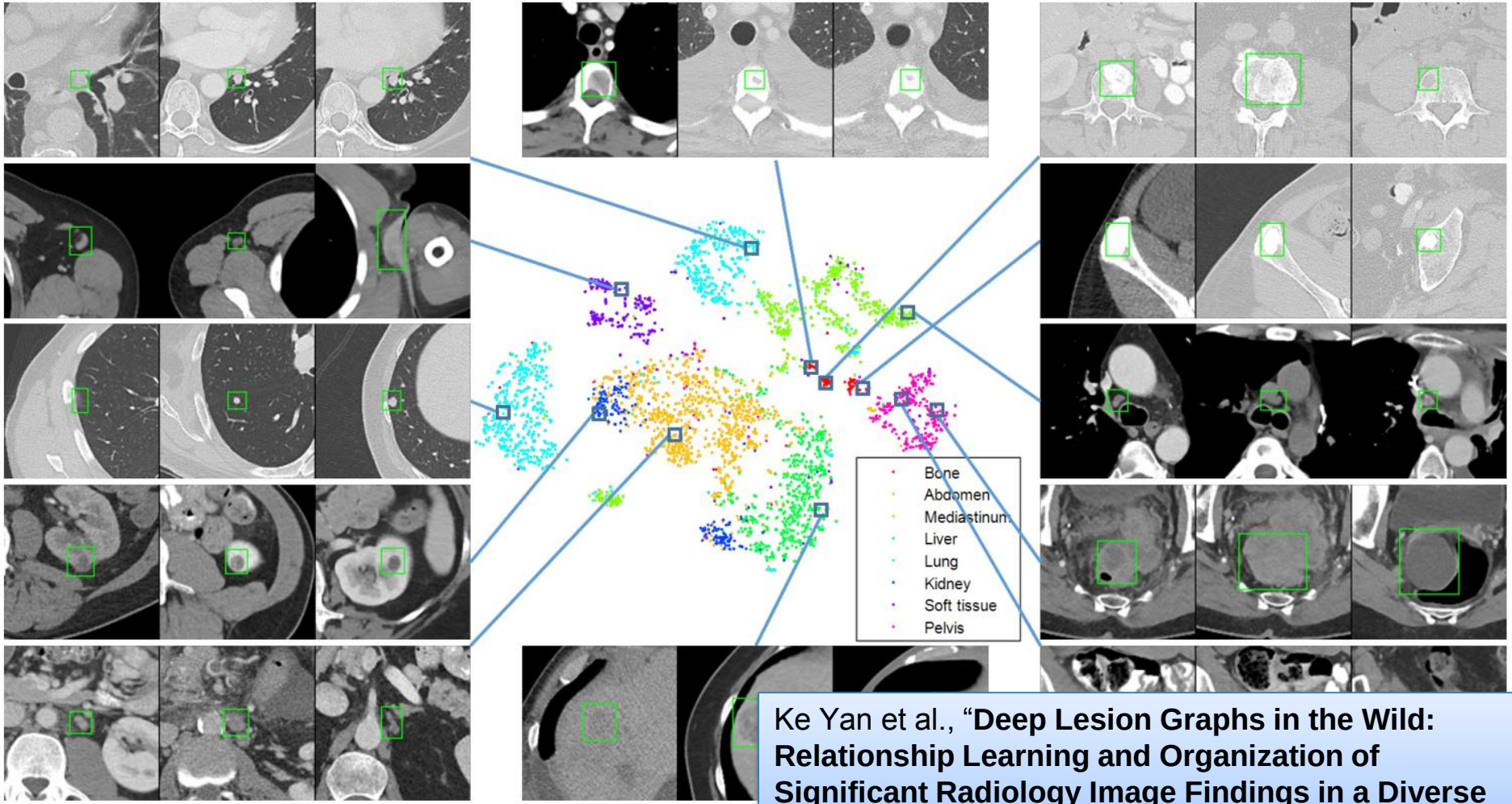
$$m_3 > m_2 > m_1 > 0$$

- Iterative refinement learning

Algorithm

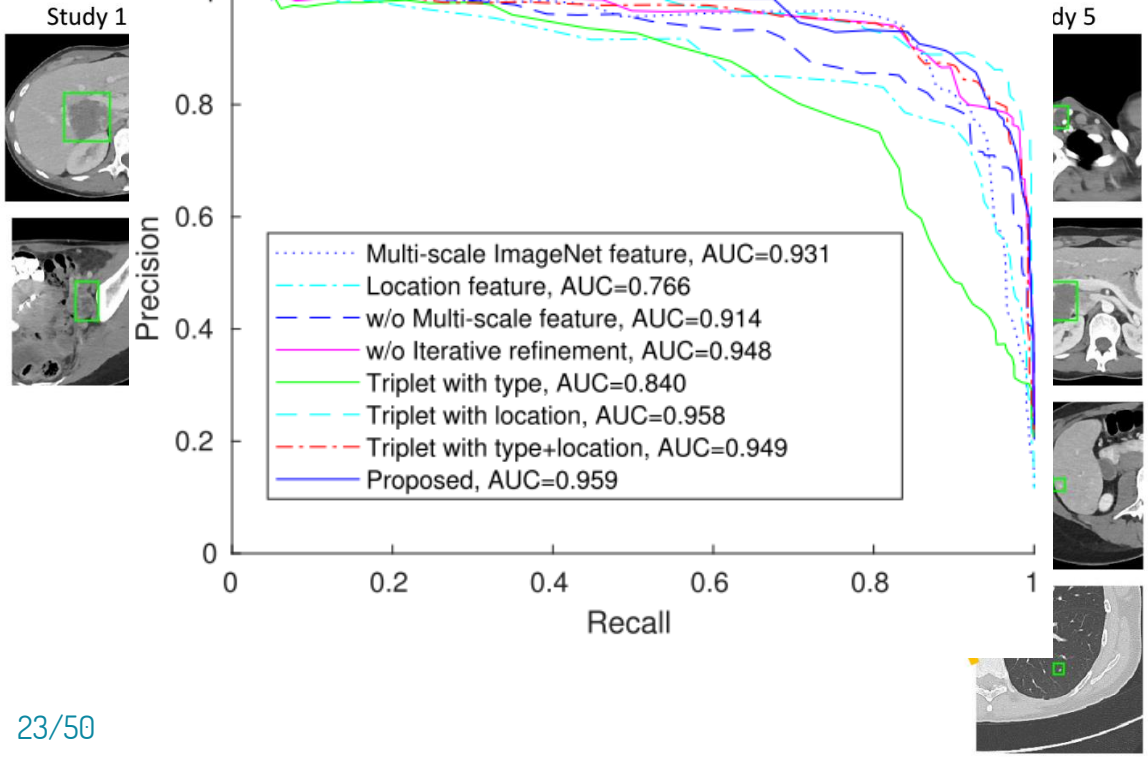
- Backbone: VGG-16
- Multi-scale, multi-crop
- Output: a **1024D** feature embedding vector for each lesion instance

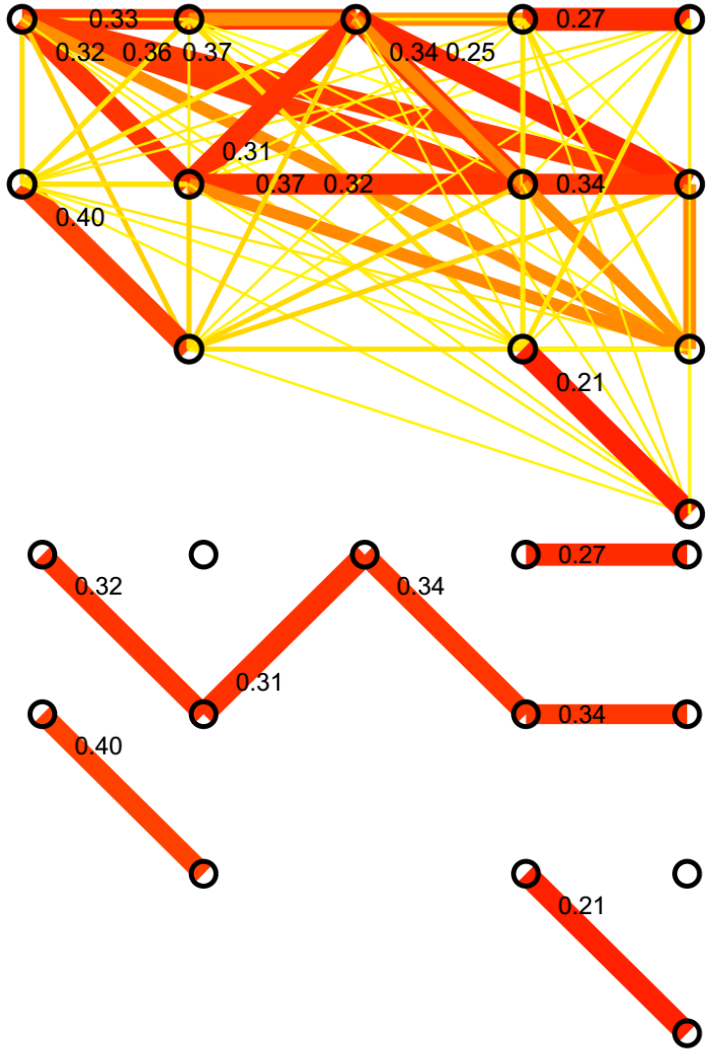




Ke Yan et al., "**Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-scale Lesion Database**," CVPR 2018.

Lesion matching



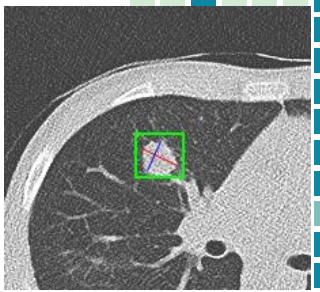


Lesion Classification

K. Yan, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, and R. M. Summers, "**Holistic and comprehensive annotation of clinically significant findings on diverse CT images: Learning from radiology reports and label ontology**," in CVPR, 2019.

Motivation

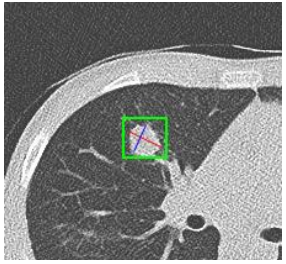
Where
What
How



- Problem
  - Fine-grained semantic information is missing
- Purpose
  - Predict semantic labels of a lesion
  - Assist diagnostic decision making
  - Generate structured reports
  - Collect lesion datasets
  - Find similar lesions

Motivation

- Aim: Given a lesion image, predict a fine-grained set of relevant labels, such as the lesion's **body part, type, and attributes**

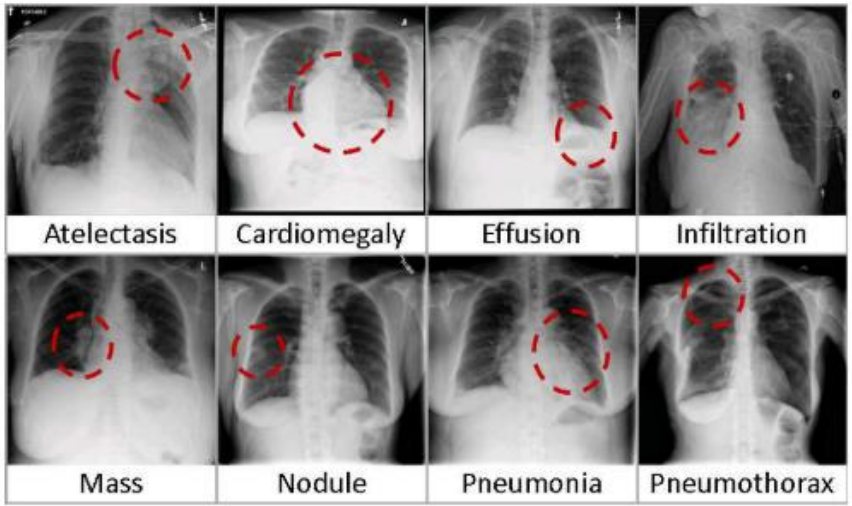


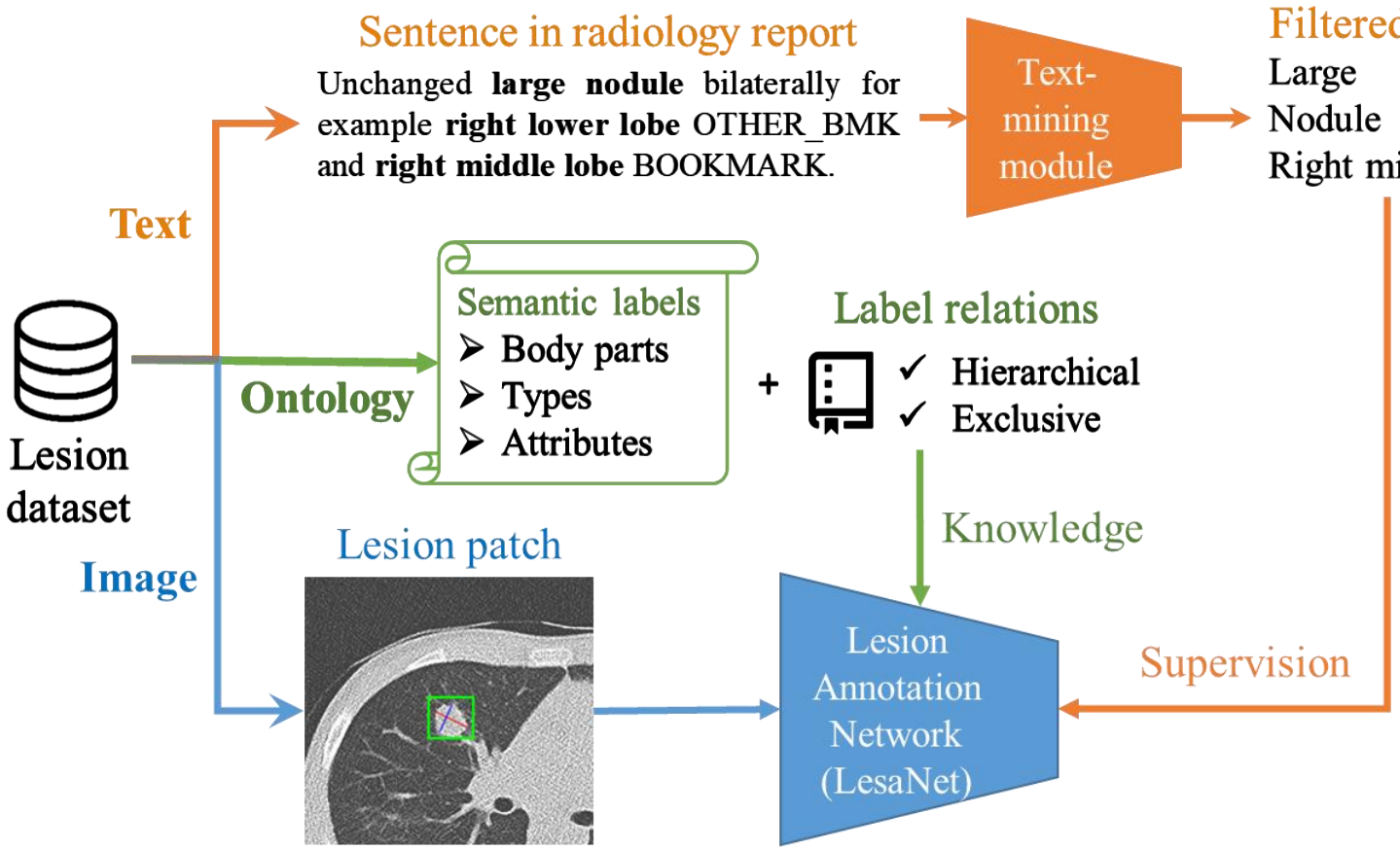
Nodule: 0.93
Right mid lung: 0.92
Lung mass: 0.89
Perihilar: 0.64
...

- Approach: Mine labels from radiological reports

Related work: mine labels from reports

- Only image-level labels are available
  - Not sufficient for lesion-level prediction
- Label set can be improved
  - Label size is limited
  - Label relation is not considered




Sentence in radiology report
Unchanged large nodule bilaterally for example right lower lobe OTHER_BMK and right middle lobe BOOKMARK.
Text-mining module
Filtered labels
Large
Nodule
Right mid lung
Text
Lesion dataset
Ontology
Semantic labels
Body parts
Types
Attributes
+
Label relations
Hierarchical
Exclusive
Knowledge
Image
Lesion patch
Lesion Annotation Network (LesaNet)
Supervision

Radiology lexicon

- Source: RadLex v3.15
  - 46,658 terms related to radiology
- Keep labels related to body part, lesion type, and attributes
- Add some missing synonyms (e.g. adjectives)
- Sentence (w/ bookmark) tokenization
- Whole-word string matching

Lesion ontology

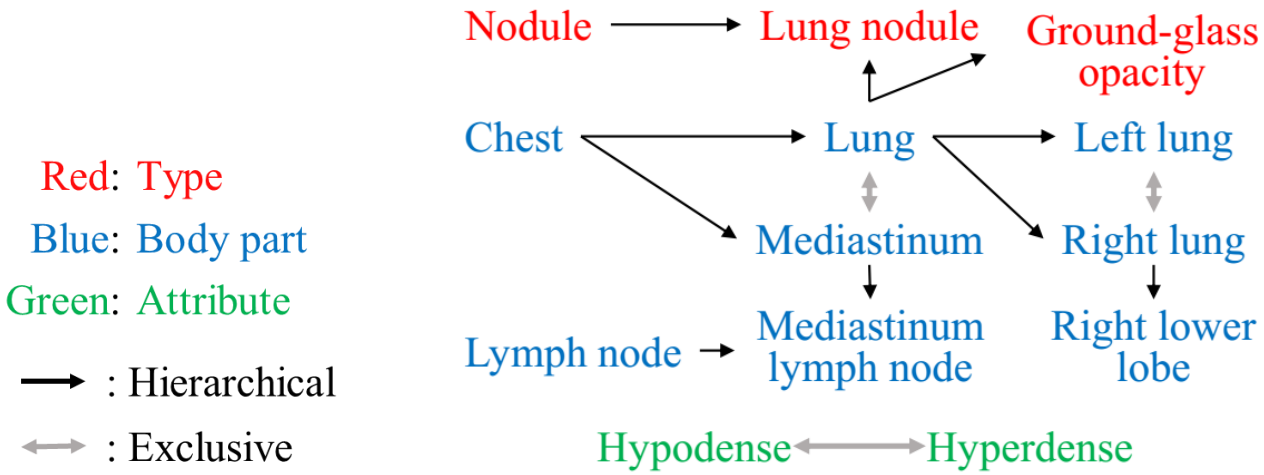
- Body parts (115)
  - coarse-level (e.g., chest, abdomen)
  - organs (lung, lymph node)
  - fine-grained organ parts (right lower lobe, pretracheal LN)
  - other body regions (porta hepatis, paraspinal)
- Types (27)
  - general terms (nodule, mass)
  - more specific ones (adenoma, liver mass)
- Attributes (29)
  - intensity, shape, size, etc. (hypodense, spiculated, large)

Label relation

- Hierarchical relation
  - A fine-grained body part is *part of* a coarse-scale one (left lung < lung)
  - A type is *sub-type* of another one (hemangioma < neoplasm)
  - A type is *located in* a body part (lung nodule < lung)
  - Extraction from RadLex → manual correction, 137 parent-child pairs

Label relation

- Mutually exclusive relation
  - Manually annotate, 4,461 pairs



Relevant label extraction

- Some labels in the sentence is irrelevant or uncertain

Unchanged **large nodule** bilaterally for example ~~**right lower lobe**~~ OTHER_BMK and **right middle lobe** BOOKMARK.

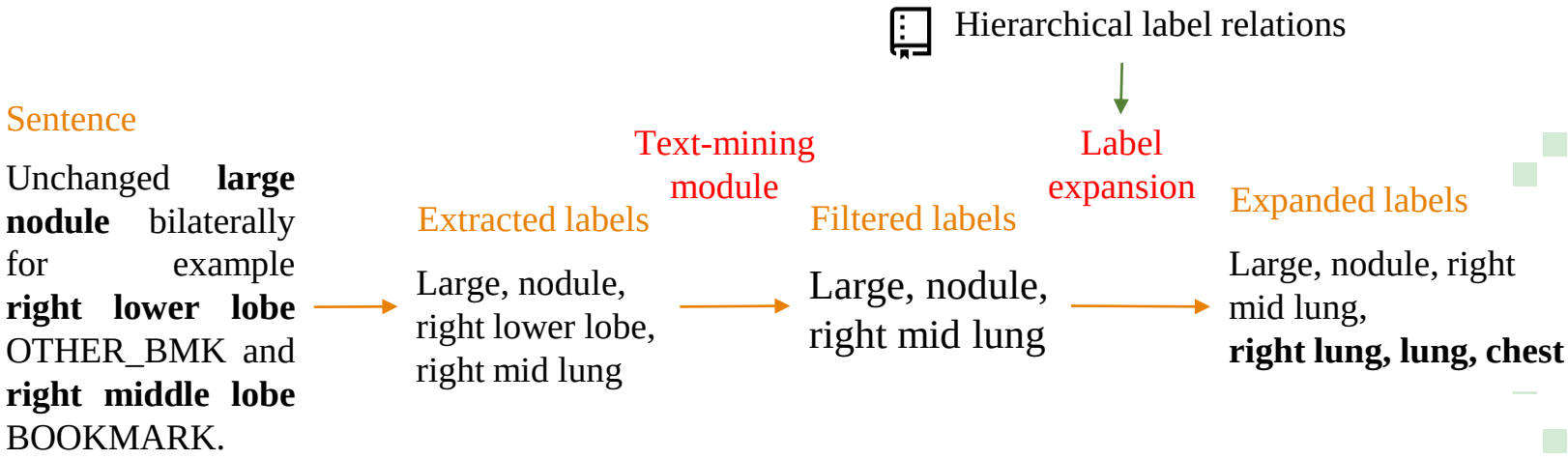
Dense or ***enhancing*** lower right **liver** lesion BOOKMARK possibly due to ***hemangioma.***

- To remove irrelevant labels, we propose a text-mining module: relation extraction CNN followed by rule filters

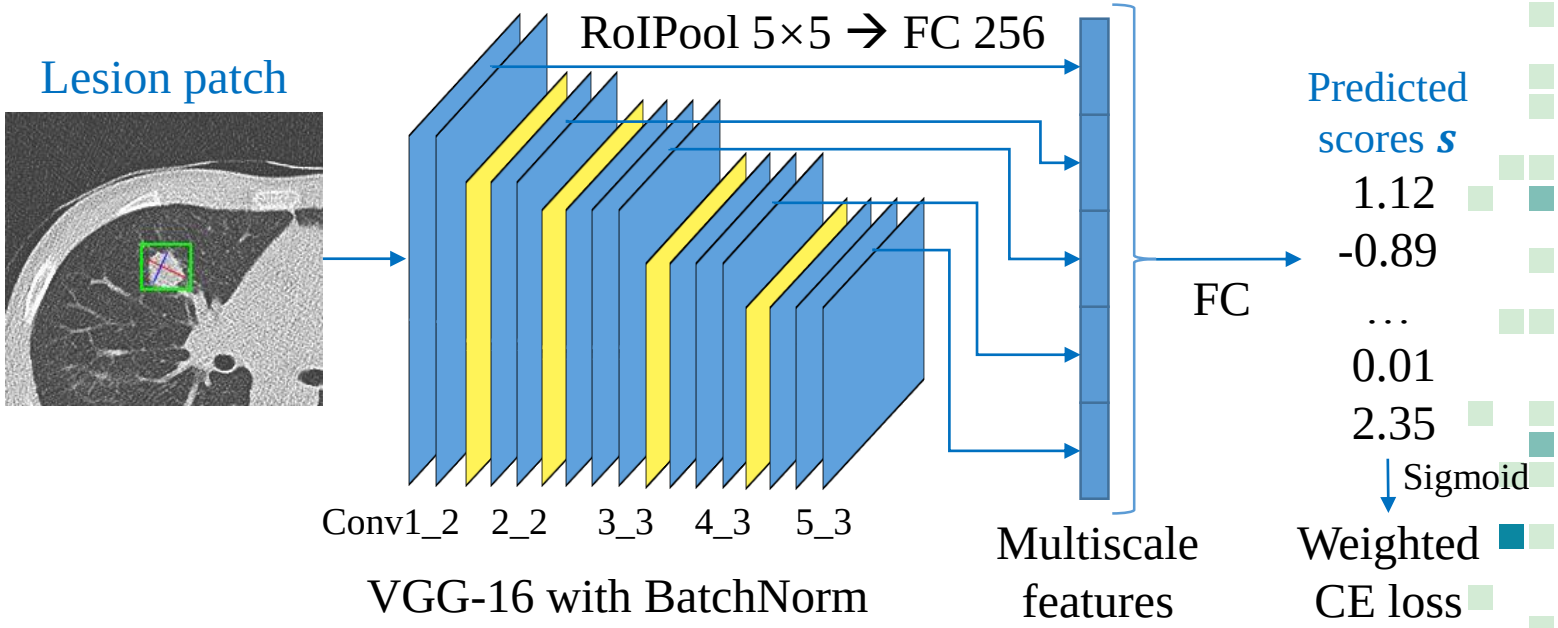
Yifan Peng et al., "**A self-attention based deep learning method for lesion attribute detection from CT reports**," IEEE International Conference on Healthcare Informatics (ICHI), 2019.

Label expansion

- Infer the missing parent labels



LesaNet: Multiscale multilabel CNN

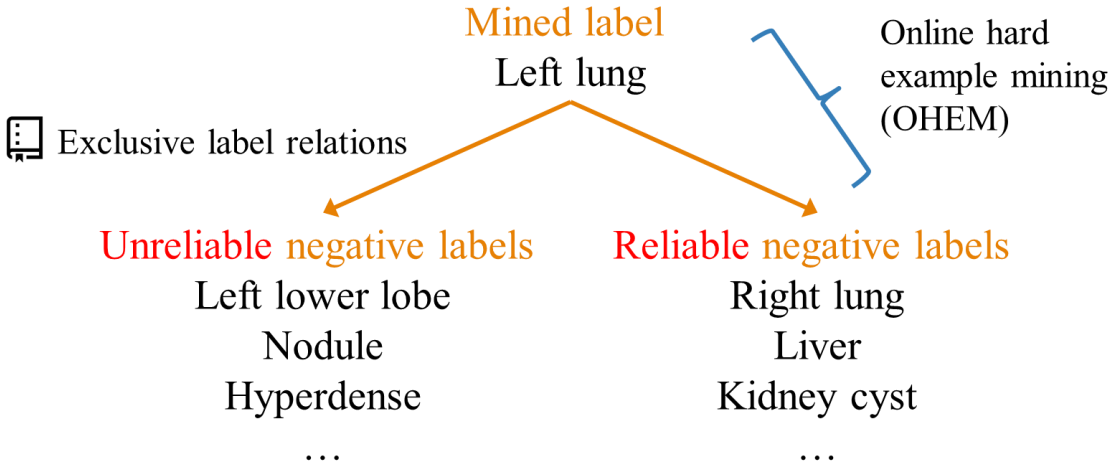


Relational hard example mining (RHEM)

- Motivation
  - Some labels/samples are difficult to learn
- Idea
  - Online hard example mining (OHEM)
- Problem
  - Mined labels are incomplete, so the negative labels may be unreliable
  - OHEM may treat missing labels as hard negatives

Relational hard example mining (RHEM)

- Solution
  - Use mutually exclusive label relation to infer reliable negative labels
  - OHEM is only performed on reliable labels → RHEM



Relational hard example mining (RHEM)

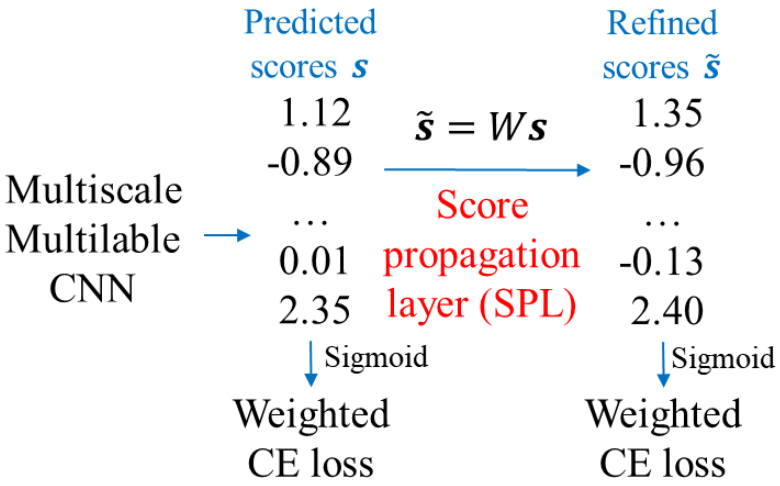
- Stochastic sampling strategy
  - Online difficulty of reliable label *c* of lesion *i*

$$\delta_{i,c} = |\sigma_{i,c} - y_{i,c}|^{\gamma}$$

  - Randomly sample examples (lesion-label pairs) in a minibatch according to δ
  - Examples with large δ are emphasized
- RHEM also works as a dynamic weighting mechanism for imbalanced labels

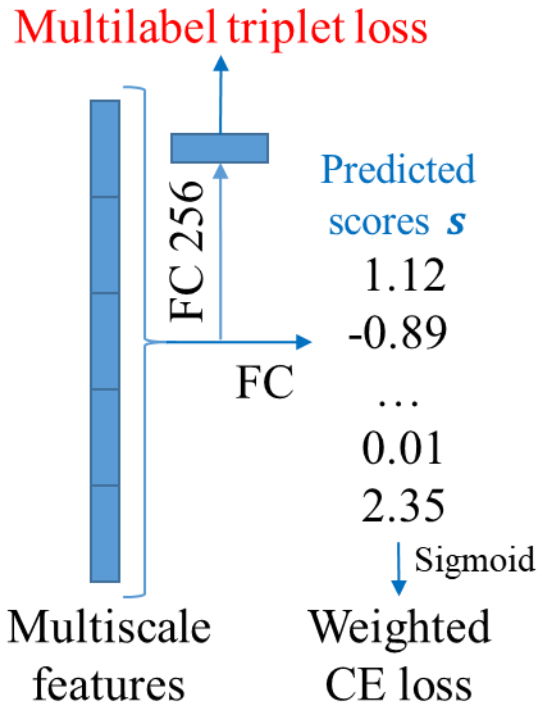
Score propagation layer

- Learn to capture the first-order correlation between labels
- W is initialized with an identity matrix



Joint classification and retrieval

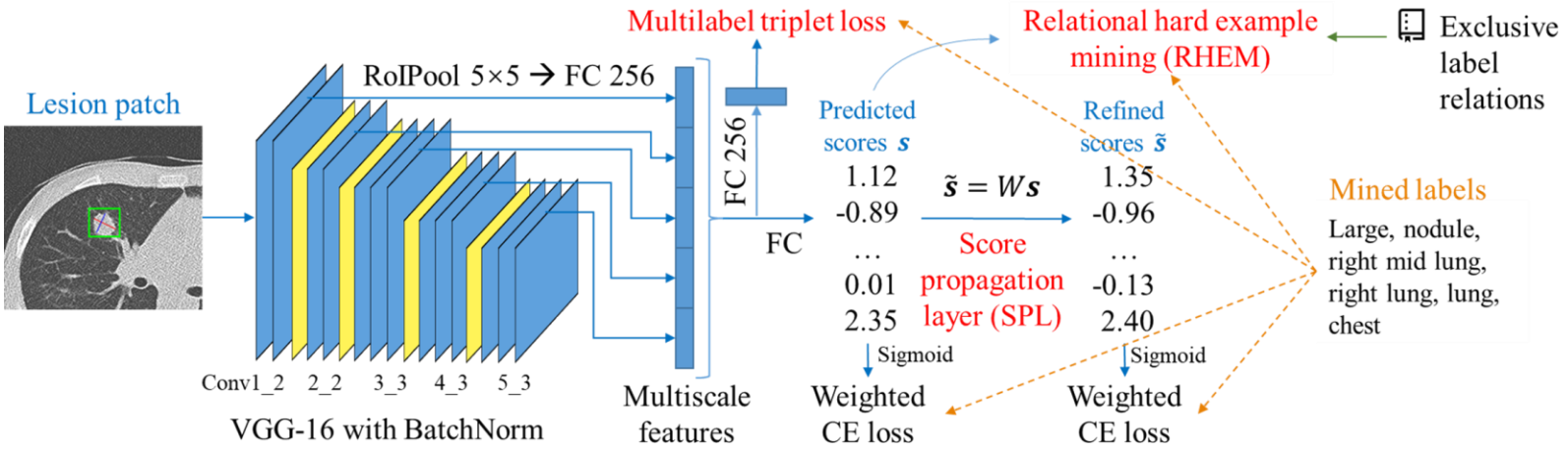
- Aim
  - Find lesions with similar semantic labels
  - Increase interpretability



Overall framework of LesaNet

- Loss function

$$L = L_{\text{WCE}} + L_{\text{CE, RHEM}} + L_{\text{WCE, SPL}} + \lambda L_{\text{triplet}}$$

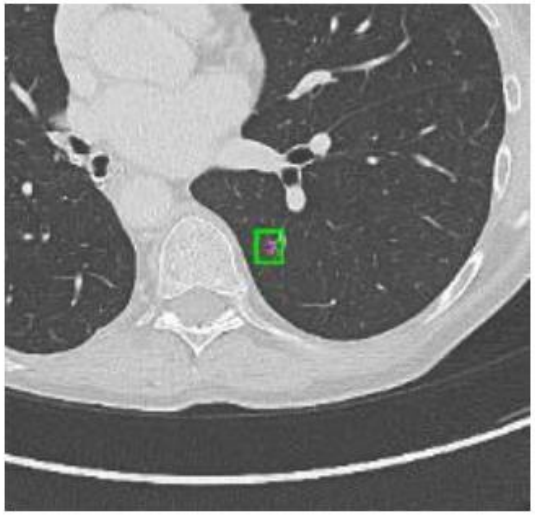


Dataset

- Training set: 19,213 lesions with sentences; validation: 1,852; test: 1,759 (**text-mined test set**)
- Two radiologists further manually annotated 500 random lesions in the test set **(hand-labeled test set**)
- **Input**: 120mm^2 3-channel lesion image patch

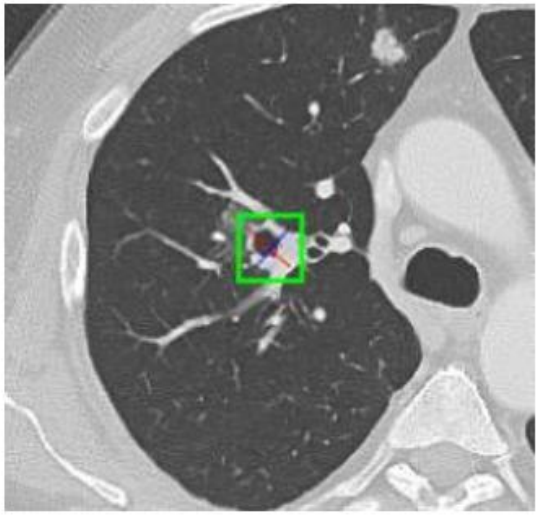
Ablation study

Method	Text-mined test set				Hand-labeled test set			
	AUC	Precision	Recall	F1	AUC	Precision	Recall	F1
LesaNet	**0.9344**	0.3593	0.5327	**0.3423**	**0.9398**	0.4737	0.5274	**0.4344**
w/o score propagation layer	0.9275	**0.3680**	0.4733	0.3233	0.9326	0.4833	0.4965	0.4092
w/o RHEM	0.9338	0.2983	**0.5550**	0.3178	0.9374	0.4341	0.5327	0.4303
w/o label expansion	0.9148	0.3523	0.5104	0.3270	0.9236	0.4503	0.5420	0.4205
w/o text-mining module	0.9334	0.3365	0.5350	0.3324	0.9392	**0.4869**	0.5361	0.4250
w/o triplet loss	0.9312	0.3201	0.5394	0.3274	0.9335	0.4645	**0.5624**	0.4337



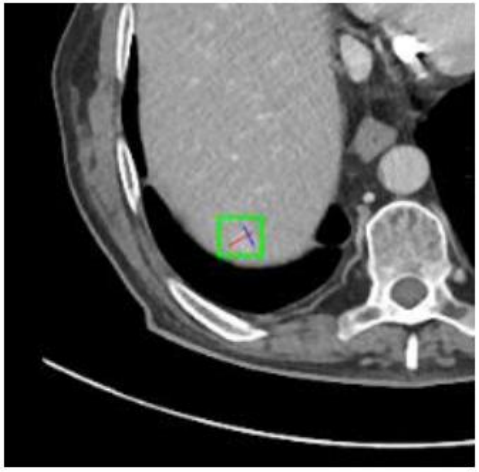
(c) Lesion #30088

TP: ground-glass opacity	0.9667
TP: nodule	0.9645
TP: left lower lobe	0.9617
TP: lung nodule	0.9108
FP: left upper lung	0.8122



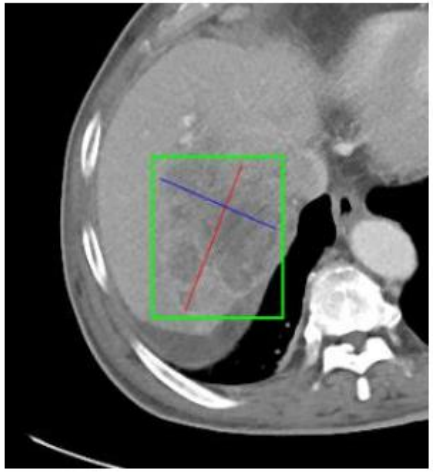
(d) Lesion #22789

TP: cavitary	0.9587
TP: right upper lobe	0.9430
FP: lung mass	0.8625
FP: perihilar	0.8205
FP: lobular	0.7320
FN: nodule	0.3876



(g) Lesion #15628

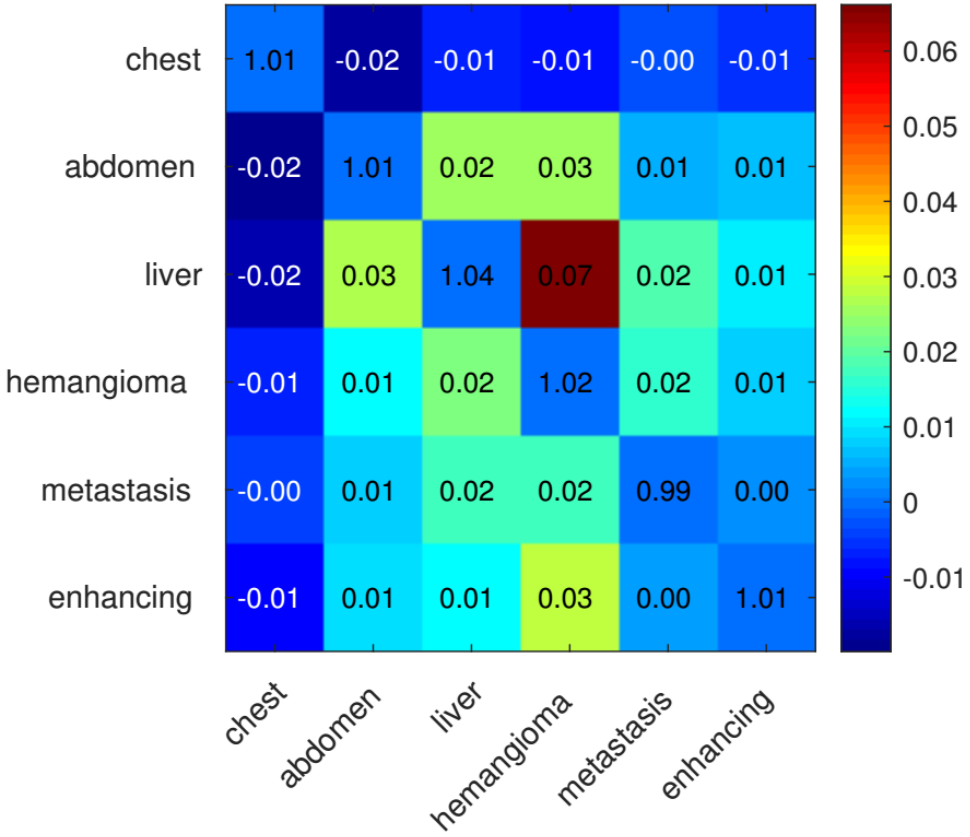
TP: liver	0.9849
TP: hemangioma	0.9508
TP: enhancing	0.9071
TP: indistinct	0.8703
FP: metastasis	0.8549
TP: hyperdense	0.8061

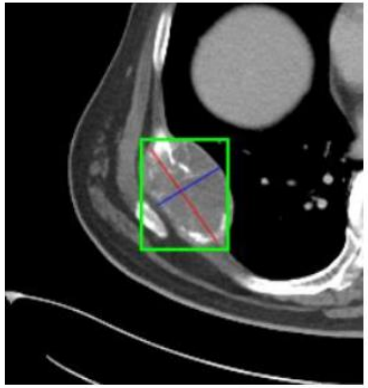


(h) Lesion #27443

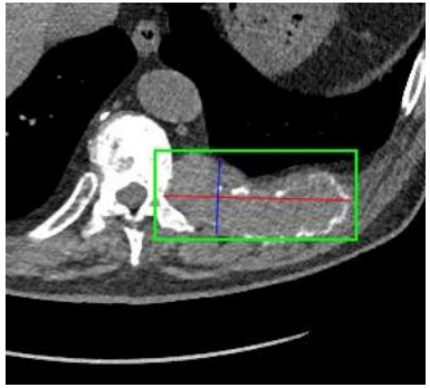
TP: liver mass	0.9151
TP: metastasis	0.8832
TP: conglomerate	0.8277
TP: lobular	0.7826
FP: indistinct	0.7699
FN: heterogeneous	0.8851
FN: large	0.8206
FN: enhancing	0.7320

Insights of the score propagation weights

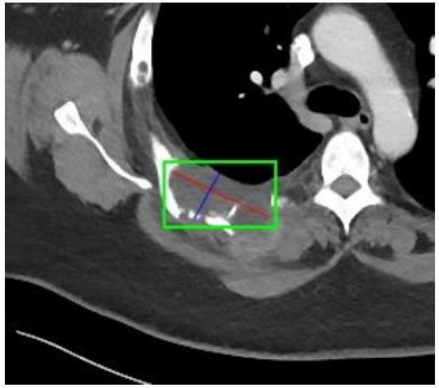




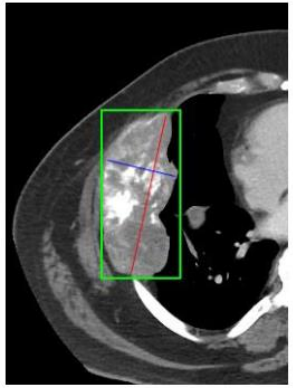
(c) Expanded **right** posterior **rib** lesion



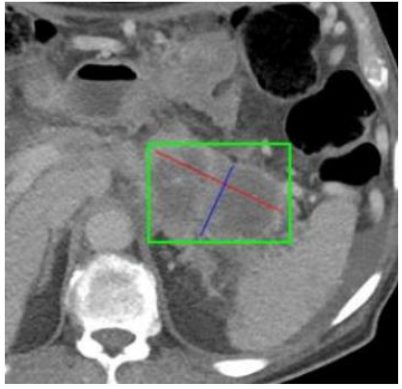
Posterior **left rib mass**



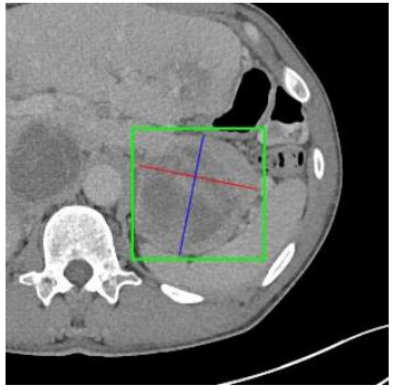
Right chest wall mass



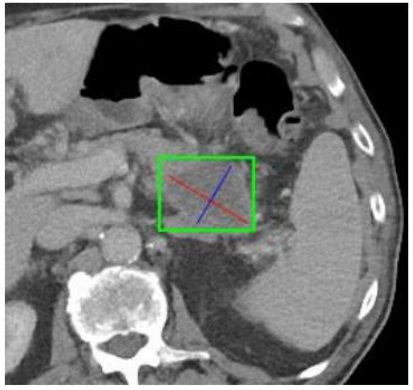
Unchanged large **right 7th rib expansile mass**



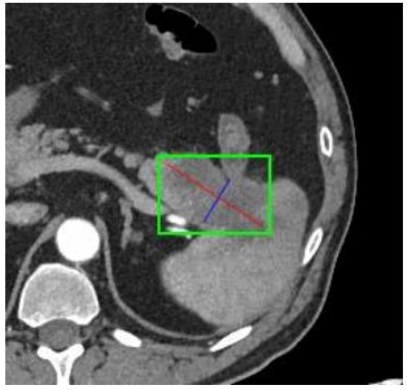
(d) Complex **retroperitoneal mass** involving the region of the **tail and body of the pancreas**



Pancreatic tail mass



Centrally **hypoattenuating mass** within the **pancreatic tail**



Low attenuation pancreatic tail mass

Summary

- Try to mine data and label from existing databases and reports
- If manual labels are not available, use weak labels to organize the data
- Leverage expert knowledge, e.g. label ontology

- Future work
  - Combining multiple lesion datasets

Thank you!

Acknowledgment

- This work was supported by the Intramural Research Program of the NIH Clinical Center and National Library of Medicine.
- We thank NVIDIA for the donation of GPUs.

Qualitative analysis

- LesaNet succeeded in predicting fine-grained body parts, lesion types, and attributes
- Errors may occur at:
  - Similar body parts and types, e.g. “left lower lobe” and “left upper lung” in (c), “hemangioma” and “metastasis” in (g)
  - Rare and/or variable labels were not learned very well, such as “conglomerate” and “necrosis” in (b)
  - Some labels may not have a clear definition, such as “mass” and “nodule” in (d)

Sentence tokenization

1. Find the "bookmark"
   - Hyperlinks (~20K)
   - Sizes and slice number references (~6K, detected using regular expressions)
2. Tokenize the sentence using NLTK
3. Use rules to fix some missing periods

FINDINGS:
Lungs, pleurae: Unchanged diffuse ground-glass opacity to the point of air bronchograms in lower lobes. **Unchanged reticular and nodular juxtapleural features for example left upper lobe BOOKMARK (1.0 cm x 0.9 cm) (series 4, image 136) and left lower lobe associated pleural thickening.**
Cardiac, Vascular: coronary, aorta, great vessels: unremarkable
Decreased lymphadenopathy for example axilla BOOKMARK (1.5 cm x 1.2 cm) (series 2, image 8)
Mediastinum: Unchanged mediastinal adenopathy
Upper abdomen: Unchanged splenomegaly BOOKMARK (15.2 cm) (series 2, image 58)
Bones, soft tissues: no evidence of suspicious sclerotic or lytic lesions

Label extraction

1. Text preprocessing on sentences
   - To lower-case, remove non-ASCII characters
   - Para aortic, para-aortic, paraaortic → paraaortic
   - Word tokenization
   - Lemmatize: plural to singular
2. Whole-word string matching based on RadLex
3. Keep 171 frequent labels (≥10 in training set and 1 in val/test set)

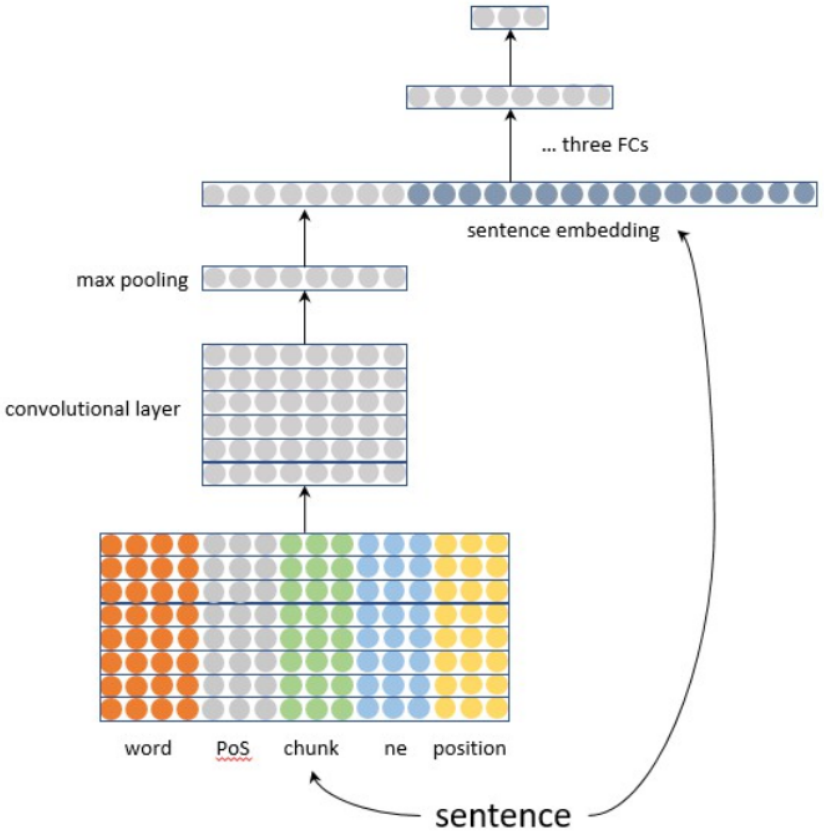


Figure 3. Framework of our CNN text-mining model.

Type	Regular Expression
Irrelevant	1. (no evidence of | no evidence of developing | no evidence of abdominal | not | poorly | previously seen | without | without evidence of) LABEL 2. (adjacent to | arising from | above | anterior to | abutting | beneath | close to | encasing | left of | left of this | near | posterior to | right of) LABEL 3. (other) LABEL
Uncertainty	1. (or | and / or | / | likely | possibly) LABEL 2. (dome of | portion of | tail of) LABEL

Table 1. Regular expressions to detect irrelevant and uncertain labels.

Method	Precision	Recall	F-score
Rule-based	**0.813**	0.507	0.545
CNN	0.788	0.783	0.784
CNN + Rule-based	0.798	**0.815**	**0.806**

Table 2. Performance of text-mining relevant labels.

Yifan Peng et al., "**A self-attention based deep learning method for lesion attribute detection from CT reports**," IEEE International Conference on Healthcare Informatics (ICHI), 2019.

Multiscale multilabel CNN

- Weighted CE loss: address imbalanced labels
  - Positive cases are sparse for most labels

$$L_{\mathrm{WCE}} = \sum_{i=1}^{B}\sum_{c=1}^{C}\left(\beta_c^{\mathrm{p}} y_{i,c}\log\sigma_{i,c} + \beta_c^{\mathrm{n}}(1-y_{i,c})\log(1-\sigma_{i,c})\right)$$

$$\beta_c^{\mathrm{p}} = |P_c + N_c|/|2P_c|, \beta_c^{\mathrm{n}} = |P_c + N_c|/|2N_c|$$

Implementation Details

- **Input**: 120mm^2 3-channel lesion image patch
- **Weighted CE loss**: clamped the weights β to be at most 300
- **RHEM**: $\gamma = 2$ and $S = 10^4$
- **Triplet loss**: $T = 5000$, loss weight $\lambda = 5$
- PyTorch, trained from scratch (BatchNorm helps)
  - Batch size 128
  - SGD lr=0.01 for 10 epochs then 0.001 for 5 epochs

Dataset

- Two radiologists further manually annotated 500 random lesions in the test set **(hand-labeled test set)**
  - Reduce missing annotations
  - In average, there are 4.2 labels per lesion in the text-mined test set, and 5.4 in the hand-labeled test set

Evaluation metric

- Per-class averaged **AUC**
- Per-class averaged **precision, recall, and F1**

Label-wise analysis

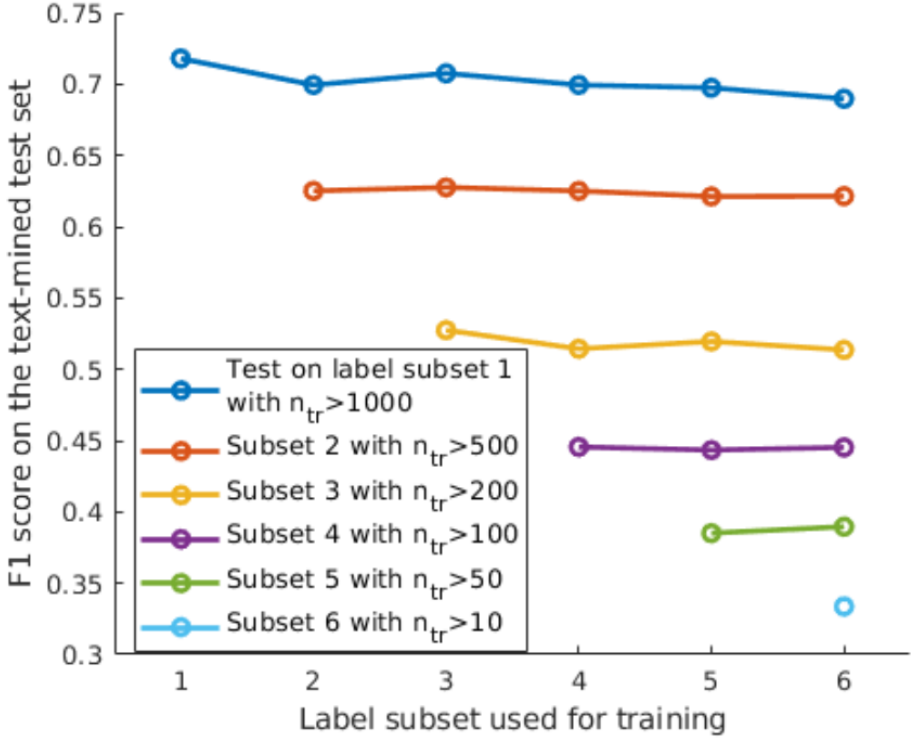
- Why F1s are low?
  - Many labels have few positive cases in the test set
  - Missing annotations

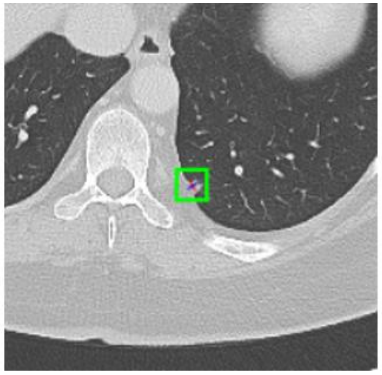
Label	AUC	F1	Label	AUC	F1
Chest	96.2	90.2	Nodule	89.1	66.9
Lung	98.6	92.0	Cyst	96.0	40.7
Liver	98.6	78.8	Adenoma	99.9	30.8
Lymph node	93.7	76.2	Metastasis	74.0	10.7
Adrenal gland	99.5	76.2	Hypodense	87.7	50.9
Right mid lung	98.7	56.6	Sclerotic	99.7	75.4
Pancreatic tail	97.5	35.3	Cavitary	94.9	25.0
Paraspinal	97.5	9.8	Large	80.6	17.5

Table 2. Accuracies (%) of typical body parts, types, and attributes.

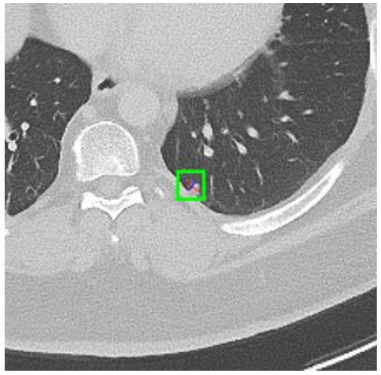
Label-wise analysis

- Is holistic learning good?
- Conclusion:
  - Learning more labels jointly does not affect accuracies of single labels significantly
  - Rare labels generally have low F1s

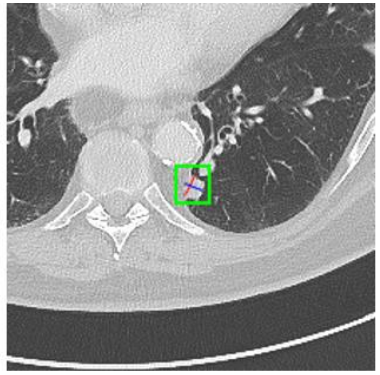




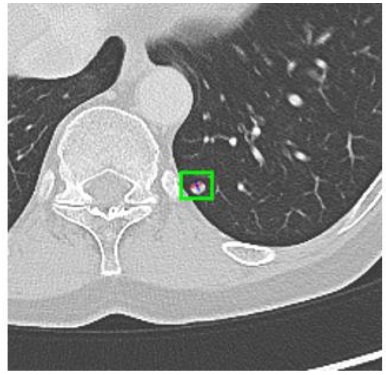
(a) Unchanged **pulmonary nodule** at the **left lower lobe**



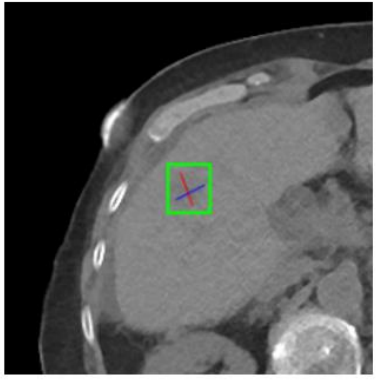
At least 2 **subcentimeter peripheral left lower lung focus**



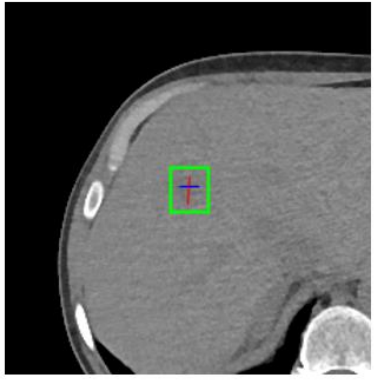
Left lower lung mass unchanged



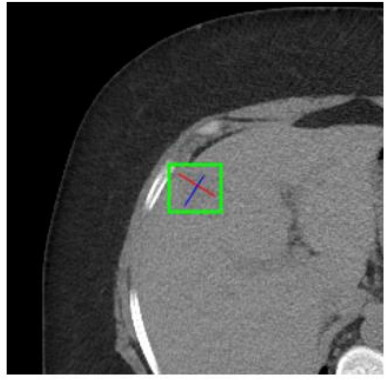
Noncalcified left lower lung mass unchanged



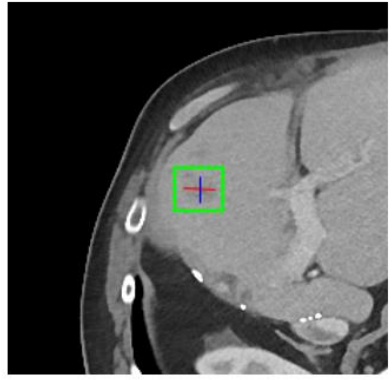
(b) Abnormality likely represent **metastasis** including focal **mass right lobe liver**



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**