Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-scale Lesion Database

Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam P. Harrison
Mohammadhadi Bagheri, Ronald M. Summers
Imaging Biomarkers and Computer-Aided Diagnosis Laboratory
National Institutes of Health Clinical Center, 10 Center Drive, Bethesda, MD 20892
{ke.yan, xiaosong.wang, ling.zhang3, mohammad.bagheri, rms}@nih.gov,
{lel, aharrison}@nvidia.com

Abstract

Radiologists in their daily work routinely find and annotate significant abnormalities on a large number of radiology images. Such abnormalities, or lesions, have collected over years and stored in hospitals' picture archiving and communication systems. However, they are basically unsorted and lack semantic annotations like type and location. In this paper, we aim to organize and explore them by learning a deep feature representation for each lesion. A large-scale and comprehensive dataset, DeepLesion, is introduced for this task. DeepLesion contains bounding boxes and size measurements of over 32K lesions. To model their similarity relationship, we leverage multiple supervision information including types, self-supervised location coordinates, and sizes. They require little manual annotation effort but describe useful attributes of the lesions. Then, a triplet network is utilized to learn lesion embeddings with a sequential sampling strategy to depict their hierarchical similarity structure. Experiments show promising qualitative and quantitative results on lesion retrieval, clustering, and classification. The learned embeddings can be further employed to build a lesion graph for various clinically useful applications. An algorithm for intra-patient lesion matching is proposed and validated with experiments.

1. Introduction

Large-scale datasets with diverse images and dense annotations [9, 12, 22] play an important role in computer vision and image understanding, but often come at the cost of vast amounts of labeling. In computer vision, this cost has spurred efforts to exploit weak labels [48, 19, 5], *e.g.*, the enormous amount of weak labels generated everyday on the web. A similar situation exists in the medical imaging domain, except that annotations are even more time consuming and require extensive clinical training, which precludes approaches like crowd-sourcing. Fortunately, like web data in computer vision, a vast, loosely-labeled, and largely untapped data source does exist in the form of hospital picture archiving and communication systems (PACS). These archives house patient images and accompanying radiological reports, markings, and measurements performed during clinical duties. However, data is typically unsorted, unorganized, and unusable in standard supervised machine learning approaches. Developing means to fully exploit PACS radiology database becomes a major goal within the field of medical imaging.

This work contributes to this goal of developing an approach to usefully mine, organize, and learn the relationships between lesions found within computed tomography (CT) images in PACS. Lesion detection, characterization, and retrieval is an important task in radiology [46, 11, 23, 21]. The latest methods based on deep learning and convolutional neural networks (CNNs) have achieved significantly better results than conventional hand-crafted image features [15, 23]. However, large amounts of training data with high quality labels are often needed. To address this challenge, we develop a system designed to exploit the routine markings and measurements of significant findings that radiologists frequently perform [10]. These archived measurements are potentially highly useful sources of data for computer-aided medical image analysis systems. However, they are basically unsorted and lack semantic labels, *e.g.*, lung nodule, mediastinal lymph node.

We take a feature embedding and similarity graph approach to address this problem. First, we present a new dataset: DeepLesion, which was collected from the PACS of a major medical institute. It contains 32,120 axial CT slices from 10,594 CT imaging studies of 4,427 unique patients. There are 1–3 lesions in each image with accompanying bounding boxes and size measurements. The lesions are diverse but unorganized. Our goal is to understand them

and discover their relationships. In other words, can we organize them so that we are able to (1) know their type and location; (2) find similar lesions in different patients, *i.e.*, content-based lesion retrieval; and (3) find similar lesions in the same patient, *i.e.*, lesion instance matching for disease tracking?

As Fig. 1 illustrates, the above problems can be addressed by learning feature representations for each lesion that keeps a proper similarity relationship, *i.e.*, lesions with similar attributes should have similar embeddings. To reduce annotation workload and leverage the intrinsic structure within CT volumes, we use three weak cues to describe each lesion: type, location, and size. Lesion types are obtained by propagating the labels of a small amount of seed samples to the entire dataset, producing pseudo-labels. The 3D relative body location is obtained from a self-supervised body-part regression algorithm. Size is directly obtained by the radiological marking. We then define the similarity relationship between lesions based on a hierarchical combination of the cues. A triplet network with a sequential sampling strategy is utilized to learn the embeddings. We also apply a multi-scale multi-crop architecture to exploit both context and detail of the lesions, as well as an iterative refinement strategy to refine the noisy lesion-type pseudo-labels.

Qualitative and quantitative experimental results demonstrate the efficacy of our framework for several highly important applications. **1)**, we show excellent performance on content-based lesion retrieval [28, 47, 41, 21]. Effective solutions to this problem can help identify similar case histories, better understand rare disorders, and ultimately improve patient care [23]. We show that our embeddings can be used to find lesions similar in type, location, and size. Most importantly, the embeddings can match lesions with semantically similar body structures that are not specified in the training labels. **2)**, the embeddings are also successfully applied in intra-patient lesion matching. Patients under therapy typically undergo CT examinations (studies) at intervals to assess the effect of treatments. Comparing lesions in follow-up studies with their corresponding ones in previous studies constitutes a major part of a radiologist's workload [25]. We provide an automated tool for lesion matching which can significantly save time, especially for patients with multiple follow-up studies [31].

2. Related work

Deep Metric Learning: Metric learning can be beneficial whenever we want to keep certain similarity relationship between samples [2]. The siamese network [3] is a seminal work in deep metric learning, which minimizes the distance between a pair of samples with the same label and pushes samples with different labels apart. It was improved by the triplet network [29], which considers relative distances. The triplet network requires three samples to compute a loss: an anchor A, a positive sample P with the same label as A, and a negative sample N with a different label. The network learns embeddings that respect the following distance relationship:

$$\|f(A) - f(P)\|_2^2 + m < \|f(A) - f(N)\|_2^2, \qquad (1)$$

where f is the embedding function to be learned and m is a predefined margin. Various improvements to the standard triplet network have been proposed [49, 36, 4, 38, 37]. Three key aspects in these methods are: how to define similarity between images, how to sample images for comparison, and how to compute the loss function. Zhang et al. [49] generalized the sampling strategy and triplet loss for multiple labels with hierarchical structures or shared attributes. Son et al. [37] employed label hierarchy to learn object embeddings for tracking, where object class is a high-level label and detection timestamp is low-level. Our sequential sampling strategy shares the similar spirit with them, but we lack well-defined supervision cues in the dataset, so we proposed strategies to leverage weak cues, *e.g.* self-supervised body-part regressor and iterative refinement.

Lesion Management: Great efforts have been devoted to lesion detection [40, 46], classification [6, 11], and retrieval [28, 47, 41, 21]. Recently, CNNs have become the method of choice over handcrafted features due to the former's superior performance [33, 39, 15, 23]. Our work is in line with content-based medical image retrieval, which has been surveyed in detail by [21]. Existing methods generally focus on one type of lesion (*e.g.* lung lesion or mammographic mass) and learn the similarity relationship based on manually annotated labels [47, 41] or radiology reports [28]. To the best of our knowledge, no work has been done on learning deep lesion embeddings on a large comprehensive dataset with weak cues. Taking a different approach, [16, 44] cluster images or lesions to discover concepts in unlabeled large-scale datasets. However, they did not leverage multiple cues to explicitly model the semantic relationship between lesions. Several existing works on intra-patient lesion matching focus on detecting follow-up lesions and matching them pixel by pixel [17, 26, 34, 43], which generally requires organ segmentation or time-consuming nonrigid volumetric registration. Besides, they are designed for certain types of lesions, whereas our lesion embedding can be used to match all kinds of lesions.

3. DeepLesion Dataset

DeepLesion dataset consists of over 32K clinically significant findings mined from a major institute's PACS. To the best of our knowledge, this dataset is the first to automatically extract lesions from challenging PACS sources. Importantly, the workflow described here can be readily scaled up and applied to multiple institutional PACS, providing a means for truly massive scales of data.

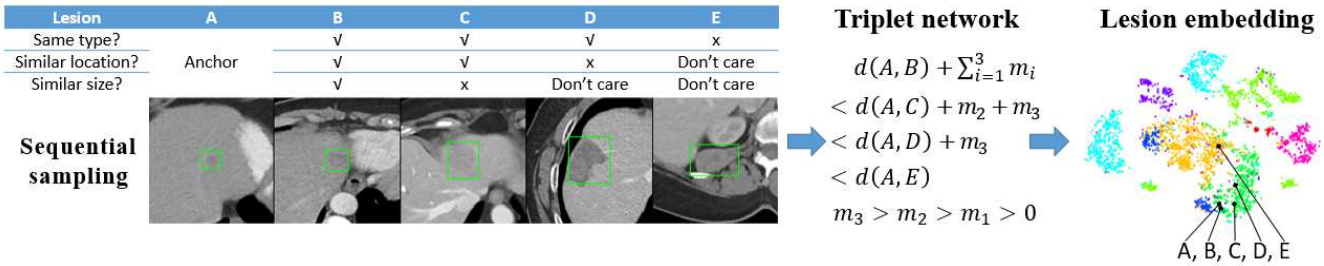

Figure 1. *The proposed framework. Using a triplet network, we learn a feature embedding for each lesion in our comprehensive DeepLesion dataset. Training samples A–E are selected with a sequential sampling strategy so as to make the embeddings respects similarity in type, location, and size.*

Radiologists routinely annotate clinically meaningful findings in medical images using arrows, lines, diameters or segmentations. These images, called "bookmarked images", have been collected over close to two decades in our institute's PACS. Without loss of generality, we study one type of bookmark in CT images: lesion diameters. As part of the RECIST guidelines [10], which is the standard in tracking lesion progression in the clinic, lesion diameters consist of two lines, one measuring the longest diameter and the second measuring its longest perpendicular diameter in the plane of measurement. We extract the lesion diameter coordinates from the PACS server and convert them into corresponding positions on the image plane. After removing some erroneous annotations, we obtain 32,120 axial CT slices (mostly 512×512) from 10,594 studies of 4,427 unique patients. There are 1–3 lesions in each image, adding up to 32,735 lesions altogether. We generate a box tightly around the two diameters and add a 5-pixel padding in each direction to capture the lesion's full spatial extent. Samples of the lesions and bounding boxes can be found in Fig. 1.

The 12-bit CT intensity range is rescaled to floating-point numbers in $[0, 255]$ using a single windowing covering the intensity ranges in lungs, soft tissues, and bones. Each image is resized so that the spacing is 1 mm/pixel. For each lesion, we crop a patch with 50 mm padding around its bounding box. To encode 3D information, we use 3 neighboring slices (interpolated at 2 mm slice intervals) to compose a 3-channel image. No data augmentation was used. DeepLesion will be publicly released. More introduction of the dataset can be found in the supplementary material.

4. Learning Lesion Embeddings

To learn lesion embeddings, we employ a triplet network with sequential sampling, as illustrated in Fig. 1. The cues used to supervise the network and the training strategy are described below.

4.1. Supervision Cues

Supervision information, or cues, are key in defining the similarity relationship between lesions. Because it is prohibitively time-consuming to manually annotate all lesions in a PACS-based dataset like DeepLesion, a different approach must be employed. Here we use the cues of lesion type, relative body location, and size. **Size information** (lengths of long and short lesion diameters) has been annotated by radiologists and ranges from 0.2 to 343 mm with a median of 15.6 mm. They are significant indicators of patients' conditions according to the RECIST guideline [10]. For example, larger lymph nodes are considered lesions while those with short diameters < 10 mm are treated as normal [10]. While size can be obtained directly from radiologists' markings, type and relative body location require more complex approaches.

Lesion Type: Among all 32,735 lesions, we randomly select 30% and manually label them into 8 types: lung, abdomen, mediastinum, liver, pelvis, soft tissue, kidney, and bone. These are coarse-scale attributes of the lesions. An experienced radiologist verified the labels. The mediastinum class mainly consists of lymph nodes in the chest. Abdomen lesions are miscellaneous ones that are not in liver or kidney. The soft tissue class contains lesions in the muscle, skin, fat, etc. Among the labeled samples, we randomly select 25% as training seeds to predict pseudo-labels, 25% as the validation set, and the other 50% as the test set. There is no patient-level overlap between all subsets.

The type of a lesion is related to its location, but the latter information cannot replace the former because some lesion types like bone and soft tissue have widespread locations. Neighboring types such as lung/mediastinum and abdomen/liver/kidney are hard to classify solely by location. The challenge with using PACS data is that there are no annotated class labels for each lesion in DeepLesion. Therefore, we use labeled seed samples to train a classifier and apply it to all unlabeled samples to get their pseudo-labels [20]. Details on the classifier are provided in Sec. 4.3.

Relative Body Location: Relative body location is an important and clinically relevant cue in lesion characterization. While the x and y coordinates of a lesion are easy to acquire in axial CT slices, the z coordinate (*e.g.* 0–1 from head to toe) is not as straightforward to find. The slice indices in the volume cannot be used to compute z because

CT volumes often have different scan ranges (start, end), not to mention variabilities in body lengths and organ layouts. For this reason, we use the self-supervised body part regressor (SSBR), which provides a relative z coordinate based on context appearance.

SSBR operates on the intuition that volumetric medical images are intrinsically structured, where the position and appearance of organs are relatively aligned. The superior-inferior slice order can be leveraged to learn an appearance-based z. SSBR randomly picks m equidistant slices from a volume, denoted $j, j+k, \ldots, j+k(m-1)$, where j and k are randomly determined. They are passed through a CNN to get a score s for each slice, which is optimized using the following loss function:

$$\begin{aligned} L_{\text{SSBR}} &= L_{\text{order}} + L_{\text{dist}}; \\ L_{\text{order}} &= -\sum_{i=0}^{m-2} \log h\left(s_{j+k(i+1)} - s_{j+ki}\right); \\ L_{\text{dist}} &= \sum_{i=0}^{m-3} g(\Delta_{i+1} - \Delta_i), \\ \Delta_i &= s_{j+k(i+1)} - s_{j+ki}, \end{aligned} \tag{2}$$

where h is the sigmoid function, g is the smooth L1 loss [14]. L_{order} requires slices with larger indices to have larger scores. L_{dist} makes the difference between two slice scores proportional to their physical distance. The order loss and distance loss terms collaborate to push each slice score towards the correct direction relative to other slices. After convergence, slices scores are normalized to $[0, 1]$ to obtain the z coordinates without having to know which score corresponds to which body part. The framework of SSBR is shown in Fig. 2.

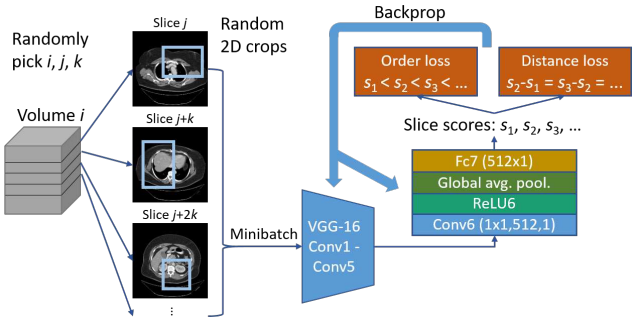

Figure 2. *Framework of the self-supervised body part regressor (SSBR).*

In DeepLesion, some CT volumes are zoomed in on a portion of the body, *e.g.* only the left half is shown. To handle them, we train SSBR on random crops of the axial slices. Besides, SSBR does not perform well on body parts that are rare in the training set, *e.g.* head and legs. Therefore, we train SSBR on all data first to detect hard volumes by examining the correlation coefficient (r) between slice indices and slice scores, where lower r often indicates rare body parts in the volume. Then, SSBR is trained again on a resampled training set with hard volumes oversampled.

4.2. Sequential Sampling

Similar to [49, 37], we leverage multiple cues to describe the relationship between samples. A naïve strategy would be to treat all cues equally, where similarity can be calculated by, for instance, averaging the similarity of each cue. Another strategy assumes a hierarchical structure exists in the cues. Some high-level cues should be given higher priority. This strategy applies to our task, because intuitively lesions of the same type should be clustered together first. Within each type, we hope lesions that are closer in location to be closer in the feature space. If two lesions are similar in both type and location, they can be further ranked by size. This is a conditional ranking scheme.

To this end, we adopt a sequential sampling strategy to select a sequence of lesions following the hierarchical relationship above. As depicted in Fig. 1, an anchor lesion A is randomly chosen first. Then, we look for lesions with similar type, location, and size with A and randomly pick B from the candidates. Likewise, C is a lesion with similar type and location but dissimilar size; D is similar in type but dissimilar in location (its size is not considered); E has a different type (its location and size are not considered). Here, two lesions are similar in type if they have the same pseudo-label; they are similar in location (size) if the Euclidean distance between their location (size) vectors is smaller than a threshold T_{low}, whereas they are dissimilar if the distance is larger than T_{high}. We do not use hard triplet mining as in [29, 27] because of the noisy cues. Fig. 3 presents some examples of lesion sequences. Note that there is label noise in the fourth row, where lesion D does not have the same type with A – C (soft tissue versus pelvis).

A selected sequence can be decomposed into three triplets: ABC, ACD and ADE. However, they are not equal, because we hope two lesions with dissimilar types to be farther apart than two with dissimilar locations, followed by size. Hence, we apply larger margins to higher-level triplets [49, 4]. Our loss function is defined as:

$$\begin{aligned} L = \frac{1}{2S} \sum_{i=1}^{S} \big[&\max(0, d_{AB}^2 - d_{AC}^2 + m_1) \\ &+ \max(0, d_{AC}^2 - d_{AD}^2 + m_2) \\ &+ \max(0, d_{AD}^2 - d_{AE}^2 + m_3) \big]. \end{aligned} \tag{3}$$

$m_3 > m_2 > m_1 > 0$ are the hierarchical margins; S is the number of sequences in each mini-batch; d_{ij} is the Euclidean distance between two samples in the embedding space. The idea in sequential sampling resembles that of SSBR (Eq. 2): ranking a series of samples to make them self-organize and move to the right place in the feature space.

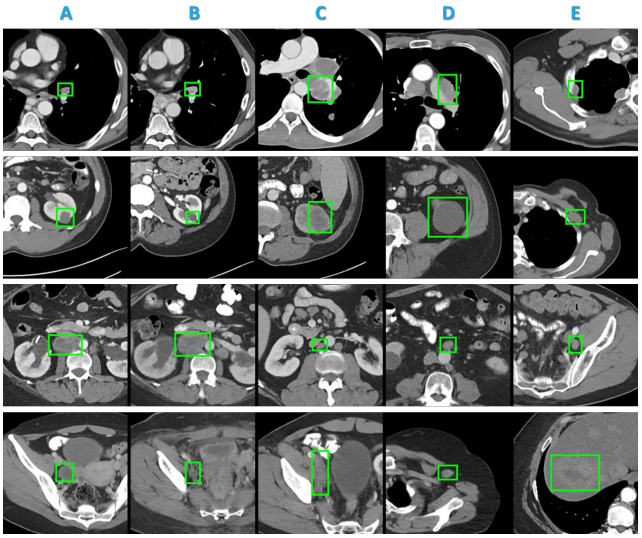

Figure 3. *Sample training sequences. Each row is a sequence. Columns 1–5 are examples of lesions A–E in Fig. 1, respectively.*

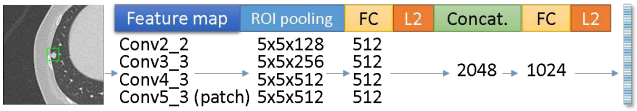

Figure 4. *Network architecture of the proposed triplet network.*

4.3. Network Architecture and Training Strategy

VGG-16 [35] is adopted as the backbone of the triplet network. As illustrated in Fig. 4, we input the 50mm-padded lesion patch, then combine feature maps from 4 stages of VGG-16 to get a multi-scale feature representation with different padding sizes [13, 18]. Because of the variable sizes of the lesions, region of interest (ROI) pooling layers [14] are used to pool the feature maps to $5 \times 5 \times num\_channel$ separately. For conv2_2, conv3_3, and conv4_3, the ROI is the bounding box of the lesion in the patch to focus on its details. For conv5_3, the ROI is the entire patch to capture the context of the lesion [13, 18]. Each pooled feature map is then passed through a fully-connected layer (FC), an L2 normalization layer (L2), and concatenated together. The final embedding is obtained after another round of FC and L2 normalization layers.

To get the initial embedding of each lesion, we use ImageNet [9] pretrained weights to initialize the convolutional layers, modify the output size of the ROI pooling layers to $1 \times 1 \times num\_channel$, and remove the FC layers in Fig. 4 to get a 1408D feature vector. We use the labeled seed samples to train an 8-class RBF-kernel support vector machine (SVM) classifier and apply it to the unlabeled training samples to get their pseudo-labels. We also tried semi-supervised classification methods [51, 1] and achieved comparable accuracy. Seed samples were not used to train the triplet network. We then sample sequences according to Sec. 4.2 and train the triplet network until convergence. With the learned embeddings, we are able to retrain the initial classifier to get cleaner pseudo-labels, then fine-tune the triplet network with a lower learning rate [44]. In our experiments, this iterative refinement improves performance.

5. Lesion Organization

The lesion graph can be constructed after the embeddings are learned. In this section, our two goals are content-based lesion retrieval and intra-patient lesion matching. The lesion graph can be used to directly tackle the first goal by finding nearest neighbors of query lesions. However, the latter one requires additional techniques to accomplish.

5.1. Intra-patient Lesion Matching

We assume that lesions in all studies have been detected by other lesion detection algorithms [40] or marked by radiologists, which is the case in DeepLesion. In this section, our goal is to match the same lesion instances and group them for each patient. Potential challenges include appearance changes between studies due to lesion growth/shrinkage, movement of organs or measurement positions, and different contrast phases. Note that for one patient not all lesions occur in each study because the scan ranges vary and radiologists only mark a few target lesions. In addition, one CT study often contains multiple series (volumes) that are scanned at the same time point but differ in image filters, contrast phases, etc. To address these challenges, we design the lesion matching algorithm described in Algo. 1.

The basic idea is to build an intra-patient lesion graph and remove the edges connecting different lesion instances. The Euclidean distance of lesion embeddings is adopted as the similarity measurement. First, lesion instances from different series within the same study are merged if their distance is smaller than T_1. They are then treated as one lesion with embeddings averaged. Second, we consider lesions in all studies of the same patient. If the distance between two lesions is larger than T_2 ($T_2 > T_1$), they are not similar and their edge is removed. After this step, one lesion in study 1 may still connect to multiple lesions in study 2 if they look similar, so we only keep the edge with the minimum distance and exclude the others. Finally, the remaining edges are used to extract the matched lesion groups.

6. Experiments

Our experiments aim to show that the learned lesion embeddings can be used to produce a semantically meaningful similarity graph for content-based lesion retrieval and intra-patient lesion matching.

Algorithm 1 *Intra-patient lesion matching*

Input: Lesions of the same patient represented by their embeddings; the study index s of each lesion; intra-study threshold T_1; inter-study threshold T_2.

Output: Matched lesion groups.

1: Compute an intra-patient lesion graph $G = \langle V, \mathcal{E} \rangle$, where V are nodes (lesions) and $\mathcal{E}$ are edges. Denote d_{ij} as the Euclidean distance between nodes i, j.
2: **Merge** nodes i and j if $s_i = s_j$ and $d_{ij} < T_1$.
3: **Threshold**: $\mathcal{E} \leftarrow \mathcal{E} - \mathcal{D}, \mathcal{D} = \{\langle i, j \rangle \in \mathcal{E} | d_{ij} > T_2\}$.
4: **Exclusion**: $\mathcal{E} \leftarrow \mathcal{E} - \mathcal{C}, \mathcal{C} = \{\langle i, j \rangle \,|\, \langle i, j \rangle \in \mathcal{E}, \langle i, k \rangle \in \mathcal{E}, s_j = s_k, \text{and } d_{ij} \geq d_{ik}\}$.
5: **Extraction**: Each node group with edge connections is considered as a matched lesion group.

6.1. Implementation Details

We empirically choose the hierarchical margins in Eq. 3 to be $m_1 = 0.1, m_2 = 0.2, m_3 = 0.4$. The maximum value of each dimension of the locations and sizes is normalized to 1. When selecting sequences for training, the similarity thresholds for location and size are $T_{\text{low}} = 0.02, T_{\text{high}} = 0.1$. We use $S = 24$ sequences per mini-batch. The network is trained using stochastic gradient descent (SGD) with a learning rate of 0.002, which is reduced to 0.0002 in iteration 2K. After convergence (generally in 3K iterations), we do iterative refinement by updating the pseudo-labels and fine-tuning the network with a learning rate of 0.0002. This refinement is performed only once because we find that more iterations only add marginal accuracy improvements. For lesion matching, the intra-study threshold T_1 is 0.1 and we vary the inter-study threshold T_2 to compute the precision-recall curve. Due to space limits, the details of SSBR are given in the supplementary material.

6.2. Content-based Lesion Retrieval

First, we qualitatively investigate the learned lesion embeddings in Fig. 5, which shows the Barnes-Hut t-SNE visualization [42] of the 1024D embedding and some sample lesions. The visualization is applied to our manually labeled test set, where we have lesion-type ground truth. As we can see, there is a clear correlation between data clusters and lesion types. It is interesting to find that some types are split into several clusters. For example, lung lesions are separated to left lung and right lung, and so are kidney lesions. Bone lesions are split into 3 small clusters, which are found to be mainly chest, abdomen, and pelvis ones, respectively. Abdomen, liver, and kidney lesions are close both in real-world location and in the feature space. These observations demonstrate the embeddings are organized by both type and location. The sample lesions in Fig. 5 are roughly similar in type, location, and size.

Fig. 6 displays several retrieval results using the lesion embeddings. They are ranked by their Euclidean distance with the query one. We find that the top results are mostly the same lesion instances of the same patient, as shown in the first row of Fig. 6. It suggests the potential of the proposed embedding on lesion matching, which will be further evaluated in the following section. To better exhibit the ability of the embedding in finding semantically similar lesions, rows 2–4 of Fig. 6 depict retrieved lesions from different patients. Spiculated nodules in the right lung and left paraaortic lymph nodes are retrieved in rows 2 and 3, respectively. Row 4 depicts lesions located on the tail of the pancreas, and also some failure cases marked in red. Note that our type labels used in supervision are too coarse to describe either abdomen lymph nodes or pancreas lesions (both are covered in the abdomen class). However, the framework naturally clusters lesions from the same body structures together due to similarity in type, location, size, and appearance, thus discovering these sub-types. Although appearance is not used as supervision information, it is intrinsically considered by the CNN-based feature extraction architecture and strengthened by the multi-scale strategy. To explicitly distinguish sub-types and enhance the semantic information in the embeddings, we can either enrich the type labels by mining knowledge from radiology reports [32, 8, 45, 50], or integrate training samples from other medical image datasets with more specialized annotations [7, 30]. These new labels may be incomplete or noisy, which fits the setting of our system.

Quantitative experimental results on lesion retrieval, clustering, and classification are listed in Table 1. For retrieval, the three supervision cues are thoroughly inspected. Because location and size (all normalized to 0–1) are continuous labels, we define an evaluation criterion called average retrieval error (ARE):

$$\text{ARE} = \frac{1}{K} \sum_{i=1}^{K} \|\mathbf{t}^Q - \mathbf{t}_i^R\|_2, \tag{4}$$

where $\mathbf{t}^Q$ is the location or size of the query lesion and $\mathbf{t}_i^R$ is that of the ith retrieved lesion among the top-K. On the other hand, the ARE of lesion type is simply $1 - precision$. Clustering and classification accuracy are evaluated only on lesion type. Purity and normalized mutual information (NMI) of clustering are defined in [24]. The multi-scale ImageNet feature is computed by replacing the 5×5 ROI pooling to 1×1 and removing the FC layers.

In Table 1, the middle part compares the results of applying different supervision information to train the triplet network. Importantly, when location and size are added as supervision cues, our network performs best on lesion-type retrieval—even better than when only lesion-type is used as the cue. This indicates that location and size provides important supplementary information in learning similarity embeddings, possibly making the embeddings more organized and acting as regularizers. The bottom part of the table shows results of ablation studies, which demon-

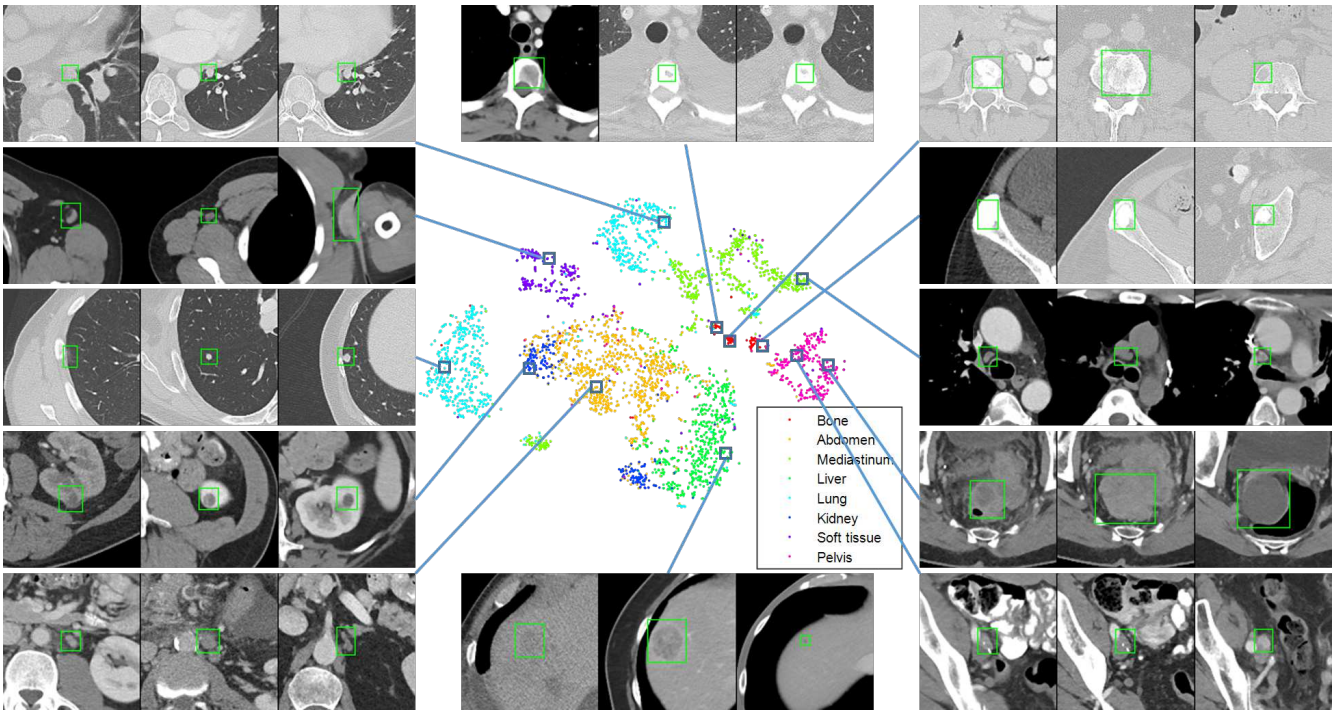

Figure 5. *t-SNE visualization of the lesion embeddings on the test set (4,927 samples) of DeepLesion. Colors indicate the manually labeled lesion types. We also split the samples to 128 clusters using K-means and show 3 random lesions in 12 representative clusters. We did not choose to show closest samples because they are very similar. Best viewed in color.*

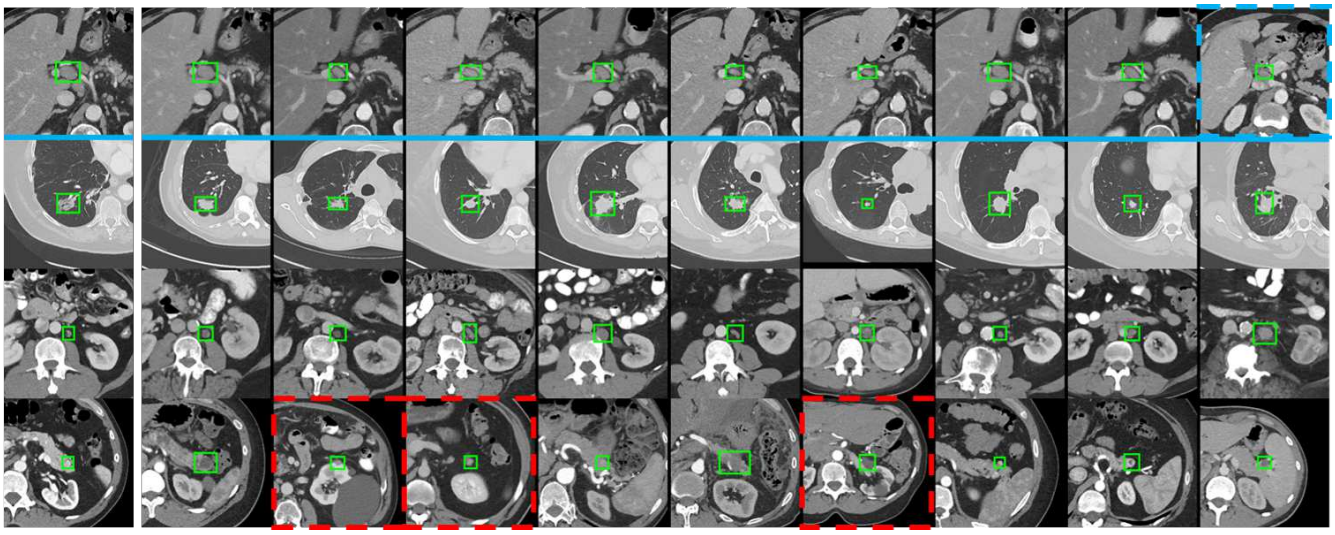

Figure 6. *Examples of query lesions (first column) and the top-9 retrieved lesions on the test set of DeepLesion. In the first row, the blue dashed box marks the lesion from a different patient than the query one, whereas the other 9 are all from the same patient. In rows 2–4, we constrain that the query and all retrieved lesions must come from different patients. Red dashed boxes indicate incorrect results, see text.*

strate the effectiveness of multi-scale features and iterative refinement, highlighting the importance of combining visual features from different context levels. When only coarse-scale features (conv5, conv4) are used, location ARE is slightly better because location mainly relies on high-level context information. However, fusing fine-level features (conv3, conv2) significantly improves type and size prediction, which indicates that details are important in these aspects. We also inspected the confusion matrix for lesion classification (Fig. 7). The most confusing types are mediastinum/lung lesions, and abdomen/liver/kidney lesions, since some of them are similar in both appearance and location. More visual results are presented in the supplementary material.

Feature representation	Average retrieval error			Clustering		Classification
	Type	Location	Size	Purity	NMI	Accuracy
Baseline: Multi-scale ImageNet feature	15.2	9.6	6.9	58.7	35.8	86.2
Baseline: Location feature	22.4	**2.5**	8.8	51.6	32.6	59.7
Triplet with type	8.8±0.2	10.8±0.2	5.7±0.1	84.7±2.8	71.5±1.7	89.5±0.3
Triplet with location	13.0±0.1	6.5±0.1	6.2±0.1	61.1±4.4	39.5±4.3	87.8±0.5
Triplet with type + location	8.7±0.2	7.2±0.1	6.0±0.1	81.3±4.7	68.0±2.4	89.9±0.3
Triplet with type + location + size	**8.5±0.1**	7.2±0.0	**5.1±0.0**	**86.0±3.9**	**72.4±4.6**	**90.5±0.2**
w/o Multi-scale feature: conv5	11.5±0.2	7.1±0.1	6.3±0.0	79.8±0.6	64.8±1.2	86.6±0.4
w/ Multi-scale feature: conv5 + conv4	9.7±0.2	7.0±0.0	5.4±0.1	82.4±3.3	67.9±2.2	89.0±0.6
w/o Iterative refinement	8.7±0.2	7.3±0.0	5.2±0.1	85.4±2.8	69.8±2.0	90.2±0.2

Table 1. *Evaluation results on the test set (4,927 samples) of DeepLesion. For retrieval, we compute the average retrieval error (%) in type, location, and size of the top-5 retrieved lesions compared to the query one. For clustering, lesions are clustered to 8 groups using K-means to calculate the purity and NMI (%). For classification, we train a 8-way softmax classifier on the seed labeled samples and apply it on the test set. The CNN in each method was trained 5 times using different random seeds. Mean results and standard deviations are reported.*

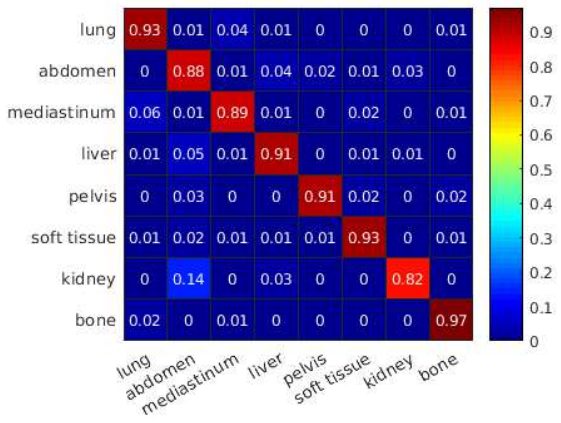

Figure 7. *The confusion matrix of lesion classification.*

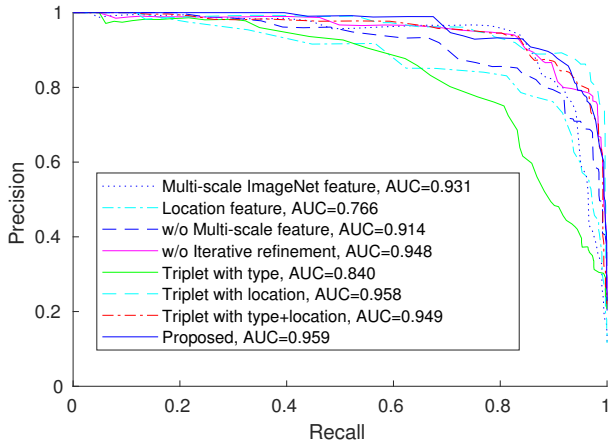

Figure 8. *Precision-recall curves of various methods on the intra-patient lesion matching task using DeepLesion. The area-under-curve (AUC) values are shown in the legends.*

6.3. Intra-patient Lesion Matching

We manually grouped 1313 lesions from 103 patients in DeepLesion to 593 groups for evaluation. Each group contains instances of the same lesion across time. There are 1–11 lesions per group. Precision and recall are defined according to [24], where a true positive decision assigns two lesions of the same instance to the same group, and a false positive decision assigns two lesions of different instances to the same group, etc. As presented in Fig. 8, our proposed embedding achieves the highest area under the curve (AUC). The location feature does not perform well because different lesion instances may be close to each other in location. This problem can be mitigated by combining location with appearance and using multi-scale features (accomplished in our triplet network). Our algorithm does not require any annotation of matched lesions for training. It is appearance-based and needs no registration or organ mask, thus is fast.

7. Conclusion and Future Work

In this paper, we present a large-scale and comprehensive dataset, DeepLesion, which contains significant radiology image findings mined from PACS. Lesion embeddings are learned with a triplet network to model their similarity relationship in type, location, and size. The only manual efforts needed are the class labels of some seed images. Promising results are obtained in content-based lesion retrieval and intra-patient lesion matching. The framework can be used as a generic lesion search engine, classifier, and matching tool. After being classified or retrieved by our system, lesions can be further processed by other specialist systems trained on data of a certain type. In the future, we plan to incorporate more fine-grained semantic information (*e.g.* from radiology reports, other specialized datasets, or active learning) to build a lesion knowledge graph.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH Clinical Center. We thank NVIDIA for the donation of GPU cards.

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006. 5

[2] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013. 2

[3] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744, 1994. 2

[4] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017. 2, 4

[5] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1431–1439, 2015. 1

[6] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Scientific Reports*, 6(1):24454, jul 2016. 2

[7] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, dec 2013. 6

[8] S. Cornegruta, R. Bakewell, S. Withey, and G. Montana. Modelling radiological language with bidirectional long short-term memory networks. *arXiv preprint arXiv:1609.08409*, 2016. 6

[9] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, jun 2009. 1, 5

[10] E. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, and Others. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009. 1, 3

[11] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. 1, 2

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[13] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1134–1142, 2015. 5

[14] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4, 5

[15] H. Greenspan, B. van Ginneken, and R. M. Summers. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, may 2016. 1, 2

[16] J. Hofmanninger, M. Krenn, M. Holzer, T. Schlegl, H. Prosch, and G. Langs. Unsupervised identification of clinically relevant clusters in routine imaging data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 192–200. Springer, 2016. 2

[17] H. Hong, J. Lee, and Y. Yim. Automatic lung nodule matching on sequential CT images. *Computers in Biology and Medicine*, 38(5):623–634, may 2008. 2

[18] P. Hu and D. Ramanan. Finding Tiny Faces. In *CVPR*, 2017. 5

[19] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, pages 301–320. Springer, 2016. 1

[20] D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013. 3

[21] Z. Li, X. Zhang, H. Müller, and S. Zhang. Large-scale retrieval for medical image analytics: A comprehensive review. *Medical Image Analysis*, 43:66–84, jan 2018. 1, 2

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[23] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, dec 2017. 1, 2

[24] C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to information retrieval*. Cambridge University Press, 2008. 6, 8

[25] J. H. Moltz, M. D'Anastasi, A. Kießling, D. P. Dos Santos, C. Schülke, and H.-O. Peitgen. Workflow-centred evaluation of an automatic lesion tracking software for chemotherapy monitoring by CT. *European radiology*, 22(12):2759–2767, 2012. 2

[26] J. H. Moltz, M. Schwier, and H.-O. Peitgen. A general framework for automatic detection of matching lesions in follow-up ct. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 843–846. IEEE, 2009. 2

[27] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016. 4

[28] J. Ramos, T. T. J. P. Kockelkorn, I. Ramos, R. Ramos, J. Grutters, M. A. Viergever, B. van Ginneken, and A. Campilho. Content-Based Image Retrieval by Metric Learning From Radiology Reports: Application to Interstitial Lung Dis-

eases. *IEEE Journal of Biomedical and Health Informatics*, 20(1):281–292, jan 2016. 2

[29] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 2, 4

[30] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical Image Analysis*, 42:1–13, 2017. 6

[31] M. Sevenster, A. R. Travis, R. K. Ganesh, P. Liu, U. Kose, J. Peters, and P. J. Chang. Improved efficiency in clinical workflow of reporting measured oncology lesions via pacs-integrated lesion tracking tool. *American Journal of Roentgenology*, 204(3):576–583, 2015. 2

[32] H.-C. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. Summers. Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation. *Journal of Machine Learning Research*, 17(1-31):2, 2016. 6

[33] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, may 2016. 2

[34] J. S. Silva, J. Cancela, and L. Teixeira. Fast volumetric registration method for tumor follow-up in pulmonary ct exams. *Journal of Applied Clinical Medical Physics*, 12(2):362–375, 2011. 2

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR 2015*, 2015. 5

[36] K. Sohn. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Neural Information Processing Systems*, pages 1–9, 2016. 2

[37] J. Son, M. Baek, M. Cho, and B. Han. Multi-Object Tracking with Quadruplet Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5620–5629, 2017. 2, 4

[38] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy. Deep metric learning via facility location. In *IEEE CVPR*, 2017. 2

[39] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, may 2016. 2

[40] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki. Automated detection of pulmonary nodules in pet/ct images: Ensemble false-positive reduction using a convolutional neural network technique. *Medical physics*, 43(6):2821–2827, 2016. 2, 5

[41] L. Tsochatzidis, K. Zagoris, N. Arikidis, A. Karahaliou, L. Costaridou, and I. Pratikakis. Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach. *Pattern Recognition*, 2017. 2

[42] L. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014. 6

[43] R. Vivanti. Automatic liver tumor segmentation in follow-up ct studies using convolutional neural networks. In *Proc. Patch-Based Methods in Medical Image Processing Workshop*, 2015. 2

[44] X. Wang, L. Lu, H.-C. Shin, L. Kim, M. Bagheri, I. Nogues, J. Yao, and R. M. Summers. Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 998–1007. IEEE, 2017. 2, 5

[45] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *CVPR*, may 2017. 6

[46] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang. *Zoom-in-Net: Deep Mining Lesions for Diabetic Retinopathy Detection*, pages 267–275. Springer International Publishing, 2017. 1, 2

[47] G. Wei, H. Ma, W. Qian, and M. Qiu. Similarity measurement of lung masses for medical image retrieval using kernel based semisupervised distance metric. *Medical Physics*, 43(12):6259–6269, nov 2016. 2

[48] H. Zhang, X. Shang, W. Yang, H. Xu, H. Luan, and T.-S. Chua. Online collaborative learning for open-vocabulary visual classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2809–2817, 2016. 1

[49] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. Embedding label structures for fine-grained feature representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1114–1123, 2016. 2, 4

[50] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. In *CVPR*, 2017. 6

[51] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU-CALD-02-107, Carnegie Mellon University*, 2002. 5