The Dissertation Committee for Yan Han
certifies that this is the approved version of the following dissertation:

Integrating Domain Knowledge and Deep Learning for Enhanced Chest X-ray Diagnosis and Localization

Committee:

Ahmed Tewfik, Co-Supervisor

Ying Ding, Co-Supervisor

Alan Bovik

Marinka Zitnik

Zhangyang (Atlas) Wang

Integrating Domain Knowledge and Deep Learning for Enhanced Chest X-ray Diagnosis and Localization

by

Yan Han

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2023

Dedicated to my family.

Acknowledgments

First and foremost, I want to express my deepest gratitude to my three supervisors, Prof. Ahmed Tewfik, Prof. Ying Ding, and Prof. Atlas Wang. Prof. Tewfik generously granted me the freedom to explore my research interests and cultivated an environment that fostered the development of my skills and independence as a researcher. His trust and guidance have been invaluable.

I am profoundly grateful to Prof. Ying Ding and Prof. Atlas Wang, who have been instrumental in shaping my growth as a researcher. Their vast knowledge, unwavering support, and persistent encouragement have been crucial to my success. Prof. Ding and Prof. Wang have not only provided intellectual guidance but have also served as true mentors. They have always been there to listen, offer advice, and inspire me to overcome challenges. Their dedication to my progress and well-being has left an indelible mark on my academic and personal life. I am honored to have had the opportunity to learn from and collaborate with such remarkable and inspiring individuals.

I would like to extend my heartfelt appreciation to my dissertation committee members, Prof. Marinka Zitnik and Prof. Alan Bovik, for their time, insightful feedback, and stimulating questions that broadened my way of thinking.

To my exceptional labmates and friends, Chongyan Chen, Greg Holste, Song Wang, Liyan Tang, Yi Wang, Xinyu Gong, Zhiwen Fan, Peihao Wang, Haotao Wang, Wenqing Zheng, Ziqi Ke, and Yitao Chen: your camaraderie and support have been indispensable and vital to my Ph.D. journey. I will always treasure the memories we created together.

I also want to express my gratitude to my mentors Ji Yan and Shan Li during my internship at LinkedIn, as well as Dr. Atlas Wang, Dr. Eddie Huang, and Dr. Nikhil Rao during my internship at Amazon. Their kindness, guidance, and support have significantly contributed to my professional growth.

Special thanks go to my parents for giving me life and their unwavering love, understanding, and support throughout this journey. Their faith in me and their constant encouragement have been a steady source of strength and motivation.

Thank you all for being essential parts of my academic journey and for your priceless contributions to my growth as a researcher.

Integrating Domain Knowledge and Deep Learning for Enhanced Chest X-ray Diagnosis and Localization

Publication No. _____

Yan Han, Ph.D.
The University of Texas at Austin, 2023

Supervisors: Ahmed Tewfik
Ying Ding

Chest X-ray imaging has become increasingly crucial for diagnosing various medical conditions, including pneumonia, lung cancer, and heart diseases. Despite the growing number of chest X-ray images, their interpretation remains a manual and time-consuming process, often leading to radiologist burnout and delays in diagnosis. The integration of domain knowledge and deep learning techniques has the potential to improve diagnosis, classification, and localization of abnormalities in chest X-rays, while also addressing the challenge of model interpretability.

This work proposes a series of novel methods combining radiomics features and deep learning techniques for chest X-ray diagnosis, classification, and localization. We first introduce a framework leveraging radiomics features and contrastive learning for pneumonia detection, achieving superior performance

and interpretability. The second method, ChexRadiNet, utilizes radiomics features and a lightweight triplet-attention mechanism for enhanced abnormality classification performance.

In addition, we present a semi-supervised knowledge-augmented contrastive learning framework that seamlessly integrates radiomic features as a knowledge augmentation for disease classification and localization. This approach leverages Grad-CAM to highlight crucial abnormal regions, extracting radiomic features that act as positive samples for image features generated from the same chest X-ray. Consequently, this framework creates a feedback loop, enabling image and radiomic features to mutually reinforce each other, resulting in robust and interpretable knowledge-augmented representations.

The Radiomics-Guided Transformer (RGT) fuses global image information with local radiomics-guided auxiliary information for accurate cardiopulmonary pathology localization and classification without bounding box annotations.

Experimental results on public datasets such as NIH ChestX-ray, CheXpert, MIMIC-CXR, and the RSNA Pneumonia Detection Challenge demonstrate the effectiveness of our proposed methods, consistently outperforming state-of-the-art models in chest X-ray diagnosis, classification, and localization tasks. By bridging the gap between traditional radiomics and deep learning approaches, this work aims to advance the field of medical image analysis and facilitate more efficient and accurate diagnoses in clinical practice.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Chest X-ray imaging has become an essential diagnostic tool in modern medicine due to its noninvasive nature and ability to visualize various medical conditions, including pneumonia, lung cancer, and heart diseases. However, despite the growing number of chest X-ray images, their interpretation remains a manual and time-consuming process. Radiologists face significant challenges in diagnosing and localizing abnormalities from these images, which often leads to burnout and delays in patient care. Consequently, there is a pressing need for automated and accurate methods to assist radiologists in interpreting chest X-ray images.

The integration of domain knowledge and deep learning techniques has the potential to improve the diagnosis, classification, and localization of abnormalities in chest X-rays while also addressing the challenge of model interpretability. In recent years, radiomics, a subfield of radiology that focuses on extracting quantitative features from medical images, has demonstrated its potential to facilitate medical imaging diagnosis. The rise of deep learning has further enhanced the ability to analyze chest X-ray images. However, the explainability of deep learning models often remains opaque, making it difficult

for medical professionals to trust and adopt these models in clinical practice.

In this work, we propose a series of novel methods that combine radiomics features and deep learning techniques for chest X-ray diagnosis, classification, and localization. These methods aim to provide accurate and interpretable results while minimizing the reliance on manually annotated labels and pixel regions. Our first method, a framework leveraging radiomics features and contrastive learning, detects pneumonia in chest X-rays with superior performance and interpretability. The second method, ChexRadiNet, uses radiomics features and a lightweight triplet-attention mechanism to enhance abnormality classification performance.

Furthermore, we present a semi-supervised knowledge-augmented contrastive learning framework that seamlessly integrates radiomic features as a knowledge augmentation for disease classification and localization. By leveraging Grad-CAM to highlight crucial abnormal regions and extracting radiomic features, this framework creates a feedback loop that enables image and radiomic features to mutually reinforce each other, yielding robust and interpretable knowledge-augmented representations.

The Radiomics-Guided Transformer (RGT) fuses global image information with local radiomics-guided auxiliary information for accurate cardiopulmonary pathology localization and classification without bounding box annotations.

Experimental results on public datasets such as NIH ChestX-ray, CheX-

pert, MIMIC-CXR, and the RSNA Pneumonia Detection Challenge demonstrate the effectiveness of our proposed methods, consistently outperforming state-of-the-art models in chest X-ray diagnosis, classification, and localization tasks. By bridging the gap between traditional radiomics and deep learning approaches, this work aims to advance the field of medical image analysis and facilitate more efficient and accurate diagnoses in clinical practice.

Chapter 2

Pneumonia Detection on Chest X-ray using Radiomic Features and Contrastive Learning

2.1 Introduction

Pneumonia is the leading cause of people hospitalized in the US [5]. It requires timely and accurate diagnosis for immediate treatment. As one of the most ubiquitous diagnostic imaging tests in medical practice, chest X-ray plays a crucial role in pneumonia diagnosis in clinical care and epidemiological studies [6]. However, rapid pneumonia detection in chest X-rays is not always available, particularly in the low-resource settings where there are not enough trained radiologists to interpret chest X-rays. There is, therefore, a critical need to develop an automated, fast, and reliable method to detect pneumonia on chest X-rays.

With the great success of deep learning in various fields, deep neural networks (DNNs) have proven to be powerful tools that can detect pneumonia to augment radiologists [7, 8, 9, 10]. However, most of the DNNs lacks explainability due to their black-box nature. Thus researchers still have a limited understanding of DNNs' decision-making process.

One method of increasing the explainability of DNNs in chest radio-

graphs is to leverage radiomics. Radiomics is a novel feature transformation method for detecting clinically relevant features from radiological imaging data that are difficult for the human eye to perceive. It has proven to be a highly explainable and robust technique because it is related to a specific region of interest (ROI) of the chest X-rays [11]. However, directly combining radiomic features and medical image hidden features provides only marginal benefits, a result mostly due to the lack of correlations at a "mid-level"; it can be challenging to relate raw pixels to radiomic features. In efforts to make more efficient use of multimodal data, several recent studies have shown promising results from contrastive representation learning [12, 13]. But, to the best of our knowledge, no studies have exploited the naturally occurring pairing of images and radiomic data.

In this study, we proposed a framework that leverages radiomic features and contrastive learning to detect pneumonia in chest X-ray. Our framework improves chest x-ray representations by maximizing the agreement between true image-radiomics pairs versus random pairs via a bidirectional contrastive objective between the image and human-crafted radiomic features. Experiments on the RSNA Pneumonia Detection Challenge dataset [14] show that our methods can fully utilize unlabeled data, provide a more accurate pneumonia diagnosis, and remedy the black-box's transparency.

Our contribution in this chapter is three-fold: (1) We introduce a framework for pneumonia detection that combines the expert radiographic knowledge (radiomic features) with deep learning. (2) We improve chest X-ray

representations by exploring the use of contrastive learning. Our model thus has the advantages of utilizing the paired radiomic features requiring no additional radiologist input. (3) We find that our models significantly outperform baselines in pneumonia detection with improved model explainability.

2.2 Method

Inspired by recent contrastive learning algorithms [12], our model learns representations by maximizing agreement between radiomics features related to pneumonia ROI of the chest X-rays and the image features extracted by the attention-based convolutional neural network (CNN) model, via a contrastive loss in the latent space. Since radiomics can be considered as the quantified prior knowledge of radiologists, we deem that our model is more interpretable than others. As illustrated in Figure 2.1, our framework consists of three phases: contrastive training, supervised fine-tuning, and testing.

Contrastive training. The model is given two inputs, x_u and x_v. x_u is the original chest X-rays without a corresponding paired bounding box. x_v is the original chest X-rays with an additional paired bounding box. For normal chest X-rays, we take the whole image as a bounding box.

For x_u, we utilize the pre-trained attention-based CNN models, Residual Attention Network (ResNet-18Attention) [15] pre-trained on CIFAR-10 [16], as the backbone of the network. We replace the last fully-connected layer with a multilayer perceptron (MLP) to generate a 128-dimensional image fea-

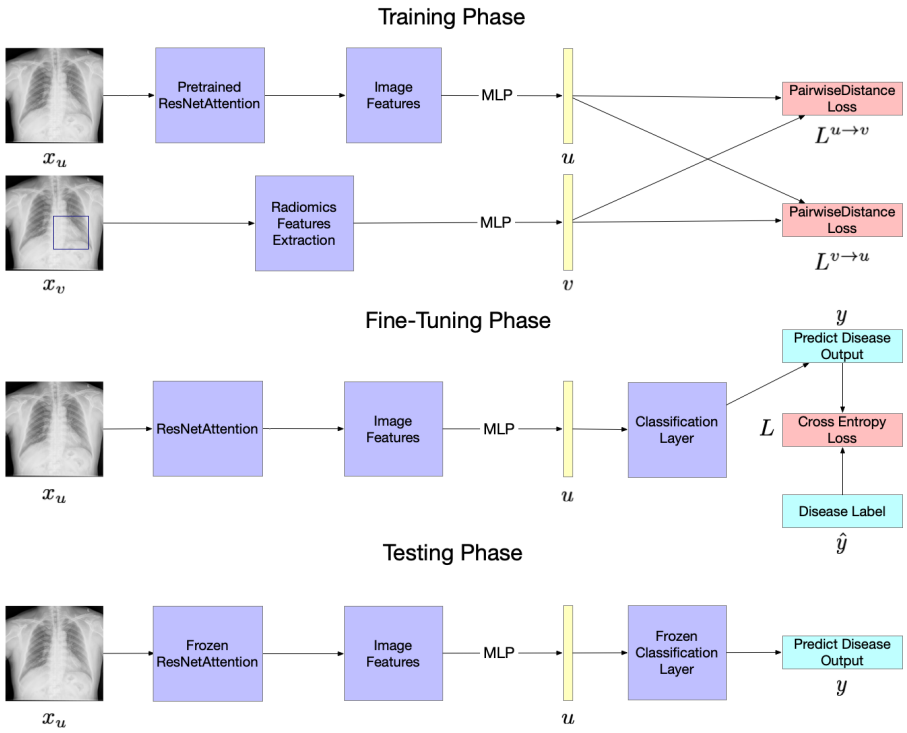

Figure 2.1: An overview of the proposed model.

tures vector u. For x_v, we apply the PyRadiomics[1] to extract 102-dimensional quantitative features, and [17] showed the details of these quantitative features and extraction process. We then use an MLP to map the features to a 128-dimensional radiomics feature v.

At each epoch of training, we sample a mini-batch of N input pairs (X_u, X_v) from the training data, and calculate their image features and radiomics features pairs (U, V). We use (u_i, v_i) to denote the ith pair. The training loss function will be divided into two parts. The first part is a con-

[1] `https://pyradiomics.readthedocs.io/en/latest/`

trastive image-to-radiomics loss:

$$L_i^{u \to v} = -\log \frac{\exp(< u_i, v_i > /\tau)}{\sum_{k=1}^{N} \exp(< u_i, v_k > /\tau)} \tag{2.1}$$

where $< u_i, v_i >$ represents the pairwise distance, i.e. $[\sum (u_i - v_i)^p]^{\frac{1}{p}}$ and p represents the norm degree, e.g., $p = 1$ and $p = 2$ represent the Taxicab norm and Euclidean norm, respectively; and $\tau \in \mathbb{R}^+$ represents a temperature parameter. In our model, we set p to 2 and τ to 0.1. Like previous work [12], which uses a contrastive loss between inputs of the different modalities, our image-to-radiomics contrastive loss is also asymmetric for each input modality. We thus define a similar radiomics-to-image contrastive loss as:

$$L_i^{v \to u} = -\log \frac{\exp(< v_i, u_i > /\tau)}{\sum_{k=1}^{N} \exp(< v_i, u_k > /\tau)} \tag{2.2}$$

Our final loss is then computed as a weighted combination of the two losses averaged over all pairs in each minibatch where $\lambda \in [0, 1]$ is a scalar weight

$$L_{train} = \frac{1}{N} \sum_{i=1}^{N} (\lambda L_i^{u \to v} + (1 - \lambda) L_i^{v \to u}) \tag{2.3}$$

Supervised fine-tuning. We follow the work of Zhang et al. [12] by fine-tuning both the CNN weights and the MLP blocks together, which closely resembles how the pre-trained CNN weights are used in practical applications. In this process, the loss function is the cross-entropy loss where $\hat{y}$ and y represent the true and predicted disease label, respectively.:

$$L_{fine-tune} = -(\hat{y} \log y + (1 - \hat{y}) \log(1 - y)) \tag{2.4}$$

Testing. The model is only given one input, the original chest X-rays x_u without a corresponding paired bounding box. Image features are extracted then mapped into the 128-dimensional feature representation u. Finally, the predicted output is calculated based on u.

2.3 Experiments

2.3.1 Dataset and Experimental Settings

To evaluate the performance of our proposed model, we conducted experiments on a public Kaggle dataset: RSNA Pneumonia Detection Challenge[2]. It contains 30,227 frontal-view images, out of which 9,783 images has pneumonia with a corresponding bounding box. We used 75% imaged for training and fine-tuning and 25% for testing.

We used SGD as our optimizer and set the initial learning rate as 0.1. We iterated the training and fine-tuning process for 200 epochs with batch size 64 and early stooped if the loss did not decrease. We reported accuracy, F1 score, and the area under the receiver operating characteristic curve (AUC).

2.3.2 Results

We compared four models: (1) ResNet-18, (2) ResNet-18 with radiomics features (ResNet-18Radi), (3) ResNet-18 with the attention mechanism (ResNet-18Att), and (4) ResNet-18Attention with radiomics features (ResNet-18AttRadi).

[2] `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data`

Experimental results are shown in Table 2.1. Compared with the baseline models (ResNet-18 and ResNet-18Att), our radiomics-based models (ResNet-18Radi and ResNet-18AttRadi) achieved better performance on the pneumonia/normal binary classification task. It suggests that radiomic features can provide additional strengths over the image features extracted by the CNN model. Compared ResNet-18Att with ResNet-18 and ResNet-18AttRadi with ResNet-18Radi, we observed that the attention mechanism could effectively boost the classification accuracy. It proves our hypothesis that pneumonia is often related to some specific ROI of chest X-rays. Hence, the attention mechanism makes it easier for the CNN model to focus on those regions.

Table 2.1: Experimental results

Model	**Accuracy**	**F1 score**	**AUC**
ResNet-18	0.763	0.782	0.795
ResNet-18Att	0.815	0.826	0.848
ResNet-18Radi	0.851	0.901	0.898
ResNet-18AttRadi	**0.886**	**0.927**	**0.923**

Figure 2.2 shows the training and fine-tuning loss convergence for the ResNet-18AttRadi model on the training set. We find that the loss drops rapidly during the pre-training stage within just a few epochs, revealing that contrastive learning makes the model learn to extract image features fast and effectively.

To fairly evaluate the impact of radiomics features on ROI, we conducted additional experiments using the whole image as a bounding box to

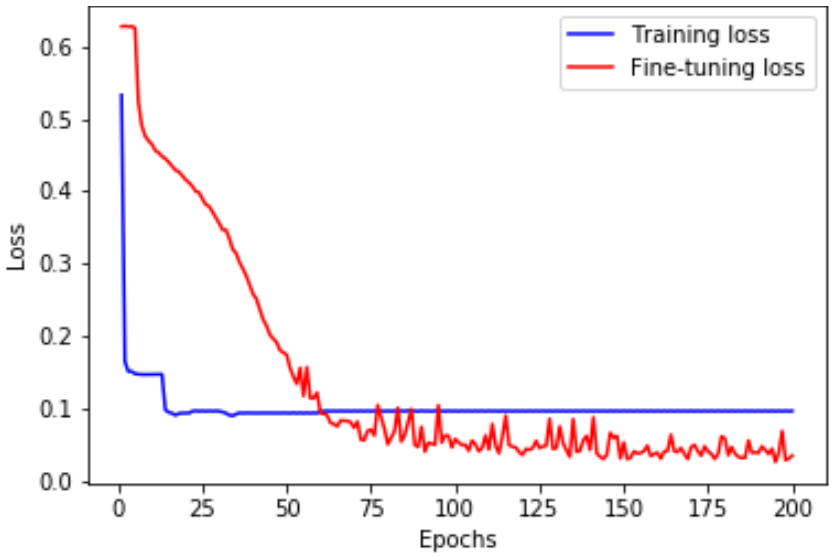

Figure 2.2: The training and fine-tuning loss convergence for the ResNet-18AttRadi model.

extract the radiomics features, denoted as ResNet-18FairRadi and ResNet-18AttFairRadi. Table 2.2 shows that even if without ROI, the radiomics features could improve the performance of the deep learning model by 5% in F1 score. This observation further demonstrates that combining radiomics features with a deep learning model for reading chest X-rays is necessary.

Table 2.2: Experimental Results Without Using Bounding Box

Model	**Accuracy**	**F1 score**	**AUC**
ResNet-18	0.763	0.782	0.795
ResNet-18FairRadi	0.821	0.841	0.864
ResNet-18Att	0.815	0.826	0.848
ResNet-18AttFairRadi	**0.854**	**0.884**	**0.877**

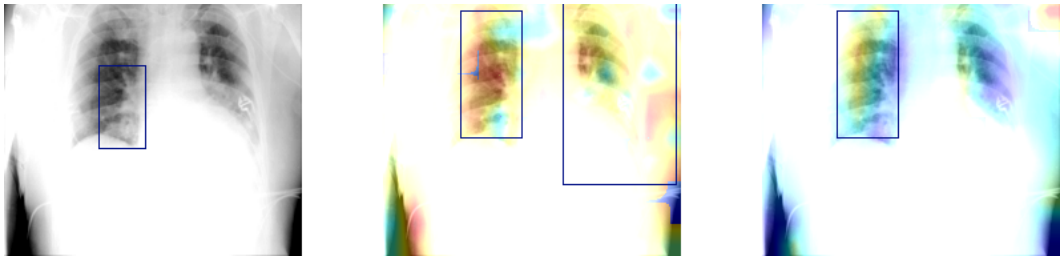

Figure 2.3: An example of visualization of attention maps. The left figure is the original **Pneumonia** chest X-ray with a bounding box. The right two figures are the attention maps of the final attention layer ResNet-18Att and ResNet-18AttRadi, respectively.

2.3.3 Visualization of the deep learning model

To demonstrate the interpretability of our model, we show some selected examples of model visualization, i.e., attention maps of ResNet-18Att and ResNet-18AttRadi. Figure 2.3 shows the original chest X-ray with a bounding box, attention map of the final attention layer of the ResNet-18Att and ResNet-18AttRadi, respectively. These examples suggest that our ResNet-18AttRadi model can focus on a more accurate area of the chest X-ray while ResNet-18Att attends to almost the whole image and contains plenty of attention noise. This illustrates that contrastive learning can help the model learn from radiomics features related to certain ROIs and thus attend more to the correct regions. And more examples of the attention maps can be found in the supplemental material.

2.4 Conclusion

In this chapter, we present a novel framework by combining radiomic features and contrastive learning to detect pneumonia from chest X-ray. Experimental results showed that our proposed models could achieve superior performance to baselines. We also observed that our model could benefit from the attention mechanism to highlight the ROI of chest X-rays. There are two limitations to this method. First, we evaluated our framework on one deep learning model (ResNet). We plan to assess the effect of radiomic features on other DNNs in the future. Second, our model relies on bounding box annotations during the training phase. We plan to leverage weakly supervised learning to automatically generate bounding boxes on large-scale datasets to ease the expert annotating process. In addition, we will compare contrastive learning with multitask learning to further exploit the integration of radiomics with deep learning. While our method only scratches the surface of contrastive learning using radiomics knowledge in the medical domain, we hope it will shed light on the development of explainable models that can efficiently use domain knowledge for medical image understanding.

Chapter 3

Using Radiomics as Prior Knowledge for Thorax Disease Classification and Localization in Chest X-rays

3.1 Introduction

The chest X-ray is one of the most common medical procedures for diagnosis, but the interpretation of chest x-ray images is subject to significant diagnosis variability for important clinical decisions. A radiologist reads about 20,000 images a year, roughly 50-100 per day, and the number is increasing. Each year, the US produces 600 billion images, and 31% of American radiologists have experienced at least one malpractice claim, often missed diagnoses [18]. The shortage of radiologists and burnout of physicians creates an urgent demand for immediate solutions. Building automatic or semi-automatic approaches to medical imaging diagnosis becomes an unavoidable next step.

The recent development of artificial intelligence, especially deep learning, offers great potential to improve medical imaging diagnosis [19]. It also sneaks into the radiology reading rooms to build a new paradigm for precision diagnosis [1, 20, 21]. Pioneering work on chest X-rays mainly focused on two problems: disease classification and localization. The recent release of large-scale datasets, such as NIH Chest X-ray [20], CheXpert [22], and MIMIC-CXR

[23], have enabled many studies using deep learning for automated chest X-ray diagnosis, such as thorax disease classification [1, 24, 25, 26] and localization [20, 27, 28].

In practice, radiologists use pattern recognition on medical images to make a diagnostic decision [29]. The knowledge of radiologists can be captured by Radiomics, which has demonstrated the effectiveness of image-based biomarkers for cancer staging and prognostication. Formally, radiomics extracts quantitative data from medical images to represent tumor phenotypes, such as spatial heterogeneity of a tumor and spatial response variations. It plays an important role in precision medicine to support evidence-based clinical decision-making. For example, radiomics can generate the detailed quantification of tumor phenotype [30] and acts as a radiographic imaging phenotype which is associated with tumor stage, metabolism, and gene or protein expression profiles [31, 32].

While radiomics offer the potential for more precise and accurate clinical predictions, it is surprising that radiomics has not been implemented in the layers of the neural networks, nor to the best of our knowledge in the deep learning workflow for X-ray analysis [33, 34]. To bridge this gap, in this chapter, we propose ChexRadiNet, a new framework that incorporates domain-specific knowledge (radiomics) into deep learning algorithms as soft constraints, and then learns end-to-end to automatically detect thorax diseases and generate bounding boxes on chest X-rays. Compared with previous studies, our proposed model does not need pre-annotated bounding boxes for

training and can achieve state-of-the-art performance for thorax disease localization. Therefore, it provides a way to introduce prior information about anticipated explanations, a technique that is widely used in the "Rationale model" [35].

For ensuring ChexRadiNet is robust and generalizable, three public benchmarking datasets were used for this purpose: NIH Chest X-ray [20], CheXpert [22], and MIMIC-CXR [23]. We demonstrate that our model outperforms baseline methods for both thorax disease classification and localization.

3.2 Method

Figure 3.1 shows our proposed ChexRadiNet, which consists of two branches. The first branch predicts whether the pathology is present or not in the image. The second branch localizes its regions using the radiomic features extracted from the first branch. ChexRadiNet utilizes a multi-task, closed-loop strategy to learn and use radiomic features as soft constraints. Formally, we are learning a two-part latent-variable model of the form $E_{z\sim p(z|x)}p(y|x,z)$, where the latent z is a radiomic-based mask over the image x with the probability $p(z|x)$. $p(y|x,z)$ is a masked version of the classification framework. Therefore, we consider the training process as a weakly-supervised learning. In this section, we first illustrate the architecture of ChexRadiNet and then present the training process.

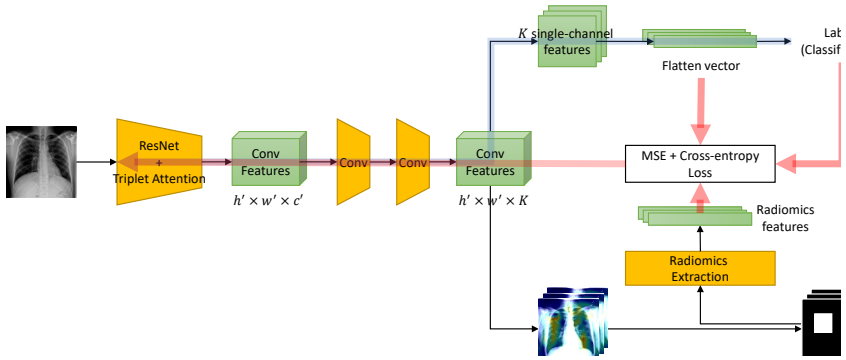

Figure 3.1: Model overview. The model contains three major parts. Blue arrows represents the feedforward multi-label classification part. The below black arrows represents the mask generation and radiomic features extraction part. Red arrows means the radiomic features regularization and backward part.

3.2.1 Model architecture

3.2.1.1 Branch I: Multi-label classification

In this branch, we label each image with a 14-dim vector $y = [y_1, \ldots, y_k, \ldots, y_K]$, $y_k \in 0, 1$, $K = 14$ for each image. y_k indicates the presence with respect to the according pathology in the image while a zero vector represents the status of "Normal" (no pathology is found in the scope of any of 14 disease categories as listed).

We use the residual neural network (ResNet) architecture [36], given its dominant performance in ILSVRC competitions and the triplet attention mechanism (see Section 3.2.1.3). However, our framework can be applied to other CNNs. ResNet-18 and ResNet-50 are used in this paper. After removing the final classification layer and global pooling layer, an input image with shape $h \times w \times c$ produces a feature tensor with shape $h' \times w' \times c'$ where h, w, and c

are the height, width, and number of channels of the input image, respectively while $h' = h/32$, $w' = w/32$, $c' = 2048$. The output of this network encodes the images into a set of abstracted feature maps. Then through an application of two convolutional layers (each followed by batch normalization and ReLU activation), the number of channels is modified to K, where K is the number of possible disease types. A perchannel probability for each disease class is then derived by a fully-connected layer with a sigmoid activation function; this is denoted $p(k|I)$, where the probability is that whether the image belongs to class k and I denotes the image. Since we intend to build K binary classifiers, we will exemplify just one class k. Note that kth binary classifiers will use the kth-channel features to do prediction. Since all images have their labels, the loss function for class k can be expressed as minimizing the binary-cross entropy as $L_k = -y_k \log p(k|I) - (1-y_k) \log(1-p(k|I))$, where y_k is the ground truth label of the k class. To enable end-to-end training across all classes, we sum up the class-wise losses to define the total loss as $L_I = \sum_k L_k$.

3.2.1.2 Branch II: Mask generation

In this branch, we generate bounding boxes (B-Box, or masks) based on the classification result of Branch I to get the most indicative areas using the class activation mappings (CAMs) [37]. The heatmap produced from the model indicates the approximate spatial location of one particular thoracic disease class each time. Due to the simplicity of intensity distributions in these resulting heatmaps, applying an ad-hoc thresholding-based B-Box generation

method for this task is found to be sufficient. Followed by the work of Wang et al. [20], the intensities in heatmaps are first normalized to $[0, 255]$ and then thresholded by $\{60, 180\}$ individually. Finally, B-Boxes are generated to cover the isolated regions in the resulting binary maps.

Radiomic features extraction. With the generated B-Boxes and original images, we extracted radiomic features to regularize the model. Quantitative radiomics can be categorized into the following subgroups:

- First-order statistics features describe the distribution of individual pixel values without concerns for spatial relationships. They are histogram-based properties using mean, median, maximum, and minimum values of the pixel intensities on the image, as well as their asymmetry, flatness, uniformity, and entropy.
- Shape features describe the shape of the region of interest (ROI) and its geometric properties (e.g., volume, maximum diameter along with different orthogonal directions, maximum surface, tumor compactness, and sphericity).
- A Gray Level Co-occurrence Matrix (GLCM) features describe the second-order joint probability function of an image region constrained by the mask. The matrix $P(i, j|\delta, \theta)$ represents the number of times the combination of levels i and j occurs in two pixels in the image, that are separated by a distance of δ pixels along angle θ.
- A Gray Level Size Zone (GLSZM) features quantify gray level zones in an

image. A gray level zone is defined as the number of connected pixels that share the same gray level intensity.

- A Gray Level Run Length Matrix (GLRLM) features quantify gray level runs, which are defined as the length in number of pixels, of consecutive pixels that have the same gray level value.
- A Neighboring Gray Tone Difference Matrix (NGTDM) features quantify the difference between a gray value and the average gray value of its neighbors within distance δ. The sum of absolute differences for gray level i is stored in the matrix.
- A Gray Level Dependence Matrix (GLDM) features quantify gray level dependencies in an image. A gray level dependency is defined as the number of connected pixels within distance δ that are dependent on the center pixel.

All above features can be extracted either directly from the images or after applying different filters or transforms (e.g., wavelet transform). In our design, we utilize the Pyradiomics tool to extract radiomic features (`https://pyradiomics.readthedocs.io/`).

Finally, we use the pairwise distance between radiomic features and image features as regularization. Therefore, the adjustable loss function is $L_{II} = L_I + \|I_F - R_F\|_p$, where I_F and R_F are the image features and radiomic features, respectively, and $\|\cdot\|$ denotes the norm and p represents the norm degree, e.g., $p = 1$ and $p = 2$ represent the Taxicab norm and Euclidean norm, respectively. In this paper, we set p to 2. Please note that although the

original shapes of I_F and R_F are not equal, we easily adapted one-layer MLP to project them into the same dimension space.

3.2.1.3 Triplet Attention

To boost the quality of masks, we integrate the triplet-attention mechanism [38]. Triplet Attention mechanism requires few learnable parameters and could capture important features by taking cross-dimension interaction into account [38]. In other words, it includes three sub-branches to respectively capture the dependency between spatial dimensions Height (H), Width (W), and the Channel (C) dimension. For the first branch, in measuring the interactions between dimension H and dimension C, it first performs a *Z-pool* operation by concatenating the result of average pooling and max pooling across dimension W. This operation can be summarized as $\chi_1^* = \text{z-pool}(\chi') = [MaxPool_w(\chi'); AvgPool_w(\chi')]$ where $\chi' \in \mathbb{R}^{W \times H \times C}$ is a 90 degree anti-clockwise rotation along the H axis from the output of the previous convolutional layer $\chi \in \mathbb{R}^{C \times H \times W}$ and $\chi_1^* \in \mathbb{R}^{2 \times H \times C}$ is the output of a Z-Pool operation. χ_1^* then passed through a standard 2D convolutional layer followed by sigmoid activation σ to get attention weights for χ_1^*. It would finally rotate back to match the original shape of χ after applying the attention weights. These steps can be represented by the following: $y_1 = r(\chi' \sigma(\text{CNN}_1(\chi_1^*)))$ where r is the rotation operation to retain the original shape of input. Similarly, y_2, y_3 are obtained from the last two branches by measuring the interactions between dimensions W and C and between dimensions W and H, respectively.

Note that the last branch is similar to the spatial attention in CBAM[39], and it requires no rotation. The refined input y is represented by averaging outputs from three branches: $y = \frac{1}{3}(y_1 + y_2 + y_3)$.

3.2.2 Training Strategy of ChexRadiNet

ChexRadiNet adopts an end-to-end multi-task training scheme. Each epoch consists of two tasks. In the first task (Branch I), we use the whole image to fine-tune the ResNet + Triplet Attention network pre-trained on ImageNet. During this process, we feed the generated masks into the radiomics extraction block to get radiomic features. In the second task (Branch II), we use radiomic features as regularization to further fine-tune the whole model. In each epoch, we use the model with the highest AUC on the validation set for testing.

3.3 Experiments

3.3.1 Datasets

Datasets	Patients	Chest X-rays
NIH Chest X-ray	30,805	112,120
CheXpert	65,240	224,316
MIMIC-CXR	227,827	377,110

Table 3.1: Descriptions of the datasets.

For the abnormality classification task, we evaluated the ChexRadiNet framework using the NIH Chest X-ray [20], CheXpert [22], and MIMIC-CXR [23] datasets (Table 3.1). The Chest X-ray dataset contains 112,120 X-ray

images collected from 30,805 patients. The disease labels were extracted from radiological reports with Natural Language Processing tools [40]. There are 15 classes, one for "No findings" and 14 diseases: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural thickening, Pneumonia, and Pneumothorax. The disease labels are expected to have above 90% accuracy. In addition, the Chest X-ray dataset includes 984 bounding boxes for 8 types of chest diseases annotated for 880 images by radiologists.

CheXpert dataset is another large-scale public chest X-ray dataset currently available, which contains 224,316 X-ray scans of 65,240 patients. This dataset was labeled for the presence of 14 observations, including 12 common thoracic pathologies. Each observation can be assigned to either positive (1), negative (0), or uncertain (-1). To simplify the task, we choose to ignore all the uncertain samples. In addition, to compare with previous literature, we follow the same evaluation protocol over 5 observations: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. MIMIC-CXR is also a large-scale CXR dataset, which contains 377,110 chest X-rays associated with 227,827 imaging studies. Images are provided with 13 labels. Similar to CheXpert, each label can be assigned to either positive (1), negative (0), or uncertain (-1).

3.3.2 Evaluation metrics and experimental settings

For the abnormality detection task, we randomly split each dataset into training (70%), validation (10%), and test (20%) sets. Note that there is no patient overlap between the sets. We use AUC scores, the area under the ROC curve, to measure the disease identification accuracy. A higher AUC score indicates better performance.

For the abnormality localization task, following the work of Li et al [27], we only consider 8 diseases for the evaluation of mask generation because only eight types of diseases are provided with bounding boxes in the NIH Chest X-ray dataset. We use intersection over union (IoU) to evaluate the predicted disease regions against the ground truth bounding boxes.

We use ResNet-50 as the backbone model. We set the batch size as 256 and train the model for 20 epochs. The model is optimized using the stochastic gradient descent (SGD) optimizer with a learning rate of 0.1 and decay the learning rate by 0.1 every 5 epochs of training. We trained our model on AWS with 16 Nvidia K80 GPUs. The model is implemented in PyTorch.

3.3.3 Results

3.3.3.1 Disease classification

Table 3.2 shows the AUC of each class and a mean AUC across the 14 chest diseases. We used ResNet-50 pre-trained on ImageNet as the backbone. Our ChexRadiNet outperforms other models in terms of mean AUC. For every single class, our proposed framework is better than all other models

except with DensNet-121 for Fibrosis, Hernia, Mass, Nodule, Pneumonia, and Pneumothorax. Possible reasons can be that Rajpurkar et al's backbone is much deeper than our ResNet-50 [1], which enables it to capture more discriminative features than our ResNet-50. In addition, "Mass" and "Nodule" parts are small and hard to detect. For "Fibrosis" and "Hernia", they are not annotated with bounding boxes and diffuse, and thus we cannot apply the weakly-supervised learning with radiomic features.

Table 3.2: AUC results on the NIH Chest X-ray dataset.

Method	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion
Wang et al., 2017[20]	0.716	0.807	0.708	0.835	0.784
Wang et al., 2018[21]	0.732	0.844	0.701	0.829	0.793
Yao et al., 2018[25]	0.772	0.904	0.788	0.882	0.859
Rajpurkar et al., 2017[1]	0.821	0.905	0.794	0.893	0.883
Kumar et al., 2017[41]	0.762	0.913	0.784	0.888	0.864
ChexRadiNet	**0.831**	**0.934**	**0.817**	**0.906**	**0.892**
Method	Emphysema	Fibrosis	Hernia	Infiltration	Mass
Wang et al., 2017[20]	0.815	0.769	0.767	0.609	0.706
Wang et al., 2018[21]	0.865	0.796	0.876	0.666	0.725
Yao et al., 2018[25]	0.829	0.767	0.914	0.695	0.792
Rajpurkar et al., 2017[1]	**0.926**	**0.804**	**0.939**	0.720	**0.862**
Kumar et al., 2017[41]	0.898	0.756	0.802	0.692	0.750
ChexRadiNet	0.925	0.798	0.882	**0.734**	0.846
Method	Nodule	Pleural Thickening	Pneumonia	Pneumothorax	**Mean**
Wang et al., 2017[20]	0.671	0.708	0.633	0.806	0.738
Wang et al., 2018[21]	0.685	0.735	0.720	0.847	0.772
Yao et al., 2018[25]	0.717	0.765	0.713	0.841	0.803
Rajpurkar et al., 2017[1]	**0.777**	0.814	**0.763**	**0.893**	0.842
Kumar et al., 2017[41]	0.666	0.774	0.715	0.859	0.795
ChexRadiNet	0.748	**0.867**	0.737	0.889	**0.843**

3.3.3.2 Disease localization

We compare our disease localization accuracy under varying IoU to other state-of-the-art models, shown in Table 3.3. Our model predicts well

not only for easy tasks but also for hard tasks like localizing "Mass" and "Nodule", where the disease localization is within a small area. When the IoU is set to 0.1, our model outperforms other models in terms of Atelectasis, Cardiomegaly, Effusion, and Pneumothorax. As the IoU threshold increases, our framework is superior to other models in terms of better accuracy and maintains great performance. For instance, when IoU is set to 0.3, our result for "Cardiomegaly" is 0.73 while the reference model is only 0.46. We get more than 0.15 accuracy improvement for Effusion, Infiltration, Mass, Pneumonia, and Pneumothorax. When IoU is set to 0.5, our result for "Cardiomegaly" is still as high as 0.59 while the reference model drops to barely 0.18.

Following Li et al.[27], we prefer a higher IoU threshold, i.e., IoU = 0.7, for disease localization because we expect high-accuracy disease localization application in clinical use. To this end, the method we proposed is superior to the baseline by a large margin.

Please note that for some diseases, e.g., Pneumonia and Infiltration, the localization of disease can appear in multiple places while only one bounding box is provided for each image. Thus, it is reasonable that our model doesn't align well with the ground truth when the threshold is as small as 0.1, especially for Pneumonia and infiltration. Overall, our model outperforms the reference models for all IoU thresholds except for T(IoU)=0.1 (probably because ground truth has missing annotation while ours does not).

Table 3.3: Disease localization under varying IoU on the NIH Chest X-ray dataset. Please note that since our model doesn't use any ground truth bounding box information, to fairly evaluate the performance of our model, we only consider the previous methods' results under the same setting, therefore, for the case T(IoU)=0.1, we have two baselines, but for the rest cases, we only have one baseline.

T(IoU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	**Mean**
0.1	Wang et al., 2017 [20]	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.569
	Li et al., 2018 [27]	0.63	0.89	0.78	0.91	0.70	0.29	0.31	0.44	0.619
	ChexRadiNet	0.72	0.96	0.81	0.88	0.67	0.33	0.59	0.47	**0.679**
0.2	Wang et al., 2017 [20]	0.47	0.68	0.45	0.48	0.26	0.05	0.35	0.23	0.371
	ChexRadiNet	0.49	0.84	0.62	0.54	0.46	0.21	0.43	0.39	**0.498**
0.3	Wang et al., 2017 [20]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.221
	ChexRadiNet	0.28	0.73	0.54	0.43	0.38	0.15	0.35	0.32	**0.398**
0.4	Wang et al., 2017 [20]	0.09	0.28	0.20	0.12	0.07	0.01	0.08	0.07	0.115
	ChexRadiNet	0.17	0.65	0.42	0.32	0.29	0.09	0.21	0.19	**0.293**
0.5	Wang et al., 2017 [20]	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.061
	ChexRadiNet	0.11	0.59	0.29	0.15	0.12	0.07	0.14	0.08	**0.194**
0.6	Wang et al., 2017 [20]	0.02	0.08	0.05	0.02	0.00	0.01	0.02	0.03	0.029
	ChexRadiNet	0.06	0.37	0.09	0.06	0.08	0.04	0.05	0.05	**0.100**
0.7	Wang et al., 2017 [20]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.011
	ChexRadiNet	0.02	0.21	0.04	0.02	0.07	0.01	0.03	0.04	**0.055**

3.4 Discussion

3.4.1 Ablation study

We conducted an ablation study to demonstrate the performance of radiomics on NIH Chest X-ray (Table 3.4), CheXpert (Table 3.5), and MIMIC-CXR (Table 3.6). We tried ResNet50+Triplet Attention without radiomic features. Table 3.4 shows that AUC will drop significantly when not using radiomic features. We observe the same trend in the other two datasets. This demonstrates that it is beneficial to include radiomic features.

We also report results of ChesxRadiNet using ResNet-18, a relevant small network, as a backbone. Table 3.7 shows the results with and without using the radiomic features in three datasets. We observe the AUCs drop significantly when not using radiomic features in all cases. This suggests that the generalizability of our proposed method in smaller networks. In addition, the ResNet-18 version still performs better than other models in Table 3.2 except Rajpurkar et al[1]. It indicates the superior of our proposed method for using radiomic features.

3.4.2 Qualitative analysis

Figure 3.2 shows the attention map of our model against the ground truth bounding boxes. The visualization provides better explainability of our model. In Figure 3.2 we visualized our results for Cardiomegaly, Mass, and Pneumonia.

Cardiomegaly is considered to be present if the cardiothoracic rate is

Table 3.4: Comparison of AUC on the NIH Chest X-ray dataset.

Method	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion
w/o radiomics	0.751	0.850	0.777	0.867	0.833
ChexRadiNet	0.831	0.934	0.817	0.906	0.892
Method	Emphysema	Fibrosis	Hernia	Infiltration	Mass
w/o radiomics	0.783	0.733	0.804	0.670	0.694
ChexRadiNet	0.925	0.798	0.882	0.734	0.846
Method	Nodule	Pleural Thickening	Pneumonia	Pneumothorax	**Mean**
w/o radiomics	0.643	0.699	0.700	0.792	0.757
ChexRadiNet	0.748	0.867	0.737	0.889	**0.842**

Table 3.5: Comparison of AUC on the CheXpert dataset.

Method	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion	**Mean**
w/o radiomics	0.781	0.813	0.893	0.918	0.921	0.865
ChexRadiNet	0.831	0.848	0.920	0.930	0.921	**0.890**

larger than 50% (cardiothoracic Ratio equals "Maximum horizontal cardiac width" over "Maximum horizontal thoracic width"), which means an enlarged heart. The 2nd image in the 1st row as well as the 2nd image in the 2nd row in Figure 3.2 shows that our model successfully detects cardiomegaly, an enlarged heart, perfectly, and aligns with the yellow bounding box well.

A lung mass is an abnormal spot in the lungs that is more than 3 centimeters. Our results (4th images in the 1st and 2nd rows), although focusing on larger areas, can capture some clues of lung mass.

Note that in the chest X-ray 14 dataset, only one bounding box is annotated for one disease image. Though some patients are diagnosed with several diseases, only the most important disease is annotated on the radiology image. This means that ground truth has missing annotations (shown by

Table 3.6: Comparison of AUC on the MIMIC-CXR dataset.

Method	Atelectasis	Cardiomegaly	Consolidation	Edema	Enlarged Card.
w/o radiomics	0.841	0.824	0.859	0.906	0.748
ChexRadiNet	0.851	0.831	0.866	0.900	0.767
Method	Fracture	Lung Lesion	Lung Opacity	Pleural Effusion	Pneumonia
w/o radiomics	0.713	0.782	0.775	0.923	0.753
ChexRadiNet	0.735	0.814	0.810	0.933	0.831
Method	Pneumothorax	Pleural Other	Support Devices	**Mean**	
w/o radiomics	0.909	0.850	0.931	0.832	
ChexRadiNet	0.919	0.909	0.937	**0.854**	

Table 3.7: Comparison of mean AUC on three datasets using ResNet-18 as a backbone.

	NIH Chest X-ray	CheXpert	MIMIC-CXR
w/o radiomics	0.749	0.854	0.822
ChesxRadiNet (ResNet-18)	0.810	0.883	0.837

Pneumonia). Pneumonia inflames the air sacs in one or both lungs. For Pneumonia detection, radiologists will look for white spots in the lungs. For the 6th image in the 2nd row, both lungs are infected and white spots are shown in both lungs. However, the bounding box of the 6th image only annotates the right lung while our model successfully localizes Pneumonia for both lungs.

Overall, our results show that the predicted disease localizations have a great alignment with the ground truth and can even serve as a supplement to the ground truth.

3.5 Conclusions

In this chapter, we propose a framework that jointly learns radiomic features and predicts 14 thoracic diseases. We evaluated our model on three

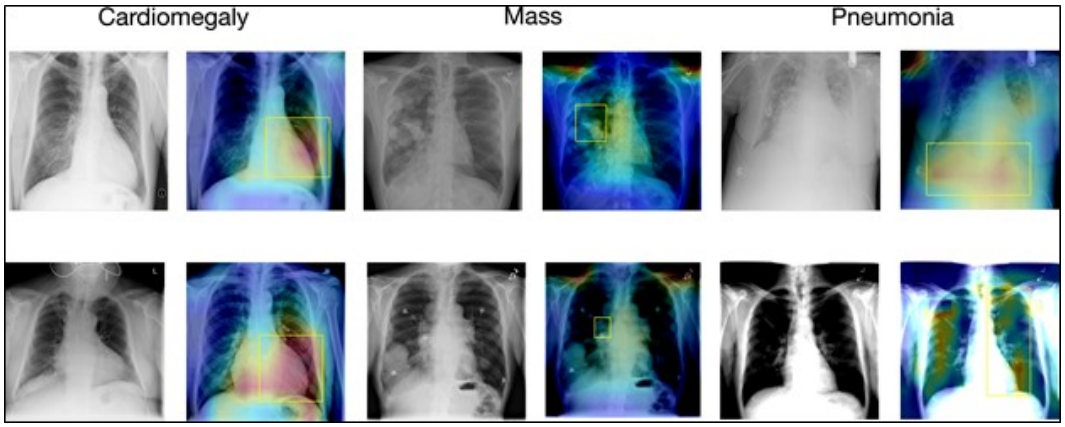

Figure 3.2: Visualization of the disease localization on the test images with ChexRadiNet and ground truth bounding boxes. The attention maps are generated from the final output tensor and overlapped on the original radiology images. The left image in each pair is the chest X-ray image and the right one is the generated attention map and the ground truth (in the yellow box).

publicly available corpora. We showed that both our disease identification and localization outperform state-of-the-art models in the quantitative and qualitative analysis. Our proposed framework has two main limitations. First, chest X-rays are very different from natural images, but we rely on deep learning models (ResNet) that work better on natural images. Second, the robustness of radiomic features relies on the accuracy of bounding boxes, in our work, the bounding boxes are generated by heatmaps. It is not guaranteed that the generated heatmaps are always good and accurate. Our future work will continue to solve these two limitations. Automatically generating correct bounding boxes can be a milestone to push the agenda for AI-driven medical imaging diagnosis. It can abruptly increase the annotated medical images at a much

lower cost so that better CNN models can be trained, therefore better diagnosis models can be obtained. Bounding boxes can increase the interpretability of AI solutions by locating the abnormalities as the visual evidence in medical images, which can build trust between doctors and patients.

Chapter 4

Knowledge-Augmented Contrastive Learning for Abnormality Classification and Localization in Chest X-rays with Radiomics using a Feedback Loop

4.1 Introduction

The chest X-ray is one of the most common radiological examinations for detecting cardiothoracic and pulmonary abnormalities. Due to the demand for accelerating chest X-ray analysis and interpretation along with the overall shortage of radiologists, there has been a surging interest in building automated systems of chest X-ray abnormality classification and localization [1]. While the class (i.e., outcomes) labels are important, the localization annotations, or the tightly-bound local regions of images that are most indicative of the pathology, often provide richer information for clinical decision making (either automated or human-based).

Automatic robust image analysis of chest X-rays currently faces many challenges. First, recognizing abnormalities in chest X-rays often requires expert radiologists. This process is therefore time-consuming and expensive to generate annotations for chest X-ray data, in particular the localized bounding box region labeling. Second, unlike natural images, chest X-rays have very

subtle and similar image features. The most indicative features are also very localized. Therefore, chest X-rays are sensitive to distortion and not amendable to typical image data augmentations such as random cropping or color jittering. Moreover, in addition to high inter-class variance of abnormalities seen in chest X-rays (i.e., feature differences between different diseases), chest X-rays also have large intra-class variance (i.e., differences in presentation among individuals of the same diseases). The appearance of certain diseases in X-rays are often vague, can overlap with other diagnoses, and can mimic many other benign abnormalities. Last but not least, the class distribution of chest X-rays is also highly imbalanced for available datasets.

Recently, contrastive learning has emerged as the front-runner for self-supervised learning, demonstrating superior ability to handle unlabelled data. Popular frameworks include MoCo [42, 43], SimCLR [13, 44], PIRL [45] and BYOL [46]. They all have achieved prevailing success in natural image machine learning tasks, such as image classification and object detection. Further, contrastive learning appears to be robust for semi-supervised learning when only few labeled data are available [44]. Recent works also found contrastive learning to be robust to data imbalance [47, 48].

Contrastive learning may offer a promising avenue for learning from the mostly unlabeled chest X-rays, but leveraging it for this task is not straightforward. One most important *technical barrier* is that most contrastive learning frameworks [42, 43, 13, 44, 46] critically depend on maximizing the similarity between two "views", i.e., an anchor and its **positive sample**, often being

generated by applying random data augmentations to the same image. This data augmentation strategy, however, does not easily translate to chest X-rays. In addition, the simultaneous demand for both classification and localization-aware features further complicates the issue. Fortunately, classical chest X-ray analysis has introduced **radiomic features** [49] as an auxiliary knowledge augmentation. The radiomic features can be considered as a strong *prior*, and therefore can potentially be utilized to guide learning of deep feature extractors. However, the extraction of reliable radiomic features via Pyradiomic[1] tool [17] heavily depends on the pathology localization – hence we will run into an intriguing "chicken-and-egg" problem, when trying to incorporate radiomic features into contrastive learning, whose goal includes learning the localization from unlabeled data.

This chapter presents an innovative holistic framework of **Knowledge-Augmented Contrastive Learning**, which seamlessly integrates radiomic features as the other contrastive knowledge-augmentation for the chest X-ray image. As the *main difference* from existing frameworks, the two "views" that we contrast now are from two different domain knowledge characterizing the same patient: the chest X-ray image and the radiomic features. Notably, the radiomic features have to be extracted from the learned pathology localizations, which are not readily available. As these features will be dynamically updated, forming a "feedback loop" during training in which both modalities' learning mutually reinforce each other. The *key enabling technique* to

[1] `https://pyradiomic.readthedocs.io/`

link this feedback loop is a novel module we designed, called *Bootstrap Your Own Positive Samples* (**BYOP**). For an unannotated X-ray image, we utilize Grad-CAM [50] to generate the input heatmap from the image modality backbone, which yields the estimated bounding box after thresholding; and we then extract the radiomic features within this estimated bounding box, which becomes the alternative view to contrast with the image view. The usage of radiomic features also adds to the model interpretability. Our contributions are outlined as follows:

- A brand-new **framework** dedicated to improving abnormality identification and localization in (mostly unannotated) chest X-rays by knowledge-augmented contrastive learning, which highlights exploiting radiomic features as the auxiliary knowledge augmentation to contrast with the images, given the inability to perform classical image data augmentation.

- An innovative **technique** called BYOP to enable the effective generation of radiomic features, which is necessary as the true bounding boxes are often absent. BYOP leverages an interpretable learning technique to supply estimated bounding boxes dynamically during training.

- Excellent experimental **results** achieved on the NIH Chest X-ray benchmark [7], using very few annotations. Besides improving the disease classification AUC from 82.8% to 83.8%, our framework significantly boosts the localization results, by an average of 2% over different IoU thresholds, compared to reported baselines. Figure 4.1 provides a visualization example showing

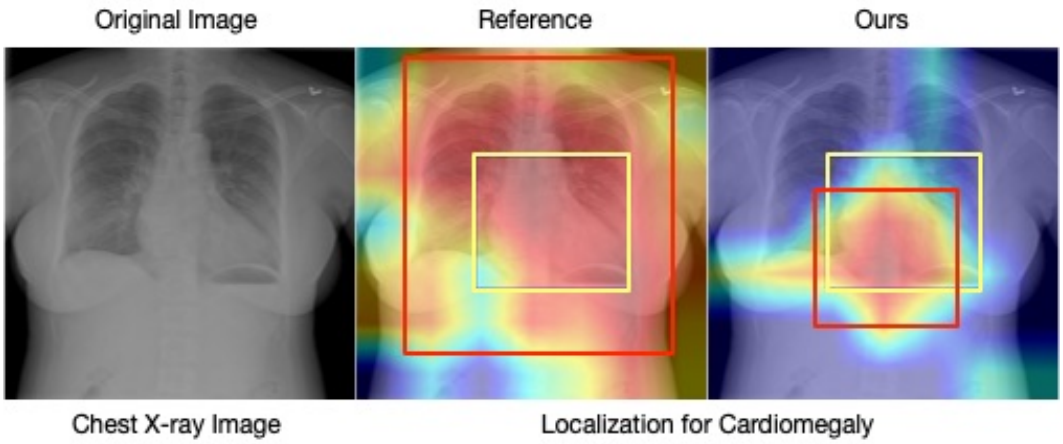

Figure 4.1: Visualization of heatmaps of chest X-rays with ground-truth bounding box annotations (yellow) and its prediction (red) for localize Cardiomegaly in one test chest X-ray image. The visualization is generated by rendering the final output tensor as heatmaps and overlaying it on the original images. The left image is the original chest X-ray image, the middle is the visualization result by CheXNet [1] and the right is our model's attempt. Best viewed in color.

our localization results to be more robust and accurate than the previous results from CheXNet [1],

4.2 Method

The Framework. Our goal is to learn an image representation y_i which can then be used for disease classification and localization. Our framework uses two neural networks to learn: the image and radiomics networks. The image network consists of an encoder f_i (ResNet-18) and a projector g_i (two-layer MLPs with ReLU). The radiomics network has a similar architec-

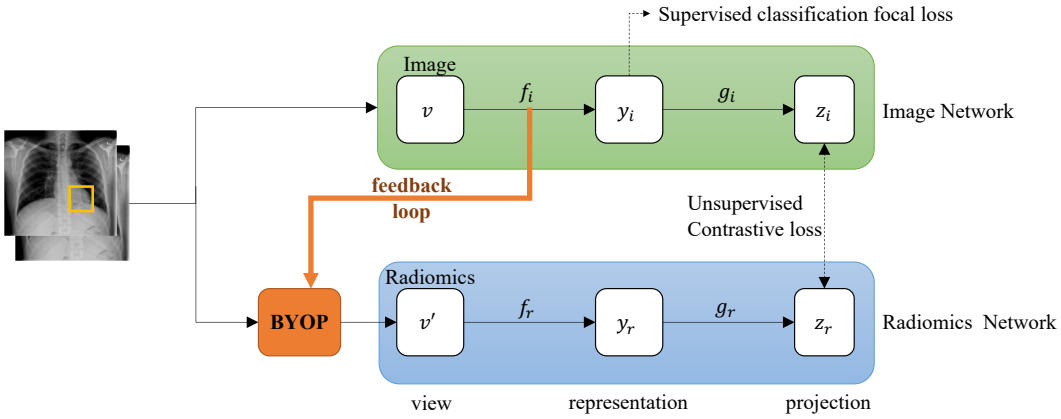

Figure 4.2: Overview of our proposed framework. During training, given a set of images, very few images have annotations, our framework provides two views: the image and the radiomic features (generated by the BYOP module, the detail view is shown in Figure 4.3). From the image view v, we output a representation $y_i = f_i(v)$ and a projection $z_i = g_i(y_i)$ via an image encoder f_i and image projection head g_i, respectively. Similarly, from the radiomic view v', we output $y_r = f_r(v')$ and the radiomic projection $z_r = g_r(y_r)$ via a radiomic encoder f_r and radiomic projection head g_r, respectively. We maximize agreement between z_i and z_r via a contrastive loss (NT-Xnet). In addition, we minimize the classification errors from representation y_i via a focal loss. During testing, only the image encoder is kept and applied to the new X-rays.

ture as the image network, but uses another three-layer MLPs for radiomic encoder f_r and a different set of weights for the projector g_r. The proposed architecture is summarized in Figure 4.2.

The primary innovation of our method lies in how we select positive and negative examples, which will be expanded below in Section 4.2.1 and Section 4.2.2. We also formulate the semi-supervised loss for our problem when a small amount of annotated data is available in Section 4.2.3. The entire framework can be trained from end to end, and the representation y_i will be used for

downstream disease classification and localization tasks.

4.2.1 Finding Positive and Negative Samples: Data-Driven Learning Meets Domain Expertise

The reasons to use contrastive learning as our framework are threefold. First, contrastive learning leverages unlabeled data and we have few disease localization (bounding boxes) annotations available. Second, empirical findings [47, 48] prove that contrastive learning is robust in classification tasks with class-imbalanced datasets. In clinical settings, most medical image datasets suffer an extreme class-imbalance problem [51]. Third, contrastive learning naturally fits "multi-view" concepts. In our case, we are still comparing two different views of the same subject, but unlike classic contrastive learning where two views are from the same domain space, our views for positive sampling are from different domain knowledge ([52] proved that views from multi-domain knowledge should also align), while our negative sampling is from the same domain knowledge. In the subsequent section, we will describe our unique positive and negative sampling methodologies in more detail.

Positive Sampling. To obtain a positive pair of views, we randomly select an image labeled with a given disease and generate two views for it. The first view will be its image features and another view will be its radiomic features. We decided to leverage radiomic features for the second view as traditional image augmentation strategies cannot be leveraged here. Furthermore, radiomic features have labels, are naturally more interpretable than the image features

extracted from deep learning-based image encoders.

Obtaining the radiomic features for our dataset is a "chicken-and-egg" problem. **Radiomic features are highly sensitive and dependent on local regions** for which we do not have local bounding box annotations. Meanwhile, we need to make the image features similar to the radiomic features to learn from radiomic features to better learn localization of the abnormalities. This process means that **bounding boxes generation is dependent on radiomic features** which forms a loop cycle. To address this issue, we design the *Bootstrap Your Own Positive Samples* (**BYOP**) method using such a feedback module. For more details, see Section 4.2.2.

Negative Sampling. The original images are used for views of the negative samples because the same domain is supposed to be more similar and thus harder for the model to distinguish between the positive and negative samples, leading to a more robust model [53]. Besides, the image features focuses on local regions highlighted by the attention map rather than the whole image. To identify harder negative samples, we go one step further, by not only selecting any random image, but "hard similar" images. Here, we first get prior knowledge from the pre-constructed disease hierarchy relationship for image negative sampling, shown in Figure 4.4, defined by [2]. The pre-constructed disease hierarchy relationship is initialed with 21 nodes. In this hierarchy, each disease (green) belongs to a body part (grey). We therefore only treat normal chest X-rays or images within the same body part but with a different disease as negative examples. We call these negative examples "hard similar" images

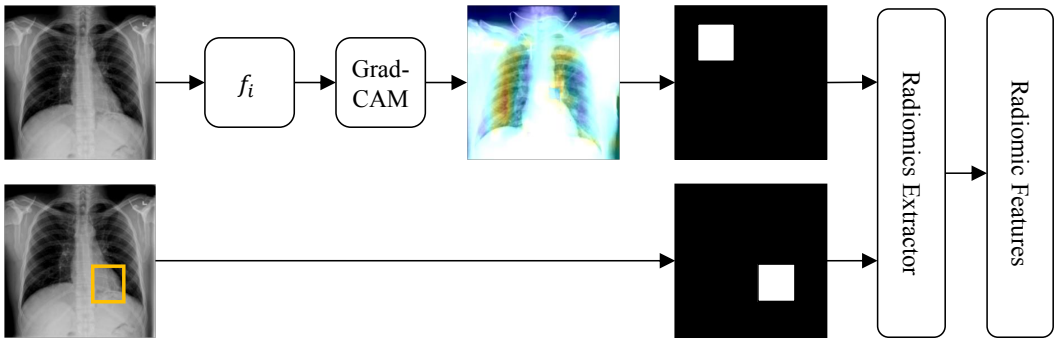

Figure 4.3: Overview of our *BYOP module*. For the unannotated images, we leverage *Grad-CAM* to generate heatmaps and apply an ad-hoc threshold to generate the bounding boxes. For the annotated images, we directly use the ground-truth bounding boxes. Then with the combination of generated bounding boxes and ground-truth bounding boxes, we use the *Pyradiomic* tool as the radiomic extractor to extract the radiomic features. Note that the generated radiomic features are the combination of the accurate and 'pseudo' radiomic features for annotated and unannotated images, respectively.

in this study. As an example, if our "anchor" image is labeled as "Pneumonia/Lung" , our "hard similar" images should include "Atelectasis/Lung", "Edema/Lung", or "Normal" but not "Bone Fractures".

4.2.2 Bootstrap Your Own Positive Samples (BYOP) with Radiomics in the Feedback Loop

The core component of our cross-modal contrastive learning is the *Bootstrap Your Own Positive Samples* (**BYOP**) module. BYOP leverages a feedback loop to learn region localization from generated radiomic features as the positive sample for the image features. The architecture of BYOP is shown in Figure 4.3. The BYOP contains two components, bounding box generation and radiomic features extraction.

Bounding Boxes Generation. We feed the fourth layer of the image encoder f_i (i.e., ResNet-18) to the Gradient-weighted Class Activate Mapping (Grad-CAM) [50] to extract attention maps and apply an ad-hoc threshold to generate bounding boxes from the attention maps.

Radiomic Features Extraction. The radiomic features are composed of the following categories:

- First-Order statistics features measure the distribution of voxel intensities within the bounding boxes. The features include energy (the measurement of the magnitude of voxel values), entropy (the measurement of uncertainty in the image values), and max/mean/median gray level intensity within the region of interest (ROI), etc.

- Shape-based features include features like Mesh Surface, Pixel Surface, Perimeter, and etc.

- Gray-level features include a gray-level features include a Gray Level Cooccurance Matrix (GLCM) features, a Gray Level Size Zone (GLSZM) features, a Gray Level Run Length Matrix (GLRLM) features, a Neighboring Gray Tone Difference Matrix (NGTDM) features, and a Gray Level Dependence Matrix (GLDM) features.

Given the original images and generated bounding boxes, we used the Pyradiomic tool to extract radiomic features [17].

4.2.3 Semi-Supervised Loss Function

Our framework is mixed with supervised classification and unsupervised contrastive learning. For the localization task, we use the knowledge-augmented contrastive loss for unsupervised contrastive learning. For the classification task, we could have used standard cross-entropy loss, but considering that the chest X-ray dataset is highly imbalanced, we instead find focal loss more helpful [54]. We briefly review the two loss functions below.

Unsupervised Knowledge-Augmented Contrastive Loss. Our cross-modal contrastive loss function extends the normalized temperature-scaled cross-entropy loss (NT-Xent). We randomly sample a minibatch of N examples and define the contrastive prediction task on pairs of augmented examples derived from the minibatch. Let v_{bd} be the image in the minibatch with disease d and body part b, and $\text{sim}(u, v)$ be the cosine similarity. The loss function $\ell_{v_{bd}}$ for a positive pair of example (v_{bd}, v'_{bd}) is defined as

$$\ell_{v_{bd}} = -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{z}_i(v_{bd}), \boldsymbol{z}_r(v'_{bd})\right)/\tau\right)}{\sum \mathbb{1}_{[k=b, l\neq d]} \exp\left(\text{sim}\left(\boldsymbol{z}_i(v_{bd}), \boldsymbol{z}_i(v_{kd})\right)/\tau\right)}$$

where $\mathbb{1}_{[k=b,l\neq d]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $k = b$ and $l \neq d$. τ is the temperature parameter. The final unsupervised contrastive loss $\mathcal{L}_{cl}$ is computed across all disease-positive images in the minibatch.

Supervised Focal Loss. We feed the output of the image encoder f_i to a simple linear classifier. The supervised classification focal loss is defined as

$$\mathcal{L}_{fl} = \begin{cases} -\alpha\left(1 - y'\right)^{\gamma} \log y', & y = 1 \\ -(1-\alpha) y'^{\gamma} \log\left(1 - y'\right), & y = 0 \end{cases}$$

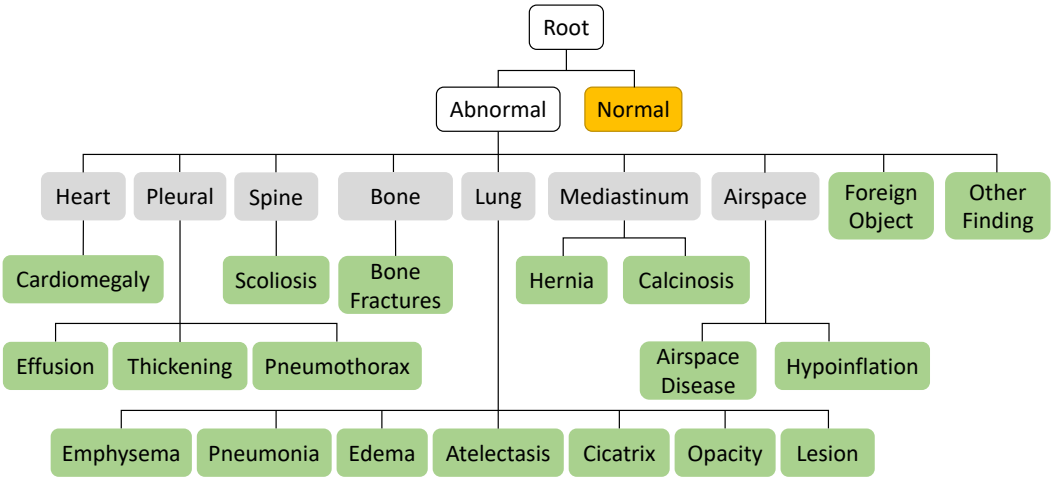

Figure 4.4: Disease hierarchy relationship predefined based on domain expertise, reprinted from [2].

α allows us to give different importance to positive and negative examples. γ is used to distinguish easy and hard samples and force the model to learn more from difficult examples.

Eventually, we treat it as multi-task learning (one task is supervised disease classification and one is unsupervised contrastive learning) and the total loss is defined as

$$\mathcal{L} = \lambda \times \mathcal{L}_{cl} + (1 - \lambda) \times \mathcal{L}_{fl}$$

4.3 Experiments

Dataset and Protocol Setting. We evaluated our framework using the NIH Chest X-ray dataset [7]. It contains 112,120 X-ray images collected from 30,805 patients. As other large chest X-ray datasets, this dataset is also

Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	**Mean**
Wang et. al.[7]	0.72	0.81	0.78	0.61	0.71	0.67	0.63	0.81	0.718
Wang et. al.[21]	0.73	0.84	0.79	0.67	0.73	0.69	0.72	0.85	0.753
Yao et. al.[25]	0.77	0.90	0.86	0.70	0.79	0.72	0.71	0.84	0.786
Rajpurkar et. al.[1]	0.82	0.91	0.88	0.72	0.86	0.78	0.76	0.89	0.828
Kumar et. al.[55]	0.76	0.91	0.86	0.69	0.75	0.67	0.72	0.86	0.778
Liu et. al.[56]	0.79	0.87	0.88	0.69	0.81	0.73	0.75	0.89	0.801
Seyyed et. al.[57]	0.81	0.92	0.87	0.72	0.83	0.78	0.76	0.88	0.821
Our model	0.84	0.93	0.88	0.72	0.87	0.79	0.77	0.90	0.838

Table 4.1: Comparison with the baseline models for AUC of each class and average AUC. For each column, red values denote the best results.

extremely class imbalanced: the healthy cases (84,321 front-view images) are far more than cases with diseases (24,624 front-view images), and different disease occurrence frequencies vary dramatically. The disease labels were extracted from radiology reports with a rule-based tool [40]. There are 9 classes, specifically one for "No findings" and 8 for diseases (Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax). The disease labels are expected to have above 90% accuracy. In addition, the dataset includes 984 bounding boxes for 8 types of chest diseases annotated for 880 images by radiologists. We separate the images with provided bounding boxes from the entire dataset. Hence, we have two sets of images called "annotated" (880 images) and "unannotated" (111,240 images).

In our experiment, we follow the same protocol of [7], to shuffle the unannotated dataset into three subsets: 70% for training, 10% for validation, and 20% for testing. For the annotated dataset, we randomly split the dataset into two subsets: 20% for training and 80% for testing. Note that there is no patient overlap between all the sets.

Evaluation Metrics. For the disease classification task, we use Area under the Receiver Operating Characteristic curve (AUC) to measure the performance of our model. For the disease localization task, we evaluate the detected regions against annotated ground truth bounding boxes, using intersection over union ratio (IoU). The localization results are only calculated on the test set of the annotated dataset. The localization is defined as correct only if IoU $>$ T(IoU), where T(*) is the threshold.

Implementation Details. We use the ResNet-18 model as the image encoder. We initialize the image encoder with the weights from the pre-trained ImageNet model except for the last fully-connected layer. We set the batch size as 64 and train the model for 30 epochs. We optimize the model by the Adam method and decay the learning rate by 0.1 from 0.001 every 5 epochs. Furthermore, we use linear warmup for the first 10 epochs only for the disease classification task, which helps the model converge faster to generate stable heatmaps. We train our model on AWS with one Nvidia Tesla V100 GPU. The model is implemented in PyTorch.

4.3.1 Disease Classification

Table 5.2 shows the AUC of each class and a mean AUC across the 8 chest diseases. Compared to a series of relevant baseline models, our proposed model achieves better AUC scores for the majority of diseases. The overall improvement in performance is remarkable when compared to other models except CheXNet [1]. One possible reason for our lack of improvement can be

that [1]'s backbone is DenseNet-121, which is much deeper than the ResNet-18 in our model. It thus, able to capture much more discriminative features than our ResNet-18. Despite the fact, our model still achieves better or comparable results than CheXNet, which demonstrates that the cross-modal contrastive learning branch boosts the robustness of the image features without the need to increase the complexity of the backbone. Specifically, the performance of our model demonstrates significant improvements for disease abnormalities with larger associated regions on the image, such as "Ateclectasis", "Cardiomegaly", and "Pneumothorax". In addition, small objects features like "Mass" and "Nodule", are recognized as well as in CheXNet. In summary, these experimental results show the superiority of our proposed model over relevant other methodologies.

4.3.2 Disease Localization

We compare our disease localization accuracy to other state-of-the-art models under different IoU thresholds (Table 4.2). Since disease localization is not an easy task in chest X-ray images, we did not find as many other methods as for disease classification task. To our knowledge, we only have two baseline methods from [7] and [27]. From these comparisons, we find our model significantly outperforms baselines by an average of 2% over different IoU thresholds. Importantly, our model is able to perform well not only on the easier tasks, but also for more difficult ones like localizing "Mass" and "Nodule", where the disease localization is within a small area. When the IoU

T(IoU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	**Mean**
0.1	*Wang et. al.*[7]	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.569
	Li et. al.[27]	0.71	0.98	0.87	0.92	0.71	0.40	0.60	0.63	0.728
	Our model	0.72	0.96	0.88	0.93	0.74	0.45	0.65	0.64	0.746
0.2	*Wang et. al.*[7]	0.47	0.68	0.45	0.48	0.26	0.05	0.35	0.23	0.371
	Li et. al.[27]	0.53	0.97	0.76	0.83	0.59	0.29	0.50	0.51	0.622
	Our model	0.55	0.89	0.78	0.85	0.62	0.31	0.52	0.54	0.633
0.3	*Wang et. al.*[7]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.221
	Li et. al.[27]	0.36	0.94	0.56	0.66	0.45	0.17	0.39	0.44	0.496
	Our model	0.39	0.85	0.60	0.67	0.43	0.21	0.40	0.45	0.500
0.4	*Wang et. al.*[7]	0.09	0.28	0.20	0.12	0.07	0.01	0.08	0.07	0.115
	Li et. al.[27]	0.25	0.88	0.37	0.50	0.33	0.11	0.26	0.29	0.374
	Our model	0.24	0.81	0.42	0.54	0.34	0.13	0.28	0.32	0.385
0.5	*Wang et. al.*[7]	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.061
	Li et. al.[27]	0.14	0.84	0.22	0.30	0.22	0.07	0.17	0.19	0.269
	Our model	0.16	0.77	0.29	0.35	0.24	0.09	0.15	0.22	0.284
0.6	*Wang et. al.*[7]	0.02	0.08	0.05	0.02	0.00	0.01	0.02	0.03	0.029
	Li et. al.[27]	0.07	0.73	0.15	0.18	0.16	0.03	0.10	0.12	0.193
	Our model	0.09	0.74	0.19	0.16	0.18	0.04	0.11	0.14	0.206
0.7	*Wang et. al.*[7]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.011
	Li et. al.[27]	0.04	0.52	0.07	0.09	0.11	0.01	0.05	0.05	0.118
	Our model	0.05	0.54	0.09	0.11	0.12	0.02	0.07	0.06	0.133

Table 4.2: Disease localization accuracy comparison under different IoU thresholds. Red numbers denote the best result for each column.

Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	**Mean**
Base	0.75	0.85	0.83	0.67	0.69	0.64	0.70	0.79	0.740
w. FL	0.78	0.84	0.80	0.68	0.76	0.72	0.72	0.82	0.765
w. BYOP	0.82	0.90	0.85	0.71	0.82	0.75	0.74	0.86	0.806
Full model	0.84	0.93	0.88	0.72	0.87	0.79	0.77	0.90	0.838

Table 4.3: Ablation studies on focal loss and BYOP module for disease classification. Red numbers denote the best result for each column.

threshold is set to 0.1, our model outperforms others on all diseases except for "Cardiomegaly". As the IoU threshold increases, our framework is superior to other models in terms of better accuracy and maintains this superior performance. For instance, when the threshold increases, the IoUs of "Cardiomegaly" decrease less than the baselines and even outperform the baselines when IoU threshold is above 0.5.

We prefer a higher IoU threshold, specifically, IoU = 0.7, for disease localization because we expect high-accuracy disease localization application

is necessary for clinical applications. To this end, the method we propose is superior to the baseline by a slight margin.

It is also worth nothing that, for some diseases, such as Pneumonia and Infiltration, the localization of disease can appear in multiple places while only one bounding box is provided for each image. Hence, it is reasonable that our model does not align well with the ground truth when the threshold is as small as 0.1, especially for Pneumonia and Infiltration. Overall, our model outperforms the reference models for almost all IoU thresholds.

4.3.3 Ablation Discussion

In this section, we study the contribution of our BYOP module on both disease classification and localization tasks.

Disease Classification. For this task, note that the use of focal loss should also boost the model with the class-imbalanced chest X-ray dataset. Thus, we compare the performance of our base model with only focal loss (labeled "w. FL") or with only the BYOP module (labeled "w. BYOP"), respectively. As shown in Table 4.3, although both focal loss and BYOP improve the model performance, BYOP contributed more strongly. This stronger contribution is expected since BYOP tends to generate more robust radiomic features, which further reinforces the image encoder to focus on the image region that contains the targeted disease.

Disease Localization. Note that our base model is a ResNet-18 image encoder, which is not as powerful as CheXNet [1] with DenseNet-121.

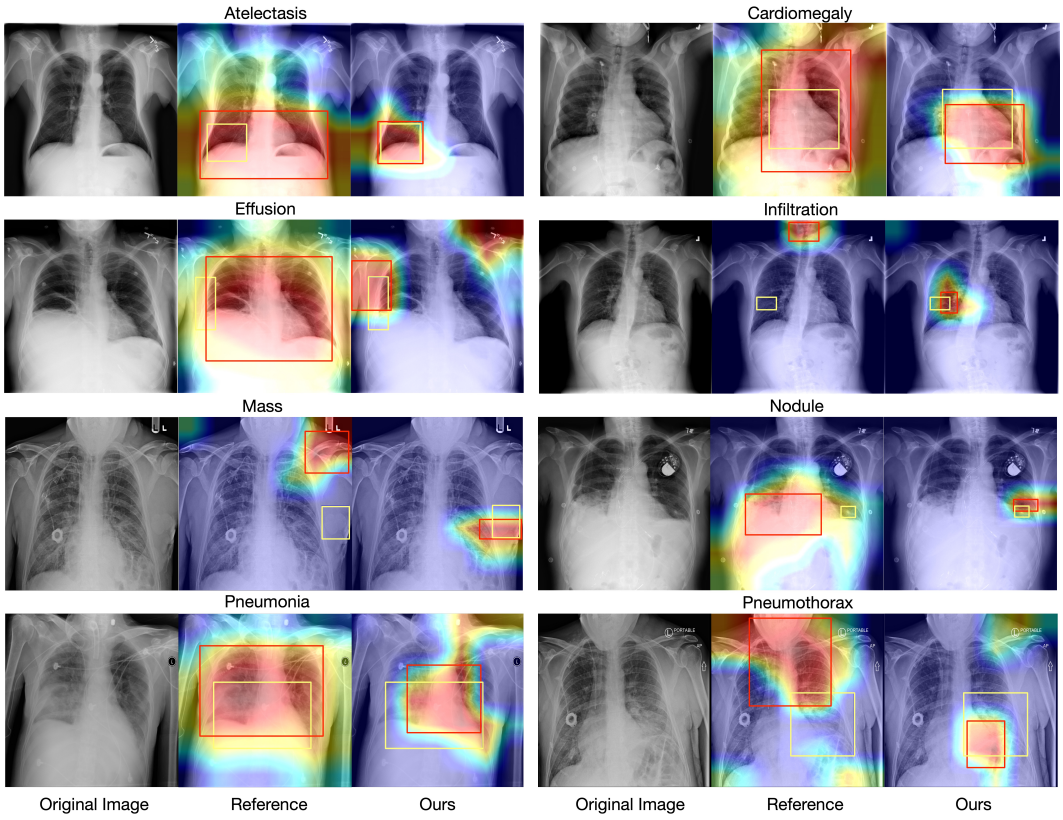

Figure 4.5: Examples of visualization of localization on the test images. We plot the results of diseases near thoracic. The attention maps are generated from the fourth layer of ResNet-18. The ground-truth bounding boxes and the predicted bounding boxes are shown in yellow and red, respectively. The left image in each pair is the original chest X-ray image, the middle one is the localization result of CheXNet [1] and the right one is our localization result. All examples are positive for corresponding disease labels. Best viewed in color.

Thus we compare the performance of our model with CheXNet. As shown in Figure 4.5, our localization result is superior to the CheXNet. For the example of 'Atelectasis', 'Cardiomegaly', 'Effusion', 'Nodule', 'Pneumonia' and 'Pneumothorax', while the baseline model tends to focus on a large area of the image, our model precisely captures the correct disease location. For harder localization cases like 'Mass' and 'Nodule', the baseline model's focus is incorrect and does not have any overlap with the ground-truth areas while our model still predicts perfectly. The results demonstrate that the BYOP module significantly boosts the model performance.

4.4 Conclusions

In this chapter, we propose a semi-supervised, end-to-end knowledge-augmented contrastive learning model that can jointly model disease classification and localization with limited localization annotation data. Our approach differs from previous studies in the choice of data augmentation, the use of radiomic features as prior knowledge, and a feedback loop for image and radiomic features to mutually reinforce each other. Additionally, the project aims to address current gaps in radiology by making prior knowledge more accessible to image data analytic and diagnostic assisting tools, with the hope that this will increase the model's interpretability. Experimental results demonstrate that our method outperforms the state-of-the-art algorithms, especially for the disease localization task, where our method can generate more accurate bounding boxes. Importantly, we hope the method developed here is inspiring

for the future research on incorporating different kinds of prior knowledge of medical images with contrastive learning.

Chapter 5

Radiomics-Guided Global-Local Transformer for Weakly Supervised Pathology Localization in Chest X-Rays

5.1 Introduction

In medicine, *radiomics* refers to the process of extracting quantitative and semiquantitative features from medical images, such as radiographs or computed tomography scans, for improved decision support [58]. These handcrafted radiomic features aim to describe a local "region of interest" such as a tumor with numeric features that assess qualities such as size, shape, texture, variations in pixel intensity, and relationships between neighboring pixels [59]. Given their advantages, researchers have explored the performance of radiomic features for chest X-ray analysis. For example, Shi *et al.* [60] and Saygılı [61] each extracted a set of radiomic features, which were then used to diagnose different types of pneumonia. Bai *et al.* [62] proposed a hybrid model to encode the combination of radiomic features and clinical information. Ghosh *et al.* [63] presented a new handcrafted feature to distinguish between severe and nonsevere patients. However, all of the above methods rely on *accurate pathology localization annotations* to extract radiomic features from a correct and clinically meaningful region of interest [17]. Such bounding boxes are usu-

ally expensive and time-consuming to acquire by humans and, if inaccurate, will tremendously degrade the reliability of radiomic features. There is thus an unmet need to automatically localize cardiopulmonary pathologies on chest X-rays to facilitate extraction of radiomic features.

Throughout the rapid development of deep learning approaches for medical image analysis, many researchers have made efforts utilizing convolutional neural networks (CNNs) to build automated systems for chest X-ray abnormality classification and localization [1, 7, 27, 56, 64, 65, 66, 67, 68, 69, 70]. However, CNN methods bear several limitations when applied to the domain of chest radiography. First, CNNs do not naturally incorporate contextual prior information, such as reason for imaging and patient history, or domain knowledge such as human anatomy and typical disease presentation on imaging. Since radiomic features are designed by humans and semantically describe local medical image regions, they represent an auxiliary modality of information embedded with domain-specific quantitative features that can enhance automated disease localization and classification. Second, chest X-rays have more subtle discriminative features compared to natural images, making their recognition more challenging. Finally, though many have studied the interpretability of deep image classifiers for other data [71, 72, 73, 74, 75, 76], deep CNNs are often criticized for their lack of human interpretability, thus posing a major barrier to their adoption by clinicians.

With this in mind, Transformers, which have seen a surge in popularity for a variety of visual recognition tasks, provide a promising alternative

to CNNs for modeling chest X-rays. The Transformer was first introduced in the context of natural language processing [77, 78, 79], followed by its recent success in computer vision [4, 80, 81] and multi-modal learning [82]. The Transformer architecture can be considered a "universal modeling tool" that can unify the feature extraction and fusion processes from different input modalities with a *single* model that does not require domain-specific architecture tweaks. For example, Arkbari *et al.* [83] demonstrated the ability to learn powerful multi-modal representations from unlabeled video, audio, and text data, using a single multimodal Transformer. Nagrani *et al.* design a bottleneck fusion technique that allows audio- and video-derived features to interact throughout their custom Transformer architecture [84]. And Shvetsova *et al.* [85] proposed a multi-modal, modality agnostic fusion Transformer to learn to exchange information between multiple modalities, such as video, audio, and text, and integrate them into a jointly multi-modal representation to obtain an embedding that aggregates multi-modal temporal information.

In the context of modeling chest X-rays, we observe the unique potential for a Transformer-based architecture to *naturally and jointly learn from two "views" of chest X-rays*: (1) raw X-ray images that contain rich contrast details, hence benefiting from the data-driven learning capacity, and (2) radiomics that encode domain-specific quantitative features, thus guiding and regularizing the learning process with handcrafted local radiomic features. However, there exists a "chicken-and-egg" problem: extraction of useful radiomic features relies on accurate pathology localization, but the pathology

localization is often absent and first needs to be learned or separately acquired.

This chapter presents **RGT**, a **R**adiomics-**G**uided **T**ransformer (Fig. 5.1). RGT consists of two Transformer-based branches, one for the raw chest X-ray and one for the radiomic features extracted from the corresponding image. Features extracted from these two "views" of the patient are then deeply fused with interaction via cross-attention layers [86]. Of note, the radiomic features need to be extracted from the learned pathology localizations, which are not readily available. The *key enabling technology* to resolve this hurdle is to construct a feedback loop, called the ***B**ring **Y**our **O**wn **A**ttention* (**BYOA**) module, which will be expanded in Sec 5.2.2. During training, the image branch leverages its learned self-attention to estimate pathology location, which is then used to extract radiomic features from the original image for further processing by the radiomics branch. In addition to a supervised classification loss, we optimize the model with a contrastive loss that rectifies the image-derived and radiomics-derived "views" of the patient, and such an end-to-end optimization loop can bootstrap accurate pathology localizations from image data with *no bounding box annotations* used for training.

Our contributions are outlined as follows:

- We leverage radiomics as an "auxiliary input modality" that both correlates with the raw image modality and encodes domain-specific quantitative features. We then propose a novel radiomics-guided cross-attention Transformer, **RGT**, to jointly extract and fuse global image features and local

radiomic features for disease localization and classification in chest X-rays.

- To resolve the key "chicken-and-egg" problem of extracting radiomic features without available cardiopulmonary pathology localization, we construct an innovative optimization loop where the learned image-level attention map is used to extract local radiomic features. Such an end-to-end loop can bootstrap accurate cardiopulmonary pathology localization from images without leveraging human-annotated bounding boxes.

- On the NIH ChestXRay benchmark [7], our approach achieves superior disease localization and classification results. RGT outperforms prior work in weakly supervised localization by an average margin of 3.6% over different intersection-over-union (IoU) thresholds.

5.2 Method

An overview of RGT is illustrated in Fig. 5.2. In the following subsections, we will first present Cross-Attention Vision Transformer (CrossViT), a recent two-branch ViT backbone on which RGT is built, and then describe the methodological innovations required to naturally incorporate domain-specific quantitative features in the form of radiomics for improved cardiopulmonary pathology localization and classification.

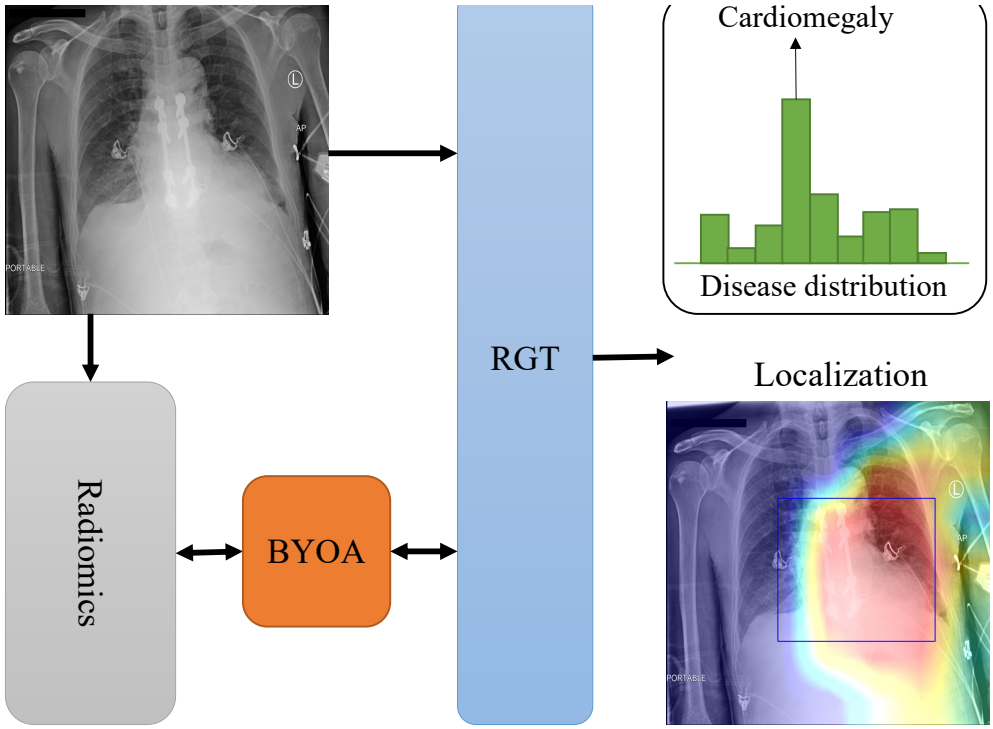

Figure 5.1: General overview of our **R**adiomics-**G**uided **T**ransformer (**RGT**) framework for weakly supervised cardiopulmonary disease localization and classification from chest X-rays. RGT takes a chest X-ray as the input and produces a heatmap for pathology localization, from which a bounding box is obtained. Radiomic features are further extracted from the bounded region and fused with image-derived features to classify the pathology present. The detailed views of **RGT** framework and ***B**ring **Y**our **O**wn **A**ttention* (**BYOA**) module are given in Fig. 5.2 and Fig. 5.3, respectively.

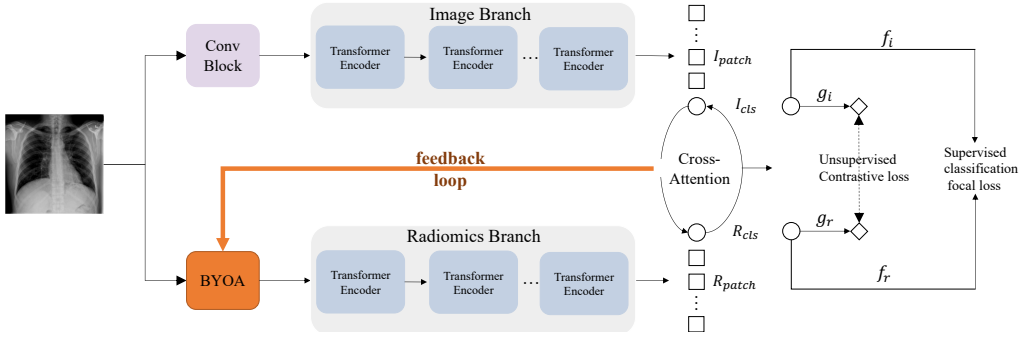

Figure 5.2: Overview of our proposed model, RGT. The image branch is a Transformer that processes a chest X-ray, and the radiomics branch is a small Transformer that processes radiomic features generated by the Bootstrap Your Own Attention (BYOA) module (Fig. 5.3). The global image representations and local radiomics representations are then fused by an efficient cross-attention module operation on each branch's `CLS` tokens. Finally, the `CLS` tokens I_{cls} (from the image branch) and R_{cls} (from the radiomics branch) are used for disease classification. We optimize the classification error with the Focal Loss [3]. We also leverage a contrastive learning strategy that aims to rectify the global image view with the local radiomics view. Specifically, RGT generates an image view $z_i = g_i(I_{cls})$ by a projection head g_i and radiomic view $z_r = g_r(R_{cls})$ by projection head g_r. We maximize the agreement between z_i and z_r via a contrastive loss (NT-Xent).

5.2.1 Preliminary: ViT and Cross-Attention

ViT first converts an image into a sequence of patch tokens by dividing the image into fixed-size patches and linearly projecting each patch into so-called "tokens". A special `CLS` (class) token is prepended to the sequence of image patches, as in the original BERT [78]. Then, all tokens are passed through stacked Transformer encoder layers. Finally, the hidden state corresponding to the `CLS` token is used as the aggregate sequence representation used for image classification.

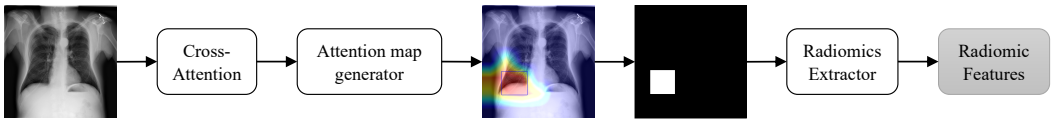

Figure 5.3: Overview of our Bootstrap Your Own Attention (BYOA) module. For the input chest X-rays, we look at the self-attention of the `CLS` token of the Image branch on the heads of the final output of the cross attention module. Then we apply a threshold of 0.1, meaning we only keep the top 10% of pixels in the generated attention map, to produce bounding boxes. Then with the generated bounding boxes, we use the *Pyradiomics* tool to extract radiomic features from the region of interest.

A Transformer encoder is composed of a sequence of blocks, where each block consists of (1) a multi-headed self-attention and (2) a feed-forward neural network. Layer normalization and residual shortcuts are, respectively, applied before and after every block. The granularity of the patch size affects the accuracy and complexity of ViT. Therefore, ViT was observed to reach greater performance with smaller (more fine-grained) patch sizes, but at the cost of higher floating-point operations (FLOPS) and memory consumption [86]. To relieve this problem, CrossViT [86] proposed a dual-branch ViT where each branch operates at a different patch size, as its own "view" of the image. The cross-attention module is then used to fuse information between the branches in order to balance the patch sizes and complexity. Similar to ViT, the final hidden vector obtained from the `CLS` tokens from the two branches are then used for image classification.

5.2.2 Our Proposed RGT Model

CrossViT supplies a graceful framework to simultaneously process and fuse two different "views" from the same input data (e.g., different-size image patches in the original paper) [86]. In RGT, we extend this idea by treating the raw image itself as one "view" and the radiomic feature extracted from this image as another "view" (Fig. 5.2). The global image representation and local radiomics representations are then fused by interacting through cross-attention. Here, a Transformer serves as the modality-agnostic backbone for both views.

Specifically, we introduce a dual-branch cross-attention Transformer where the first (primary) branch operates on the image, while the second (auxiliary) branch handles the radiomic features. To resolve the "chicken-and-egg" dilemma in extracting reliable radiomic features without bounding boxes, we have designed a novel ***B**ootstrap **Y**our **O**wn **A**ttention* (**BYOA**) module, using a feedback loop to learn pathology localization for radiomic feature extraction. A simple yet effective module is also utilized to fuse information between the branches. In the subsequent sections, we will describe the two branches, the BOYA module, and the fusion module.

Image Branch. The primary image branch uses a Progressive-Sampling ViT (PS-ViT) [87] as its backbone. Unlike the vanilla ViT that splits images into fixed-size tokens, PS-ViT utilizes an iterative and progressive sampling strategy to locate discriminative regions and avoid over-partitioning object structures. We experimentally observed PS-ViT outperforms ViT and

other variants in our framework because it generates higher-quality and more structure-aware attention maps, which are crucial for estimating the pathology localization during training.

Radiomics Branch. The complementary radiomics branch is used to process and learn deep representations of radiomic features. Handcrafted features can encompass a wide range of categories, such as first-order (basic intensity and shaped-based features), second-order (texture features extracted from various matrices), and more advanced features including those calculated from Fourier and wavelet transforms. Specifically, the 107 radiomic features utilized in this work come from the following categories described below:

- First-order statistics measure the distribution of pixel intensities within the region of interest. Such features include energy (the measurement of the magnitude of pixel values), entropy (the measurement of uncertainty in the image values), and max/mean/median gray level intensity. In total, we extract 18 first-order radiomic features.

- Shape-based features – such as mesh surface, pixel surface, and perimeter – describe the two-dimensional size and shape of the region of interest. While RGT can only produce rectangular bounding boxes for radiomics extraction, shape-based features can still be useful to quantify the size and aspect ratio of the extracted region. A total of 14 shape features are used in this work.

- Gray-level features describe statistical patterns in the pixel intensity values, drawn from the Gray Level Co-occurrence Matrix (GLCM), Gray Level Size

Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), Neighboring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM). In particular, we compute 24 GLCM features, 16 GLSZM features, 16 GLRLM features, 5 NGTDM features, and 14 GLDM features.

For this branch, we use a vanilla Transformer [88] as the radiomics encoder. Please note that the only difference is that the positional encoding module is discarded, since there does not exist any positional relationship between individual radiomic features.

Bootstrap Your Own Attention (BYOA): A Feedback Loop Module. Our main *roadblock* concerns how to generate robust radiomic features without pathology localization. On one hand, radiomic features are highly sensitive to the choice of local region of interest, for which we have no bounding box annotation. On the other hand, image features would benefit from the guidance of radiomics that encode important domain-specific quantitative features. The learning of image and radiomic features thus mutually depend on each other, forming a challenging chicken-and-egg problem.

To address this issue, we design **BYOA** to constitute an end-to-end feedback loop that can bootstrap accurate pathology localization from image data without any bounding box annotations (Fig. 5.3). BYOA contains two components: attention map generation and radiomic feature extraction.

- *Attention Map Generation.* Similar to the approach in Caron *et al.* [89], we extract self-attention of the `CLS` token from the heads of the last layer.

RGT produces two `CLS` tokens from two branches, but the attention maps *only* come from the image branch. To generate bounding boxes for radiomic features extraction, we first apply a threshold on the learned self-attention maps. This threshold, controlling the percentage of most responsive pixels kept for further processing, will influence the size of the resulting bounding box and thus the quality of radiomic features. After thresholding the attention map, image processing steps including a maximum filter and five consecutive binary dilations are used to "grow" the region of interest and smooth boundaries. Then, connected-components labeling is performed, after which we find the "center of mass" of each component. If this center of mass pixel is in the top decile of intensity values, a bounding box is drawn around it according to the mean height and width of the known bounding box annotations for the given disease class of interest. Here, we utilize one kind of prior knowledge of different diseases, e.g. Cardiomegaly usually occurs in the heart area, and localized Pneumonia usually occurs in the lung area, and the information of the average bounding boxes of these diseases could be seen as one kind of free-available prior knowledge, which could improve the accuracy of our model. And as one limitation of our method, for stable training, our method will generate per-class identically-sized bounding boxes. But during testing, we relaxed the setting of the bounding box generation.

- *Radiomic Features Extraction.* Given the original images and generated bounding boxes, we used Pyradiomics [17] to extract a variety of radiomic

features, including 18 first-order features, 14 shape-based features, and 73 gray-level features (see Appendix for full list). For feature extraction, we adopt the default settings of PyRadiomics version 3.0.1, which includes no spatial resampling, discretization, rescaling, or normalization; this is not necessary, as input radiographs have already been min-max normalized as part of model preprocessing. All features are derived from the original image (no wavelet, Laplacian of Gaussian, or other filters are applied before feature extraction).

Cross-Attention Fusion Module. To aggregate global image information with local radiomics information, this fusion step involves the `CLS` token of the image branch and patch tokens of the radiomics branch, similarly, it also involves the `CLS` token of the radiomics branch and patch tokens of the image branch. As the `CLS` token is the aggregate representation of the branch, this interaction helps include information from multiple scales. Please refer to Chen *et al.* [86] for more details about the cross-attention mechanism.

5.2.3 Semi-Supervised Loss Function

In our framework, we aim to make the learned image features from the `CLS` token similar to the learned radiomic features in order to localize pathologies in the chest X-rays. As shown in Fig. 5.2, RGT is trained using the linear combination of the supervised classification and unsupervised contrastive losses. For the supervised classification, considering that the chest X-ray dataset is usually highly imbalanced, we adopt the Focal Loss [3]. For

unsupervised contrastive learning, we use the cross-view contrastive loss [13].

Supervised Classification Focal Loss. We feed the output of the `CLS` tokens I_{cls} (from the image branch) and R_{cls} (from the radiomics branch) to a simple linear classifier. The supervised classification focal loss $\mathcal{L}_{fl}$ is defined as

$$\mathcal{L}_{fl} = \begin{cases} -\alpha\,(1-y')^{\gamma}\log y', & y=1 \\ -(1-\alpha)y'^{\gamma}\log\,(1-y')\,, & y=0 \end{cases} \tag{5.1}$$

The hyperparameter α allows us to give different importance to positive and negative examples, whereas γ is used to distinguish easy and hard samples , forcing the model to place more emphasis on difficult examples.

Unsupervised Cross-View Contrastive Loss. Our contrastive loss extends the normalized temperature scaled cross-entropy loss (NT-Xent). The difference is that we maximize agreement between two feature views extracted from different input formats, one from the image and the other from radiomic features.

Given an anchor chest X-ray in a minibatch, the positive sample will be its radiomic feature view, and the negative samples will be other chest X-rays (both image and radiomics views). Since the `CLS` token can be regarded as the representation of the input modality, we only need to maximize the agreement between each modality's `CLS` tokens. Suppose $I_{cls,k}$ and $R_{cls,k}$ are the $k-$th image features and radiomic features in the minibatch, respectively, and $sim(\cdot)$ the cosine similarity. Then the contrastive loss function $\mathcal{L}_{cl}$ is defined as

$$\mathcal{L}_{cl} = -\log \frac{\exp(sim(g_i(I_{cls,k}), g_r(R_{cls,k}))/\tau)}{\sum_{k=1}^{N}\exp(sim(g_i(I_{cls,k}), g_r(R_{cls,k}))/\tau)} \tag{5.2}$$

where τ is the temperature. The final contrastive loss is summed over all instances in the minibatch.

Overall, we treat RGT training as a weakly-supervised multi-task learning problem. In our chest X-ray setting, there exist two types of labels: disease class labels and pathology bounding box annotations. In our case, we *only* use the disease labels for training, even though the ultimate goal is to accurately localize those pathologies. Here, when we say "weakly-supervised" localization, we mean that we are able to localize pathologies only using supervision from whole-image disease labels. The combined loss function for supervised disease classification and unsupervised cross-view contrastive learning is as follows:

$$\mathcal{L} = (1 - \lambda) \times \mathcal{L}_{cl} + \lambda \times \mathcal{L}_{fl} \tag{5.3}$$

5.3 Experiments

5.3.1 Dataset and Protocol Setting

The NIH ChestXRay dataset [7] consists of 112,120 chest X-rays collected from 30,805 patients, where each image is labeled with one or more of 14 cardiopulmonary diseases. The labels are extracted from the associated radiology report using an automatic labeler [40] with a reported accuracy of 90%. For a subset of 880 images, the NIH dataset also provides bounding box localizations associated with eight disease classes: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax. The remaining six diseases are diffuse in nature, meaning it is not clinically mean-

ingful to provide a "localization" for these pathologies. Since this study aims to develop a model for weakly supervised disease localization, we only proceed with the eight diseases which have ground truth bounding box annotations. Specifically, we only use the image-level disease labels for these eight focal diseases to train RGT, binning all other classes into the already provided "No Findings" category. A significant difference between our method and existing baseline methods for pathology localization [90, 27] is that our method does not require any training data related to the bounding box while others use some percentage of these images for training.

In our experiments, we followed the same protocol as in related studies [7, 27], randomly partitioning the dataset (excluding images with bounding box annotations) into three subsets: 70% for training, 10% for validation, and 20% for testing. In order to prevent data leakage across patients, we make sure that there is no patient overlap between our train, validation, and test set.

5.3.2 Implementation Details

We build our image branch encoder based on PS-ViT [87], and apply their default hyperparameters for training. We use a shallower image encoder than the original PS-ViT, using 6 layers. For the radiomic branch encoder, since the radiomic features are already informative features, we use a small standard Transformer (2 layers) to learn representations of the radiomic features. We then add one more cross-attention layer to fuse the learned image

Table 5.1: Weakly supervised pathology localization results on the NIH ChestXRay dataset as measured by IoU accuracy at a fixed threshold. Please note that since RGT was solely supervised by disease class labels (not pathology localizations), we only compare localization performance with previous methods following the same setting for fair evaluation.

T(IoU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	**Mean**
0.1	Wang *et al.* [7]	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.569
	ViT	0.58	0.91	0.61	0.77	0.44	0.11	0.75	0.25	0.553
	RGT	0.61	0.95	0.65	0.82	0.50	0.13	0.79	0.28	**0.591**
0.2	Wang *et al.* [7]	0.47	0.68	0.45	0.48	0.26	0.05	0.35	0.23	0.371
	ViT	0.38	0.85	0.39	0.55	0.24	0.01	0.51	0.15	0.385
	RGT	0.41	0.91	0.41	0.59	0.26	0.05	0.57	0.19	**0.424**
0.3	Wang *et al.* [7]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.221
	ViT	0.20	0.45	0.19	0.32	0.06	0.00	0.21	0.02	0.181
	RGT	0.28	0.79	0.22	0.38	0.12	0.01	0.41	0.05	**0.283**
0.4	Wang *et al.* [7]	0.09	0.28	0.20	0.12	0.07	0.01	0.08	0.07	0.115
	ViT	0.10	0.21	0.03	0.05	0.02	0.00	0.04	0.00	0.056
	RGT	0.17	0.54	0.13	0.18	0.07	0.01	0.26	0.02	**0.173**
0.5	Wang *et al.* [7]	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.061
	ViT	0.05	0.15	0.01	0.04	0.02	0.00	0.03	0.00	0.034
	RGT	0.08	0.32	0.05	0.09	0.05	0.00	0.12	0.01	**0.090**
0.6	Wang *et al.* [7]	0.02	0.08	0.05	0.02	0.00	0.01	0.02	0.03	0.029
	ViT	0.01	0.03	0.01	0.01	0.01	0.00	0.01	0.00	0.010
	RGT	0.02	0.15	0.03	0.04	0.03	0.00	0.06	0.00	**0.041**
0.7	Wang *et al.* [7]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.011
	ViT	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.001
	RGT	0.01	0.04	0.01	0.02	0.01	0.00	0.03	0.00	**0.015**

features with the learned radiomic features. We set the batch size to 128 and train the model for 50 epochs. We used a cosine linear-rate scheduler with a linear warm-up of 5 epochs, an initial learning rate of 0.004, and a weight decay of 0.05. We downscale the images to 224×224 and normalize based on the mean and standard deviation of images in the ImageNet training set. We also augment the training data with random horizontal flipping. During the evaluation, we resize the image to 256×256 and take the center crop 224×224 as the input.

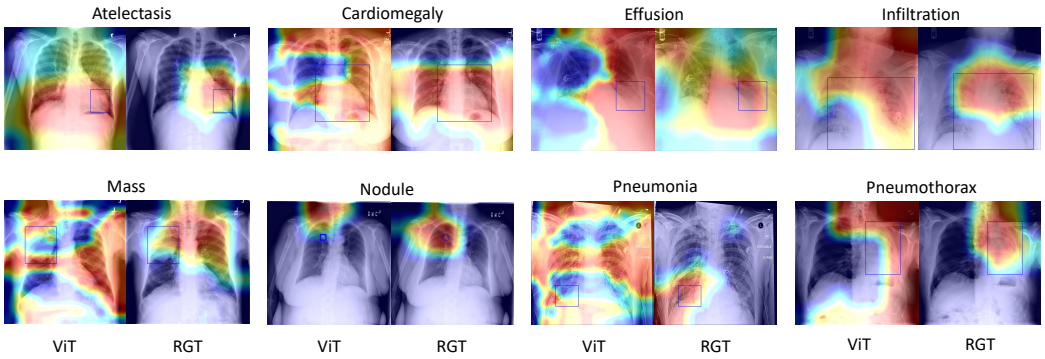

Figure 5.4: Example visualizations of pathology localization when evaluated on the 880 NIH ChestXRay images with bounding box annotations. The attention maps are generated from the self-attention maps of the `CLS` token. The ground-truth bounding boxes are shown in blue. The left image in each pair is the localization result of ViT [4], and the right one is our localization results obtained by RGT. All examples are positive for the corresponding disease labels. Best viewed in color.

5.3.3 Pathology Localization

The NIH Chest X-ray dataset contains 880 images labeled by radiologists with bounding box information, which we use to evaluate the performance of RGT for pathology localization. Many prior works [27, 90] have used a fraction of ground truth (GT) bounding boxes for training and evaluated their system on the remaining examples. Unlike these approaches, RGT *uses no bounding box annotations during training*, only using the subset of bounding box-annotated images for evaluation. Table 5.1 presents our evaluation results on all 880 images. We used [7] as our baseline to compare our localization results since it follows the same experimental setting of weakly supervised training on only disease labels.

Table 5.2: Pathology classification results for CNN- and Transformer-based methods on the NIH ChestXRay dataset, as measured by AUC. For each column, **bold** values denote the best results for the given disease class. For RGT, the average AUC per class is presented, with the standard deviation in parentheses, across three training runs with different random initializations.

Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	**Mean**
CNN									
Wang *et al.* [7]	0.72	0.81	0.78	0.61	0.71	0.67	0.63	0.81	0.718
Wang *et al.* [21]	0.73	0.84	0.79	0.67	0.73	0.69	0.72	0.85	0.753
Yao *et al.* [25]	0.77	0.90	0.86	0.70	0.79	0.72	0.71	0.84	0.786
Rajpurkar *et al.* [1]	0.82	0.91	**0.88**	0.72	0.86	0.78	0.76	**0.89**	0.828
Kumar *et al.* [55]	0.76	0.91	0.86	0.69	0.75	0.67	0.72	0.86	0.778
Liu *et al.* [56]	0.79	0.87	**0.88**	0.69	0.81	0.73	0.75	**0.89**	0.801
Seyyed *et al.* [57]	0.81	**0.92**	0.87	0.72	0.83	0.78	0.76	0.88	0.821
Han *et al.* [91]	**0.83**	**0.92**	0.87	0.76	0.85	0.76	0.77	0.86	0.828
Transformer									
ViT	0.74	0.78	0.81	0.72	0.70	0.66	0.65	0.76	0.728
CrossViT	0.69	0.71	0.72	0.72	0.74	0.79	**0.82**	0.88	0.759
PS-ViT	0.75	0.81	0.82	0.73	0.79	0.73	0.69	0.81	0.766
RGT (ours)	0.80	**0.92**	0.78	**0.86**	**0.88**	**0.88**	0.79	0.81	**0.839**
	(±0.02)	(±0.00)	(±0.01)	(±0.01)	(±0.02)	(±0.00)	(±0.01)	(±0.02)	–

5.3.3.1 Evaluation Metric

For localization, we evaluated our detected regular rectangular regions against the annotated bounding boxes, using a thresholded **IoU accuracy**, following Wang *et al.* [7]. Our localization results are only calculated for the 880 images that have ground truth annotation for 8 diseases. To compute IoU accuracy, the localization is defined as "correct" only if the observed IoU between the predicted and ground truth localization exceeds a fixed IoU threshold, T(IoU). We evaluated RGT for different thresholds ranging from {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7} as shown in Table 5.1.

5.3.3.2 Comparison with Prior Works

We compared disease localization accuracy under varying IoU with baselines following the same training setting as RGT (Table 5.1). Unlike other baselines[27, 90] that use a portion of 880 images for evaluation (because they need the remaining data for training), we used all 880 annotated images for evaluation. Therefore, no k-fold cross-validation for localization was performed. RGT average localization performance across 8 diseases is considerably higher than the baseline under all IoU thresholds. When the IoU threshold is set to 0.1, RGT outperforms the baseline [7] in the Cardiomegaly, Infiltration, Mass, and Pneumonia classes. Even with higher thresholds, our model is superior to the baseline. For example, when evaluated at T(IoU) = 0.5, our "Cardiomegaly" accuracy is 32%, while the reference model achieves 18%. Similarly, our "Pneumonia" accuracy is 12%, while the reference model reaches 3% accuracy. Note that some diseases can appear in multiple locations, but the ground truth might have mentioned only one such location. This can significantly impact the accuracy at high thresholds.

5.3.3.3 Discussion of Visualization

More importantly, we also include our own trained ViT as an additional baseline here. The quantitative results above demonstrate that, compared to the standard ViT, the additional radiomics branch and BYOA module enable RGT to learn more accurate and fine-grained pathology localizations. Example visualizations of localization results of both ViT and RGT can be

seen in Fig. 5.4. We can observe that RGT produces qualtitatively more accurate localizations than ViT for all diseases, but particularly Atelectasis, Cardiomegaly, Infiltration, Pneumonia, and Pneumothorax. Visualizations for most diseases reveal that both models often attend to regions outside the clinically relevant region of interest. However, RGT consistently attends to a smaller number of "extraneous" pixels than the standard ViT. Further, RGT always contains a significant portion of the ground truth localized region, while the ViT attention map does not – for example, see Nodule, Pneumonia, and Pneumothorax.

5.3.4 Pathology Classification

Pathology classification for chest X-rays is a multi-label classification problem. The objective is to assign one or more labels (among 8 cardiopulmonary diseases: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax) to each input image at inference time. We compared RGT with related reference approaches, which represent state-of-the-art disease classifcation performance on the NIH ChestXRay dataset. For RGT, we report the average AUC of 3 runs to show the robustness of our model.

5.3.4.1 Evaluation Metric

We used Area under the Receiver Operating Characteristic Curve (AUC) to estimate the performance of our model [92]. A higher AUC score implies a

model that is more capable of discriminating between classes. We also provide mean AUC across all the classes to highlight the overall performance of our model.

5.3.4.2 Comparison with Prior Works

AUC scores for each disease and mean AUC across eight diseases are presented in Table 5.2. We not only compared RGT with previous CNN-based state-of-the-art (SOTA) models, but also several Transformer-based models. We find that RGT outperformed all baseline approaches with respect to mean AUC across all diseases; specifically, RGT reached 0.839 mean AUC, outperforming the previous SOTA for disease classification [1] by a margin of 0.011. When considering classification performance on individual disease classes, RGT also achieved best performance on four of the eight classes. Our proposed model outperformed the next-best baseline by a margin 0.13 AUC for Infiltration, 0.09 for Nodule, and 0.02 for Mass. Compared to the Transformer-based models, the key difference is that we utilize the extracted radiomic features for disease prediction, improving the classification accuracy and enriching the model's interpretability due to the utilization of handcrafted radiomic features. Please note that Liu *et al.* [56] used 5-fold cross-validation in their model evaluation. While the problem settings are very similar, the evaluation schemes are so different that a direct comparison of this work to RGT would be inappropriate.

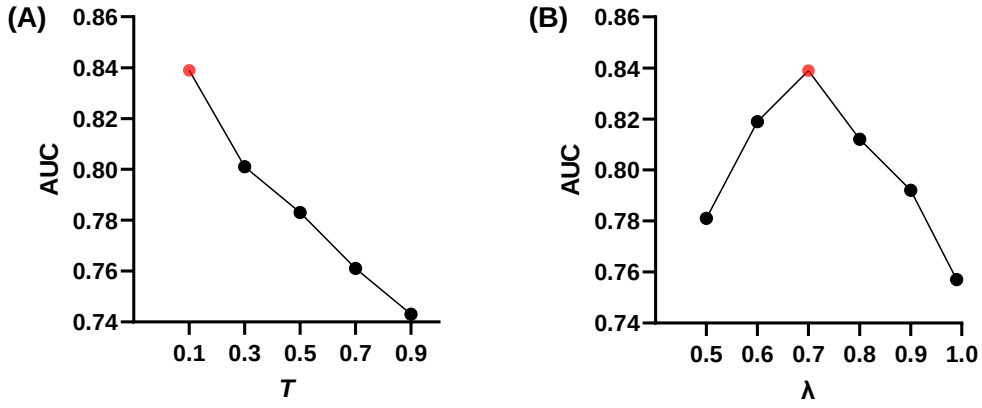

Figure 5.5: Effect of (A) varying T in attention map generation and (B) varying λ in Equation (3) on pathology classification for the NIH ChestXRay dataset.

5.3.4.3 Effect of Attention Map Threshold

We investigate the impact of the threshold (T), used in the process of attention map generation, on the performance of RGT for disease classification. Fig. 5.5A summarizes the AUC comparison of RGT for different values of T. Higher values of T imply smaller bounding boxes from which to extract radiomic features. During our experiments, we found that RGT performs better on the disease classification task when larger bounding boxes are generated. Since radiomic features are typically computed for highly localized – often small – regions of interest, this was originally an unintuitive finding. There appears to be a tradeoff between bounding box size and the resulting performance on the disease classification and localization tasks. Specifically, extracting smaller boxes that are accurately localized and ignore as much background signal as possible should lead to more robust and useful radiomic

features. However, attending to smaller regions of the image comes at the expense of decreasing the "receptive field" of learned global image features, thus degrading the quality of the classification task. This observation emphasizes the difference between disease classification and localization tasks: global information aids classification while rich local information aids localization.

5.3.4.4 Effect of Contrastive Learning

We also investigated the impact of the unsupervised contrastive loss on RGT's disease classification ability. Specifically, we evaluate the performance of RGT for disease classification by varying λ in equation (3). Fig. 5.5B summarizes the AUC comparison of RGT for different values of λ. Higher values of λ implies lower weight to contrastive loss. During our experiments, we found that RGT performs worse when small weight (1%) is given to the contrastive loss. RGT's performance improves when we increase contrastive loss weight, but after a certain point ($\lambda = 0.7$), performance considerably decreases. This confirms our hypothesis that both contrastive and focal losses are important, but that care must be taken to properly balance the objectives. The supervised classification and unsupervised constrastive losses enable RGT to learn both disease-level and patient-level discriminative visual features.

5.3.5 System Usability Study

We hired two radiologist experts to validate the usefulness of RGT's disease localizations; one expert has 3.5 years of experience, while the other

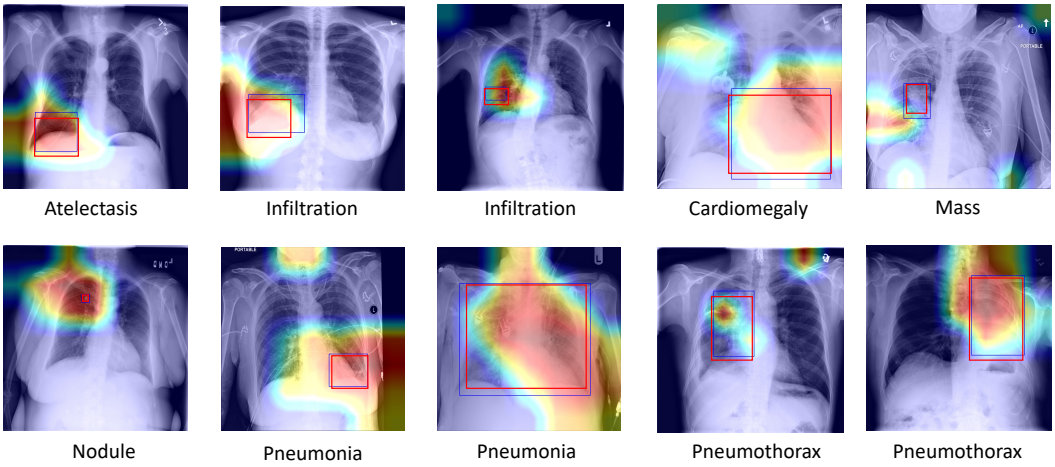

Figure 5.6: System Usability Study. Visualizations of pathology localization come from 10 randomly selected chest X-rays from the NIH ChestXRay dataset that do not have ground truth localization annotations. Saliency heat maps are generated from the self-attention maps of the `CLS` token from our trained RGT model. The red and blue represent the pathology localizations provided by two radiologists, who were instructed to draw a rectangular bounding box around the most salient image region in 90 seconds.

has 5. For this additional study, we randomly selected 10 images from the NIH ChestXRay dataset that did not have ground truth bounding box annotations. We then used RGT to predict the disease classification and localization visualization. Finally, we asked two radiologists to provide their own pathology localization for each image. Each radiologist was instructed to draw a rectangular bounding box around the clinically relevant region of interest within 90 seconds. Results can be seen in Fig. 5.6. Each image contains two human-annotated bounding boxes (red is Expert 1, and blue is Expert 2) and the extracted attention map from our RGT.

For disease classification, the two radiologists agreed with RGT's pre-

diction for all ten cases. For the localization task, we observe that the inter-rater consistency between two radiologists is very high, suggesting that they clearly agreed on the most salient image region. Overall, the radiologists found the RGT attention maps to significantly overlap with their own localizations, demonstrating the usefulness of our approach. With the exception of the "Mass" example, there is a strong agreement between the most responsive pixels in RGT's heat map and the radiologists' annotations.

5.3.6 Limitations and Discussion

There exist two main limitations to this approach. One limitation is the fact that self-attention provides only a coarse approximation of salient regions unless trained on extremely large amounts of data (e.g., see DINO [89]). Without this scale of chest radiography data available, other principled methods for saliency visualization may provide more fine-grained localizations for radiomics extraction than the native self-attention of our proposed RGT architecture. For example, the Anchors approach of Ribeiro *et al.* [73] or other input space visualization methods like LIME [72], SHAP [76], and deep Taylor decomposition [74] could be used in place of our proposed heatmap generation process. Future work will consider adapting such approaches to generate more accurate localizations for improved radiomics feature extraction, and thus better downstream disease classification and localization.

Another limitation of this approach is the fact that fixed-sized bounding boxes per target disease class are generated during RGT training. This would,

for example, make it difficult to distinguish the visual presentation of diffuse vs. localized pneumonia. However, this can be alleviated with finer granularity in the image-level disease labels used to train RGT; for instance, if "diffuse" and "localized" pneumonia were distinct class labels, then RGT would be able to provide visually distinct localizations of the two conditions. Future work may involve a module that learns the optimal bounding box dimensions for each disease in an unsupervised manner. Alternatively, the adoption of other saliency visualization methods instead of RGT's self-attention – as explained in the previous paragraph – may resolve this limitation of fixed-size bounding boxes per disease class.

5.4 Conclusion

In this chapter, we propose a radiomics-guided cross-attention Transformer, RGT, that can jointly localize and classify abnormalities in chest X-rays without supervision from bounding box annotations. Our approach differs from previous related studies in the choice of a unified Transformer architecture, the use of radiomic features, and a feedback loop for image and radiomic features to mutually interact during the training process. This work aims to bring the field of computer-aided diagnosis closer to clinical practice by making domain-specific quantitative features (in the form of radiomics) more accessible to automated medical image analysis tools, with the hope that this will increase the model's interpretability. Experimental results demonstrate that our method outperforms state-of-the-art algorithms in this weakly supervised

setting, particularly for disease localization, where our method can generate more accurate and clinically useful bounding boxes.

Bibliography

[1] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv:1711.05225*, 2017.

[2] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 910–12 917.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[5] S. Jain, W. H. Self, R. G. Wunderink, S. Fakhran, R. Balk, A. M. Bramley, C. Reed, C. G. Grijalva, E. J. Anderson, D. M. Courtney *et al.*, "Community-acquired pneumonia requiring hospitalization among

us adults," *New England Journal of Medicine*, vol. 373, no. 5, pp. 415–427, 2015.

[6] Y.-X. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B. A. Redd, C. J. Brandon, Z. Lu *et al.*, "Automated abnormality classification of chest radiographs using deep convolutional neural networks," *NPJ digital medicine*, vol. 3, p. 70, 2020.

[7] 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*, 2017, pp. 2097–2106.

[8] A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, and J. J. Rodrigues, "Identifying pneumonia in chest x-rays: A deep learning approach," *Measurement*, vol. 145, pp. 511–518, 2019.

[9] L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *arXiv:2003.09871*, 2020.

[10] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[11] B. Chen, R. Zhang, Y. Gan, L. Yang, and W. Li, "Development and clinical application of radiomics in lung cancer," *Radiation Oncology*, vol. 12, no. 1, p. 154, 2017.

[12] Y. Zhang, H. Jiang, Y. Miura *et al.*, "Contrastive learning of medical visual representations from paired images and text," *arXiv:2010.00747*, 2020.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv:2002.05709*, 2020.

[14] G. Shih, C. C. Wu, S. S. Halabi, M. D. Kohli, L. M. Prevedello, T. S. Cook, A. Sharma, J. K. Amorosa, V. Arteaga, M. Galperin-Aizenberg *et al.*, "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia," *Radiology: Artificial Intelligence*, vol. 1, no. 1, p. e180041, 2019.

[15] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *CVPR*, 2017, pp. 3156–3164.

[16] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[17] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J.

Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.

[18] E. J. Topol, *Deep medicine: how artificial intelligence can make healthcare human again*, first edition ed. New York: Basic Books, 2019.

[19] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter, and C. S. Greene, "Opportunities and obstacles for deep learning in biology and medicine." *Journal of the Royal Society, Interface*, vol. 15, no. 141, Apr. 2018.

[20] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3462–3471.

[21] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2018, pp. 9049–9058.

[22] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, and H. Marklund, "CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597.

[23] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," *arXiv preprint*, Jan. 2019.

[24] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest x-ray models," *arXiv preprint arXiv:2010.05352*, 2020.

[25] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv preprint arXiv:1710.10501*, 2017.

[26] S. Gündel, S. Grbic, B. Georgescu, S. Liu, A. Maier, and D. Comaniciu, "Learning to recognize abnormalities in chest x-rays with location-aware dense networks," in *Progress in pattern recognition, image analysis, computer vision, and applications*. Springer International Publishing, 2019, pp. 757–765.

[27] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei, "Thoracic disease identification and localization with limited supervision," in

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8290–8299.

[28] S. Hwang and H.-E. Kim, "Self-transfer learning for fully weakly supervised object localization," *arXiv preprint arXiv:1602.01625*, 2016.

[29] R. N. Bryan, Ed., *Introduction to the science of medical imaging*. Cambridge University Press, Jan. 2001.

[30] M. Nicolasjilwan, Y. Hu, C. Yan, D. Meerzaman, C. A. Holder, D. Gutman, R. Jain, R. Colen, D. L. Rubin, P. O. Zinn, S. N. Hwang, P. Raghavan, D. A. Hammoud, L. M. Scarpace, T. Mikkelsen, J. Chen, O. Gevaert, K. Buetow, J. Freymann, J. Kirby, A. E. Flanders, and M. Wintermark, "Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients," *Journal of Neuroradiology*, vol. 42, no. 4, pp. 212–221, Jul. 2015.

[31] B. Ganeshan, S. Abaleke, R. C. Young, C. R. Chatwin, and K. A. Miles, "Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage," *Cancer Imaging*, vol. 10, no. 1, pp. 137–143, 2010.

[32] B. Ganeshan, V. Goh, H. C. Mandeville, Q. S. Ng, P. J. Hoskin, and K. A. Miles, "Non–small cell lung cancer: Histopathologic correlates for texture parameters at CT," *Radiology*, vol. 266, no. 1, pp. 326–336, Jan. 2013.

[33] V. S. Parekh and M. A. Jacobs, "Deep learning and radiomics in precision medicine," *Expert Review of Precision Medicine and Drug Development*, vol. 4, no. 2, pp. 59–72, Mar. 2019.

[34] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi, "Clinically accurate chest x-ray report generation," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 249–269.

[35] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 107–117.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2921–2929.

[38] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: convolutional triplet attention module," *arXiv:2010.03045 [cs]*, Nov. 2020.

[39] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[40] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu, "NegBio: a high-performance tool for negation and uncertainty detection in radiology reports." in *AMIA Joint Summits on Translational Science proceedings*, vol. 2017, 2018, pp. 188–196.

[41] P. Kumar, M. Grewal, and M. M. Srivastava, "Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs," in *The International Conference on Computer Vision (ICCV)*, 2018, pp. 546–552.

[42] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738.

[43] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[44] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," *arXiv preprint arXiv:2006.10029*, 2020.

[45] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6707–6717.

[46] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your

own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[47] Y. Yang and Z. Xu, "Rethinking the value of labels for improving class-imbalanced learning," *arXiv preprint arXiv:2006.07529*, 2020.

[48] B. Kang, Y. Li, Z. Yuan, and J. Feng, "Exploring balanced feature spaces for rep-resentation learning."

[49] X. Chen, K. Oshima, D. Schott, H. Wu, W. Hall, Y. Song, Y. Tao, D. Li, C. Zheng, P. Knechtges, B. Erickson, and X. A. Li, "Assessment of treatment response during chemoradiation therapy for pancreatic cancer based on quantitative radiomic analysis of daily CTs: An exploratory study," *PLOS ONE*, vol. 12, no. 6, p. e0178961, Jun. 2017.

[50] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that?" *arXiv preprint arXiv:1611.07450*, 2016.

[51] L. Gao, L. Zhang, C. Liu, and S. Wu, "Handling imbalanced medical image data: A deep-learning-based one-class classification approach," *Artificial Intelligence in Medicine*, vol. 108, p. 101935, 2020.

[52] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[53] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," *arXiv preprint arXiv:2010.04592*, 2020.

[54] K. Pasupa, S. Vatathanavaro, and S. Tungjitnob, "Convolutional neural networks based focal loss for class imbalance problem: A case study of canine red blood cells morphology classification," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–17, 2020.

[55] P. Kumar, M. Grewal, and M. M. Srivastava, "Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 546–552.

[56] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 632–10 641.

[57] L. Seyyed-Kalantari, G. Liu, M. McDermott, and M. Ghassemi, "Chexclusion: Fairness gaps in deep chest x-ray classifiers," *arXiv preprint arXiv:2003.00827*, 2020.

[58] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative," *arXiv preprint arXiv:1612.07003*, 2016.

[59] V. Parekh and M. A. Jacobs, "Radiomics: a new application from established techniques," *Expert Review of Precision Medicine and Drug Devel-*

opment, vol. 1, no. 2, pp. 207–226, 2016.

[60] F. Shi, L. Xia, F. Shan *et al.*, "Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification. arxiv e-prints [preprint] 2020."

[61] A. Saygılı, "A new approach for computer-aided detection of coronavirus (covid-19) from ct and x-ray images using machine learning methods," *Applied Soft Computing*, vol. 105, p. 107323, 2021.

[62] X. Bai, C. Fang, Y. Zhou, S. Bai, Z. Liu, L. Xia, Q. Chen, Y. Xu, T. Xia, S. Gong *et al.*, "Predicting covid-19 malignant progression with ai techniques," 2020.

[63] B. Ghosh, N. Kumar, N. Singh, A. K. Sadhu, N. Ghosh, P. Mitra, and J. Chatterjee, "A quantitative lung computed tomography image feature for multi-center severity assessment of covid-19," *medRxiv*, 2020.

[64] E. Rozenberg, D. Freedman, and A. Bronstein, "Localization with limited annotation for chest x-rays," in *Machine Learning for Health Workshop*. PMLR, 2020, pp. 52–65.

[65] Y. Wang, K. Zheng, C.-T. Cheng, X.-Y. Zhou, Z. Zheng, J. Xiao, L. Lu, C.-H. Liao, and S. Miao, "Knowledge distillation with adaptive asymmetric label sharpening for semi-supervised fracture detection in chest x-rays," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 599–610.

[66] K. Yu, S. Ghosh, Z. Liu, C. Deible, and K. Batmanghelich, "Anatomy-guided weakly-supervised abnormality localization in chest x-rays," *arXiv preprint arXiv:2206.12704*, 2022.

[67] T. B. Chandra, B. K. Singh, and D. Jain, "Disease localization and severity assessment in chest x-ray images using multi-stage superpixels classification," *Computer Methods and Programs in Biomedicine*, p. 106947, 2022.

[68] C. Fernando, S. Kolonne, H. Kumarasinghe, and D. Meedeniya, "Chest radiographs classification using multi-model deep learning: A comparative study," in *2022 2nd International Conference on Advanced Research in Computing (ICARC)*. IEEE, 2022, pp. 165–170.

[69] K. Kumarasinghe, S. Kolonne, K. Fernando, and D. Meedeniya, "U-net based chest x-ray segmentation with ensemble classification for covid-19 and pneumonia." *International Journal of Online & Biomedical Engineering*, vol. 18, no. 7, 2022.

[70] D. Meedeniya, H. Kumarasinghe, S. Kolonne, C. Fernando, I. De la Torre Díez, and G. Marques, "Chest x-ray analysis empowered with deep learning: A systematic review," *Applied Soft Computing*, p. 109319, 2022.

[71] L. Jing, Y. Chen, and Y. Tian, "Coarse-to-fine semantic segmentation from image-level labels," *IEEE Transactions on Image Processing*, vol. 29, pp. 225–236, 2019.

[72] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[73] ——, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[74] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern recognition*, vol. 65, pp. 211–222, 2017.

[75] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019, pp. 1–5.

[76] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[77] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[78] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*

preprint arXiv:1810.04805, 2018.

[79] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[80] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[81] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[82] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu, "Do transformers really perform bad for graph representation?" *arXiv preprint arXiv:2106.05234*, 2021.

[83] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," 2021.

[84] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[85] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. Feris, D. Harwath, J. Glass, and H. Kuehne, "Everything at once–

multi-modal fusion transformer for video retrieval," *arXiv preprint arXiv:2112.04446*, 2021.

[86] C.-F. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," *arXiv preprint arXiv:2103.14899*, 2021.

[87] X. Yue, S. Sun, Z. Kuang, M. Wei, P. Torr, W. Zhang, and D. Lin, "Vision transformer with progressive sampling," *arXiv preprint arXiv:2108.01684*, 2021.

[88] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[89] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021.

[90] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[91] Y. Han, C. Chen, L. Tang, M. Lin, A. Jaiswal, S. Wang, A. Tewfik, G. Shih, Y. Ding, and Y. Peng, "Using radiomics as prior knowledge

for thorax disease classification and localization in chest x-rays," *arXiv preprint arXiv:2011.12506*, 2020.

[92] FawcettTom, "An introduction to roc analysis," *Pattern Recognition Letters*, 2006.

Index

Vita

Yan Han is a Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of Texas at Austin. He earned his B.Eng. from the School of Electronic, Information, and Electrical Engineering at Shanghai Jiao Tong University in 2017, followed by an M.S. in Electrical and Computer Engineering from the University of Texas at Austin in 2019. He has gained valuable industry experience through internships, working as a Machine Learning Engineer Intern at LinkedIn from May to December 2021, and as an Applied Scientist Intern at Amazon from February to December 2022. His research interests encompass medical image analysis, graph neural networks, and a broad range of machine learning and deep learning methodologies.

Permanent address: yh9442@utexas.edu

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.