Mathematical Biosciences
and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(12): 21292–21314.
DOI: 10.3934/mbe.2023942
Received: 26 October 2023
Revised: 22 November 2023
Accepted: 23 November 2023
Published: 30 November 2023

Research article

MRChexNet: Multi-modal bridge and relational learning for thoracic disease recognition in chest X-rays

Guoli Wang[1,2], **Pingping Wang**[1,2,*], **Jinyu Cong**[1,2] **and Benzheng Wei**[1,2,*]

[1] Center for Medical Artificial Intelligence, Shandong University of Traditional Chinese Medicine, Qingdao 266112, China

[2] Qingdao Academy of Chinese Medical Sciences, Shandong University of Traditional Chinese Medicine, Qingdao 266112, China

* **Correspondence:** Email: wangpingping@sdutcm.edu.cn, wbz99@sina.com.

Abstract: While diagnosing multiple lesion regions in chest X-ray (CXR) images, radiologists usually apply pathological relationships in medicine before making decisions. Therefore, a comprehensive analysis of labeling relationships in different data modes is essential to improve the recognition performance of the model. However, most automated CXR diagnostic methods that consider pathological relationships treat different data modalities as independent learning objects, ignoring the alignment of pathological relationships among different data modalities. In addition, some methods that use undirected graphs to model pathological relationships ignore the directed information, making it difficult to model all pathological relationships accurately. In this paper, we propose a novel multi-label CXR classification model called MRChexNet that consists of three modules: a representation learning module (RLM), a multi-modal bridge module (MBM) and a pathology graph learning module (PGL). RLM captures specific pathological features at the image level. MBM performs cross-modal alignment of pathology relationships in different data modalities. PGL models directed relationships between disease occurrences as directed graphs. Finally, the designed graph learning block in PGL performs the integrated learning of pathology relationships in different data modalities. We evaluated MRChexNet on two large-scale CXR datasets (ChestX-Ray14 and CheXpert) and achieved state-of-the-art performance. The mean area under the curve (AUC) scores for the 14 pathologies were 0.8503 (ChestX-Ray14) and 0.8649 (CheXpert). MRChexNet effectively aligns pathology relationships in different modalities and learns more detailed correlations between pathologies. It demonstrates high accuracy and generalization compared to competing approaches. MRChexNet can contribute to thoracic disease recognition in CXR.

Keywords: multi-label chest X-ray recognition; cross-modal fusion; pathology correlation; relational learning

1. Introduction

Thoracic diseases are diverse and imply complex relationships. For example, extensive clinical experience [1, 2] has demonstrated that pulmonary atelectasis and effusion often lead to infiltrate development, and pulmonary edema often leads to cardiac hypertrophy. This strong correlation between pathologies, known as label co-occurrence, is a common phenomenon in clinical diagnosis and is not coincidental [3], as shown in Figure 1. Radiologists need to look at the lesion area at the time of diagnosis while integrating the pathologic relationships to arrive at the most likely diagnosis. Therefore, diagnosing a massive number of Chest X-ray (CXR) images is a time-consuming and laborious reasoning task for radiologists. This has inspired researchers to utilize deep learning techniques to automatically analyze CXR images and reduce the workloads of radiologists. Multiple abnormalities may be present simultaneously in a single CXR image, making the clinical chest radiograph examination a classic multi-label classification problem. Multi-label classification means that a sample can belong to multiple categories (or labels) and that different categories are related. Relationships between pathology labels are expressed differently in different data modalities. As Figure 1 shows, pathology regions appearing simultaneously in the image reflect label relationships as features. In the word embedding of pathology labels, the label relationship is implicit in the semantic information of each label. In recent years, several advanced deep learning methods have been developed to solve this task [4–9]. According to our survey, the existing methods are divided into two classes: 1) label-independent learning methods and 2) label-correlation learning methods.

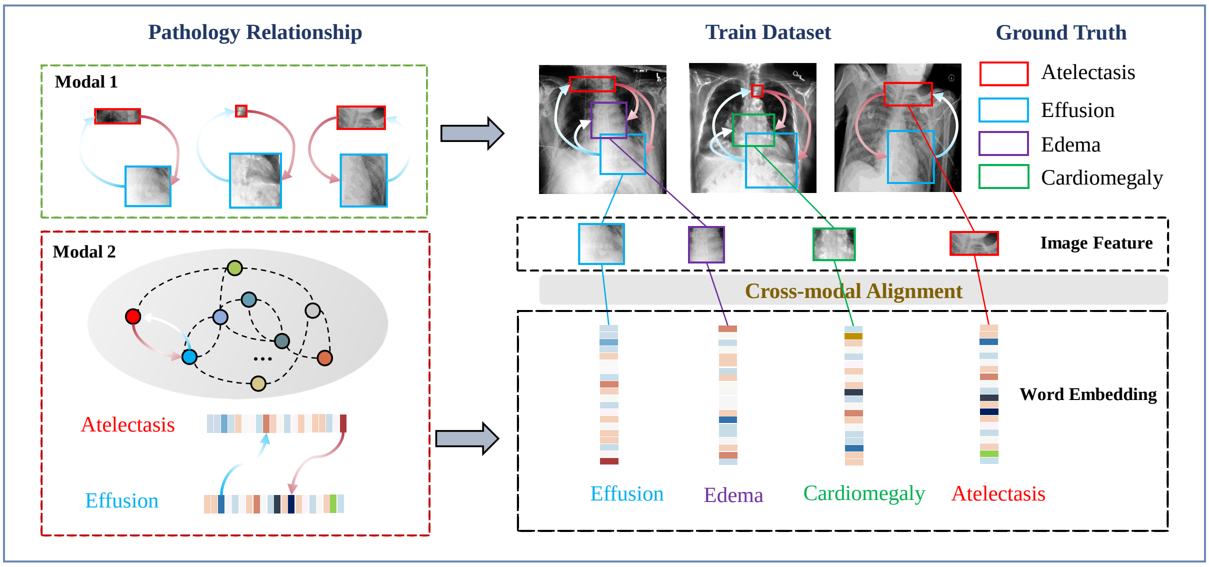


Figure 1. Illustration of pathology relationships and alignment problems in different data modals. Left: the pathology correlation within each modal. Right: we aligned the representation of pathology across modals. The transformed arrows in the figure indicate that "Pathology A → Pathology B" means that when Pathology A appears, Pathology B is likely to have occurred, but the converse does not necessarily hold.

The label-independent learning method transforms the multi-label CXR recognition task into multiple independent nonintersecting binary recognition tasks. The primary process is to train a separate binary classifier for each label on the sample to be tested. Early on, some researchers [2, 10–12] used convolutional neural networks and their variants on this task with some success by designing elaborate network structures to improve recognition accuracy. Despite their efforts and breakthroughs in this field, some things can still be improved. Since this label-independent learning method treats each label as an independent learning object, training results are susceptible to situations, such as missing sample labels and sample mislabeling. Additionally, this class of methods uses only the sample image as the main carrier of the learning object. The image as a single modal form of labeling relationships implies a particular limitation. These methods have yet to consider interlabel correlations and ignore the representation of labeling relationships in other data modalities.

Subsequently, clinical experience has shown that some abnormalities in CXR images may be strongly correlated. The literature [3] suggests that this is not a coincidence but rather one of a labeling relationship that can be called co-occurrence. The literature [1] found that edema in the lungs tends to trigger cardiomegaly. The literature [2] indicates that lung infiltrates are often associated with pulmonary atelectasis and effusion. This labeling relationship inspires the application of deep learning techniques to the CXR recognition task. In addition, this interdependent information can be used to infer missing or noisy labels from co-occurrence relationships. This improves the robustness of the model and its recognition performance.

Existing label-correlation learning methods are mainly categorized into two types: image-based unimodal learning methods and methods that additionally consider textual modal data while learning images. First, the most common technique in image-based unimodal learning methods is attention-guided. These attention-guided methods [13–15] focus on the most discriminating lesion area features in each sample CXR image. These methods capture the interdependence between labels and lesion regions implicitly, i.e., by designing attention models with different mechanisms to establish the correlation between lesion regions and the whole region. However, the above methods only locally establish label correlations on the imaging modality, ignoring the global label co-occurrence relationship. Another approach that considers textual modal data when learning images is categorized as Recurrent Neural Network (RNN)-based and Graph Convolutional Network (GCN)-based. These RNN-based methods [1, 16, 17] rely on state variables to encode label-related information and use the RNN as a decoder to predict anomalous sequences in sample images. However, this approach often requires complex computations. In addition, some researchers [18, 19] extract valuable textual embedding information from radiology reports to assist in classification. In contrast, GCN-based methods [6, 20–22] represent label-correlation information, such as label co-occurrence as undirected graph data. These methods treat each label as a graph node and use semantic word embeddings of labels as node features. However, while the above methods learn the label relations in additional modalities, they ignore the alignment between the label relation representations of different modalities, as shown on the right side of Figure 1. Moreover, these methods of modeling pathological relationships using graphs are composed so that the directed graph information is ignored, i.e., it is difficult to represent all pathological relationships in an undirected graph accurately.

In this paper, we propose a multi-label CXR classification model called MRChexNet that integrally learns pathology information in different modalities and models interpathology correlations more comprehensively. It consists of a representation learning module (RLM), a multi-modal bridge module

(MBM), and a pathology graph learning module (PGL). In RLM, we obtain image-level pathology-specific representations for lesion regions in every image. In MBM, we fully bridge the pathology representations in different modalities. The image-level pathology-specific representations from RLM align with the rich semantic information in pathology word embeddings. In PGL, we first model the undirected graph pathology correlation matrix containing all pathology relations in a data-driven manner. Second, by considering the directed information between nodes, we construct an in-degree matrix and an out-degree matrix as directed graphs by considering the out-degree and in-degree on each node as the study object, respectively. Finally, we designed a graph learning module in PGL that integrates the study of pathological information in multiple modalities. The front end of the module is designed with a graph convolution block with a two-branch symmetric structure for learning two directed graphs containing labeling relations in different directions. The back end of the module stacks graph attention layers. All labeling relations are comprehensively learned on the undirected graph pathology correlation matrix. Finally, the framework is optimized using a multi-label loss function to complete end-to-end training.

In summary, our contributions are fourfold:

1) A new RLM is proposed to obtain image-level pathology-specific representation and global image representation for image lesion regions.

2) A novel MBM is proposed that aligns pathology information in different modal representations.

3) In the proposed PGL, more accurate pathological relationships are modeled as directed graphs by considering directed information between nodes on the graph. An effective graph learning block is designed to learn the pathology information of different modalities comprehensively.

4) We developed the framework in two large-scale CXR datasets (ChestX-ray14 [2] and CheXpert [23]) and evaluated the effectiveness of MRChexNet on this basis, with average AUC scores of 0.8503 and 0.8649 for 14 pathologies. Our method achieves state-of-the-art performance in terms of classification accuracy and generalizability.

2. Related work

This section presents a summary of the relevant literature in two aspects. First, previous works on the automatic analysis of CXR images are introduced. Second, several representative works related to cross-modal fusion are presented.

2.1. Multi-label chest X-ray image recognition

To improve efficiency and reduce the workloads of radiologists, researchers are beginning to apply the latest advances in deep learning to chest X-ray analysis. In the early days of deep learning techniques applied to CXR recognition, researchers divided the CXR multi-label recognition task into multiple independent disjoint binary labeling problems. An independent binary classifier is trained for each anomaly present in the image. Wang et al. [2] used classical convolutional neural networks and transfer learning to predict CXR images. Rajpurkar et al. [10] improved the network architecture based on DenseNet-121 [11] and proposed CheXNet for anomaly classification in CXR images, which achieved good performance in detecting pneumonia. Li et al. [24] performed thoracic disease identification and localization with additional location annotation supervision. Shen et al. [12] designed a novel network training mechanism for efficiently training CNN-based automatic chest disease detection models. To dynamically capture more discriminative features for thoracic disease classification, Chen et al. [25] used a dual asymmetric architecture based on ResNet and DenseNet. However, as mentioned above, these methods do not account for the correlation between the labels.

When diagnosing, the radiologist needs to view the lesion area while integrating pathological relationships to make the most likely diagnosis. This necessity inspired researchers to start considering label dependencies. For example, Wang et al. [16] used RNN to model label relevance sequentially. Yao et al. [1] considered multi-label classification as a sequence prediction task with a fixed length. They employed long short-term memory (LSTM) [26] and presented initial results indicating that utilizing label dependency can enhance classification performance. Ypsilantis et al. [17] used an RNN-based bidirectional attention model that focuses on information-rich regions of an image and samples the entire CXR image sequentially. Moreover, some approaches have attempted to use different attentional mechanisms to correlate labels with attended areas. The work of Zhu et al. [13] and Wang et al. [14] both use an attention mechanism that only addresses a limited number of local correlations between regions on an image. Guan et al. [15] used CNNs to learn high-level image features and designed attention-learning modules to provide additional attention guidance for chest disease recognition. It is worth mentioning that as the graph data structure has become a hot research topic, some approaches use graphs to model labeling relationships. Subsequently, Chen et al. [22] introduced a workable framework in which every label represents a node, the term vector of each label acts as a node feature, and GCN is implemented to comprehend the connection among labels in an undirected graph. Li et al. [27] developed the A-GCN, which captures label dependencies by creating an adaptive label structure and has demonstrated exemplary performance. Lee et al. [20] described label relationships using a knowledge graph, which enhances image representation accuracy. Chen et al. [6] employed an undirected graph to represent the relationships between pathologies. They designed CheXGCN by using the word vectors of labels as node features of the graph, and the experiments showed promising results.

2.2. *Cross-modal fusion*

Researchers often use concatenation or elemental summation to fuse different modal features to fuse cross-modal features. Fukui et al. [28] proposed that two vectors of different modalities are made exterior product to fuse multi-modal features by bilinear models. However, this method yields high-dimensional fusion vectors. Hu et al. [29] used data within 24 hours of admission to build simpler machine-learning models for early acute kidney injury (AKI) risk stratification and obtained good results. Xu et al. [30] encouraged data on both attribute and imaging modalities to be discriminated to improve attribute-image person reidentification. To reduce the high-dimensional computation, Kim et al. [31] designed a method that achieves comparable performance to the work of Fukui et al. by performing the Hadamard product between two feature vectors but with slow convergence. It is worth mentioning that Zhou et al. [32] introduced a new method with stable performance and accelerated model convergence for the study of fusing image features and text embedding. Chen et al. [22] used ResNet to learn the image features, GCN to learn the semantic information in the label word embeddings, and finally fused the two using a simple dot product. Similarly, Wang et al. [33] designed a sum-pooling method to fuse the vectors of the two modalities after learning the image features and the semantic information of label word embeddings. It not only reduces the dimensionality of the vectors but also increases the convergence rate of the model.

3. Materials and methods

This section proposes a multi-label CXR recognition framework, MRChexNet, consisting of three main modules: the representation learning module (RLM), multi-modal bridge module (MBM), and pathology graph learning module (PGL). We first introduce the general framework of our model in Figure 2 and then detail the workflow of each of these three modules. Finally, we describe the datasets implementation details, and evaluation metrics.

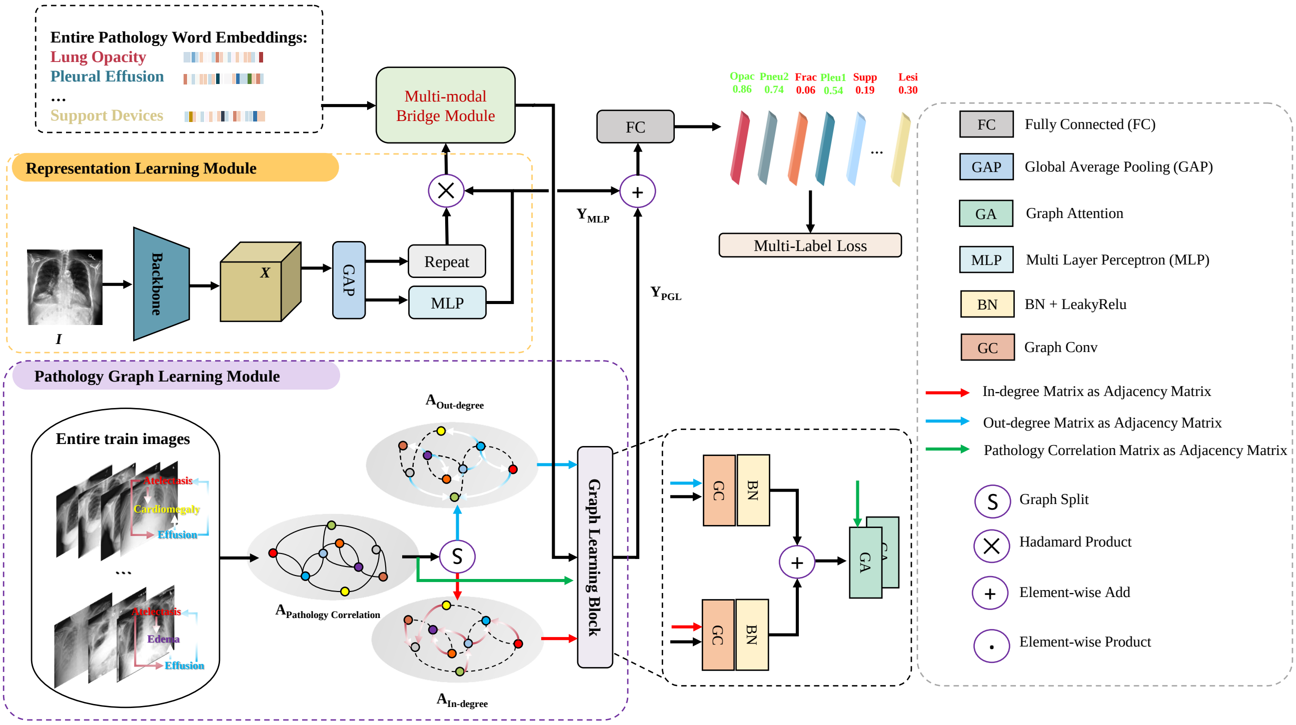


Figure 2. The overall framework of our proposed MRChexNet.

3.1. Representation learning module

Theoretically, we can use any CNN-based model to learn image features. In our experiments, following [1, 6, 25], we use DenseNet-169 [11] as the backbone for fair comparisons. Thus, if an input image I has a 224 × 224 resolution, we can obtain 1664 × 7 × 7 feature maps from the "Dense Block_4" layer of DenseNet-169. As shown in Figure 2, we perform global average pooling to obtain the image-level global feature $\boldsymbol{x} = f_{GAP}(f_{backbone}(\mathrm{I}))$, where $f_{GAP}(\cdot)$ represents the global average pooling (GAP) [34] operation. We first set up a multi layer perceptron (MLP) layer learning x to obtain an initial diagnostic score of the image, Y_{MLP}. Specifically, the MLP here consists of a layer of fully connected (FC) network + sigmoid activation function.

$$Y_{MLP} = f_{MLP}(x; \theta_{MLP}), \tag{3.1}$$

where $f_{MLP}(\cdot)$ represents the MLP layer and $\theta_{MLP} \in \mathbb{R}^{C\times D}$ is the parameter. We use the parameter θ_{MLP} as a diagnoser for each disease and filter a set of features specific to a disease from the global feature x. Each diagnoser $\theta^{C}_{MLP} \in \mathbb{R}^{D}$ extracts information related to disease C and predicts the likelihood of the appearance of disease C in the image. Then, the pathology-related feature F_{pr} is disentangled by Eq (3.2).

$$F_{pr} = f_{repeat}(x) \odot \theta_{MLP}. \tag{3.2}$$

The operation $f_{repeat}(\cdot)$ indicates that $x \in \mathbb{R}^{D}$ is copied C times to form $[X, \cdots X]^{T} \in \mathbb{R}^{C\times D}$, with $\odot$ denoting the Hadamard product. Using this method to adjust the global feature x, the adjusted x captures more relevant information for each disease.

3.2. Multi-modal bridge module

In this section, we design the MBM module to efficiently align the disease's image features and the disease's semantic word embeddings. As Figure 3 shows, the MBM module is divided into two phases: alignment + fusion and squeeze. The fixed input of the MBM module consists of two parts: $modal_1 \in \mathbb{R}^{D_1}$, which represents the image features, and $modal_2 \in \mathbb{R}^{D_2}$, which is the word embedding. First, we use two FC layers to convert $modal_1$ into $M_1 \in \mathbb{R}^{D_3}$ and $modal_2$ into $M_2 \in \mathbb{R}^{D_3}$, respectively:

$$\begin{cases} \boldsymbol{M_1} = FC_1(\boldsymbol{modal_1}) \in \mathbb{R}^{D_3} \\ \boldsymbol{M_2} = FC_2(\boldsymbol{modal_2}) \in \mathbb{R}^{D_3} \end{cases}. \tag{3.3}$$

We design a separate dropout layer for M_2 to prevent redundant semantic information from causing overfitting. After obtaining two inputs M_1, M_2 of the same dimension, the initial bilinear pooling [35] is defined as follows:

$$\boldsymbol{F} = \boldsymbol{M_1^T S_i M_2}, \tag{3.4}$$

where $\boldsymbol{F} \in \mathbb{R}^{o}$ is the output fusion feature of the MBM module and $\boldsymbol{S_i} \in \mathbb{R}^{D_3\times D_3}$ is the bilinear mapping matrix with bias terms included. $\boldsymbol{S} = [S_i, \cdots, S_o] \in \mathbb{R}^{D_3\times D_3\times o}$ can be decomposed into two low-rank matrices $\boldsymbol{u_i} = [u_1, \cdots, u_G] \in \mathbb{R}^{D_3\times G}$, $\boldsymbol{v_i} = [v_1, \cdots, v_G] \in \mathbb{R}^{D_3\times G}$. Therefore, Equation (3.4) can be rewritten as follows:

$$\boldsymbol{F_i} = \boldsymbol{1^T}\left(\boldsymbol{u_i^T M_1} \circ \boldsymbol{v_i^T M_2}\right), \tag{3.5}$$

where the value of G is the factor or latent dimension of two low-rank matrices and $1^{T} \in \mathbb{R}^{G}$ is an all-one vector. To obtain the final $\boldsymbol{F}$, two three-dimensional tensors $\boldsymbol{u_i} \in \mathbb{R}^{D_3\times G\times o}, \boldsymbol{v_i} \in \mathbb{R}^{D_3\times G\times o}$ need to be learned. Under the premise of ensuring the generality of Eq (3.5), the two learnable tensors $\boldsymbol{u}, \boldsymbol{v}$ are converted into two-dimensional matrices by matrix variable dimension, namely, $\boldsymbol{u_i} \rightarrow \boldsymbol{\tilde{u}} \in \mathbb{R}^{D_3\times Go}$ and $\boldsymbol{v_i} \rightarrow \boldsymbol{\tilde{v}} \in \mathbb{R}^{D_3\times Go}$, then Eq (3.5) simplifies to:

$$\boldsymbol{F} = f_{GroupSum}\left(\boldsymbol{\tilde{u}^T M_1} \circ \boldsymbol{\tilde{v}^T M_2}, G\right), \tag{3.6}$$

where the function $f_{GroupSum}(\boldsymbol{vector}, G)$ represents the mapping of g elements in $\boldsymbol{vector}$ into $\frac{1}{G}$ groups and outputs all G groups obtained after complete mapping as potential dimensions, $\boldsymbol{F} \in \mathbb{R}^{G}$. Furthermore, a dropout layer is added after the elementwise multiplication layer to avoid overfitting. Due to the introduction of elementary multiplication, the size of the output neuron can change drastically, and the model can converge to a local minimum that is not satisfactory. Therefore, the normalization layer ($\boldsymbol{F} \leftarrow \boldsymbol{F}/||\boldsymbol{F}||$) and power normalization layer ($\boldsymbol{F} \leftarrow \text{sign}(\boldsymbol{F})|\boldsymbol{F}|^{0.5}$) are appended. Finally, $\boldsymbol{F}$ is copied C times through operation $f_{Repeat}(\cdot)$, then $\boldsymbol{F} \in \mathbb{R}^{C \times G}$ as the final MBM output. These are the details of the MBM process.

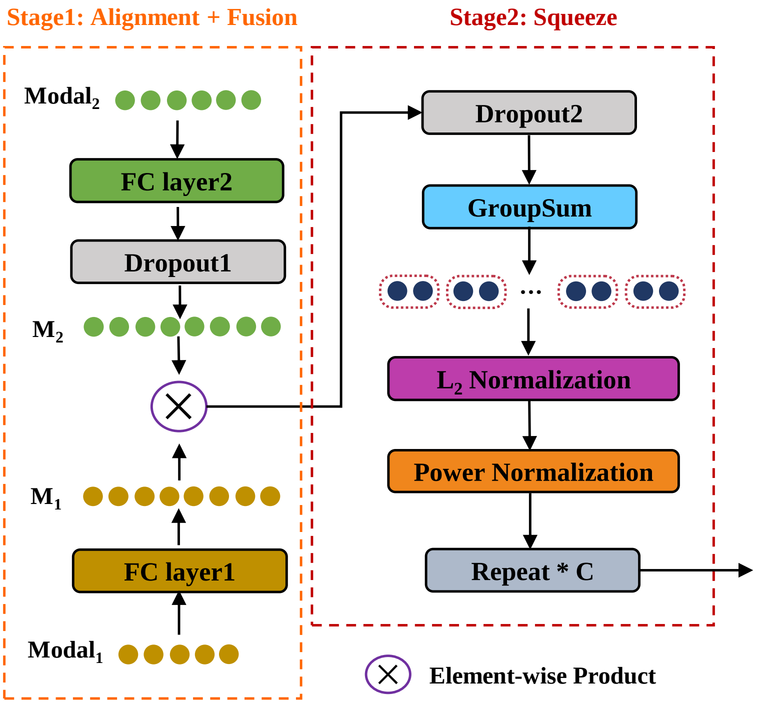


Figure 3. Architecture of multi-modal bridge module.

3.3. Pathology graph learning module

Our PGL module is built on top of graph learning. The node-level output of traditional graph learning techniques is the predicted score of each node. In contrast, the final output of our designed graph learning block is designed as the classifier for the corresponding label in our task. We use the fused features of the MBM output as the node features for graph learning. Furthermore, the graph structure (i.e., the correlation matrix) is typically predefined in other tasks. However, it is not provided in the multi-label CXR image recognition task. We need to construct the correlation matrix ourselves. Therefore, we devise a new method for constructing the correlation matrix by considering the directed information of graph nodes.

First, we capture the pathological dependencies based on the label statistics of the entire dataset and construct the pathology correlation matrix A_{pc}. Specifically, we count the number of occurrences (T_i) of the i-th pathological label (L_i) and the simultaneous occurrences of L_i and L_j (T_{ij}=T_{ji}). In addition, the label dependency can be expressed by conditional probability as follows:

$$P_{ij} = P\left(L_i|L_j\right) = \frac{T_{ij}}{T_j}, \forall i \in [1, C], \tag{3.7}$$

where P_{ij} denotes the probability that L_i occurs under the condition that L_j occurs. Note that since the conditional probabilities between two objects are asymmetric, $P_{ij} \neq P_{ji}$. The element value $A_{pc_{ij}}$ at each position in this matrix is equal to P_{ij}. Then, by considering directed information on the graph structure, we split an in-degree matrix $\boldsymbol{A}_{pc}^{in}$ and an out-degree matrix $\boldsymbol{A}_{pc}^{out}$, which are defined as follows:

$$\boldsymbol{A}_{pc}^{in} = \sum_{k} \frac{A_{pc_{ki}} A_{pc_{kj}}}{\sum_{v} A_{pc_{kv}}}, \forall i, j \in C, k, v \in C, \tag{3.8}$$

$$\boldsymbol{A}_{pc}^{out} = \sum_{k} \frac{A_{pc_{ik}} A_{pc_{jk}}}{\sum_{v} A_{pc_{vk}}}, \forall i, j \in C, k, v \in C. \tag{3.9}$$

Then, in our PGL, the dual-branch learning of the graph learning block is specifically defined as:

$$\boldsymbol{Z^{in}} = f_{gc}^{in}(A_{pc}^{in} \boldsymbol{F} \theta_{gc}^{in}), \tag{3.10}$$

$$\boldsymbol{Z^{out}} = f_{gc}^{out}(A_{pc}^{out} \boldsymbol{F} \theta_{gc}^{out}), \tag{3.11}$$

where $\boldsymbol{Z^{in}}$ and $\boldsymbol{Z^{out}}$ are the outputs of the in-degree branch and the out-degree branch, respectively. $f_{gc}(\cdot)$ denotes the graph convolutional operation, and θ_{gc} denotes the corresponding trainable transformation matrix.

To learn more about the correlations between different pathological features, we use a graph attention network (GAT) [36] to consider $\boldsymbol{Z^{in}}$ and $\boldsymbol{Z^{out}}$ jointly. We do this by using $\boldsymbol{Z^{all}} = f^{'}(\boldsymbol{Z^{in}}) + f^{'}(\boldsymbol{Z^{out}})$ as the input feature to graph attention. $f^{'}(\cdot)$ denotes the batch normalization layer and nonlinear activation operation LeakyReLU. The graph attention layer transforms the implicit features of the input nodes and aggregates the neighborhood information to the next node to improve the correlation between the information of the central node and its neighbors. The input $\boldsymbol{Z^{all}}$ to the graph attention layer is the set of node features $\left\{\boldsymbol{Z_1^{all}}, \boldsymbol{Z_2^{all}}, \cdots, \boldsymbol{Z_n^{all}}\right\} \in \mathbb{R}^d$, where d is the number of feature dimensions in each node. The attention weight coefficients $e_{i,j}$ are computed between node i and the neighborhood of node $j \in NB_i$ by a learnable linear transformation matrix W and applied to all nodes, as shown in Eq (3.12).

$$e_{i,j} = a\left[WX_i || WX_j\right], \tag{3.12}$$

where $||$ is the concatenation operation, $W \in \mathbb{R}^{\acute{d} \times d}$, $a \in \mathbb{R}^{\acute{d} \times d}$ is a learnable parameter and $\acute{d}$ denotes the dimensionality of the output features. The graph attention layer allows each node to focus on each of the other nodes. $e_{i,j}$ uses LeakyReLU as the nonlinear activation function and is normalized by the sigmoid function, which can be expressed as:

$$\alpha_{i,j} = Sigmoid_j\left(e_{i,j}\right) = \frac{\exp\left(LeakyReLU\left(e_{i,j}\right)\right)}{\sum_{k \in NB_i} \exp\left(LeakyReLU\left(e_{i,k}\right)\right)}. \tag{3.13}$$

To stabilize the learning process of the graph attention in the PGL module, we extended the multi-headed self-attention mechanism within it as follows:

$$Y_{PGL} = ||_{k=1}^{K} ReLU\left(\alpha^{(k)} \boldsymbol{Z^{all}} \boldsymbol{W^k}\right), \tag{3.14}$$

where $Y_{PGL} \in \mathbb{R}^{K\acute{D}}$ denotes the output features incorporating the pathology-correlated features, K denotes the number of attention heads, and $\alpha^{(k)}$ denotes the normalized k-th attention weight coefficient matrix. W^k denotes the transformable weight matrix under the corresponding k-th attention head. Finally, the output features are averaged and passed to the next node.

$$Y_{PGL} = ReLU\left(\frac{1}{K}\right)\sum_{K=1}^{K}\left(\alpha^{(k)}\boldsymbol{Z}^{\boldsymbol{all}}\boldsymbol{W}^{\boldsymbol{k}}\right). \tag{3.15}$$

We show through empirical studies that PGL can detect potentially strong correlations between pathological features. It improves the model's ability to learn implicit relationships between pathologies.

After obtaining Y_{MLP} and Y_{PGL}, we set the final output of our model as $Y_{Out} = Y_{MLP} + Y_{PGL}$ and then feed it into the loss function to calculate the loss. Finally, we update the entire network end-to-end using the MultiLabelSoftMargin loss (called multi-label loss) function [37]. The training loss function is described as:

$$\begin{aligned}\mathcal{L}(Y_{Out}, L) = &-\frac{1}{C}\sum_{j=1}^{C} L_j \log\left(\left(1+\exp\left(-Y_{out_j}\right)\right)^{-1}\right)\\ &+\left(1-L_j\right)\log\left(\frac{\exp\left(-Y_{outj}\right)}{\left(1+\exp\left(-Y_{out_j}\right)\right)}\right),\end{aligned} \tag{3.16}$$

where Y_{Out} and L denote the predicted pathology and the true pathology of the sample image, respectively. Y_{out_j} and L_j denote the j-th element in its predicted pathology and the j-th element in the actual pathology.

4. Experiments

In this section, we report and discuss the results on two benchmark multi-label CXR recognition datasets. Ablation experiments were also conducted to explore the effects of different parameters and components on MRChexNet. Finally, a visual analysis was performed.

4.1. Datasets

ChestX-Ray14 is a large CXR dataset. It contains 78,466 training images, 11,220 validation images, and 22,434 test images. Approximately 1.6 pathology labels from 14 semantic categories are applied to the patient images. Each image is labeled with one or more pathologies, as illustrated in Figure 4. We strictly follow the official splitting standards of ChestX-Ray14 provided by Wang et al. [2] to conduct our experiments so that our results are directly comparable with most published baselines. We use the training and validation sets to train our model and then evaluate the performance on the test set.

CheXpert is a popular dataset for recognizing, detecting and segmenting common chest and lung diseases. There are 224,616 images in the database, including 12 pathology labels and two nonpathology labels (not found and assistive device). Each image is assigned one or more disease symptoms, and the disease results are labeled as positive, negative and uncertain, as illustrated in Figure 4; if no positive disease is found in the image, it is labeled as 'no finding'. Undetermined labels in the images can be considered positive (CheXpert_1s) or negative (CheXpert_0s). On average, each image had 2.9 pathology labels for CheXpert_1s and 2.3 for CheXpert_0s. Since the data for the test set are still not published, we redivided the dataset into a training set, a validation set, and a test set at a ratio of 7:1:2.

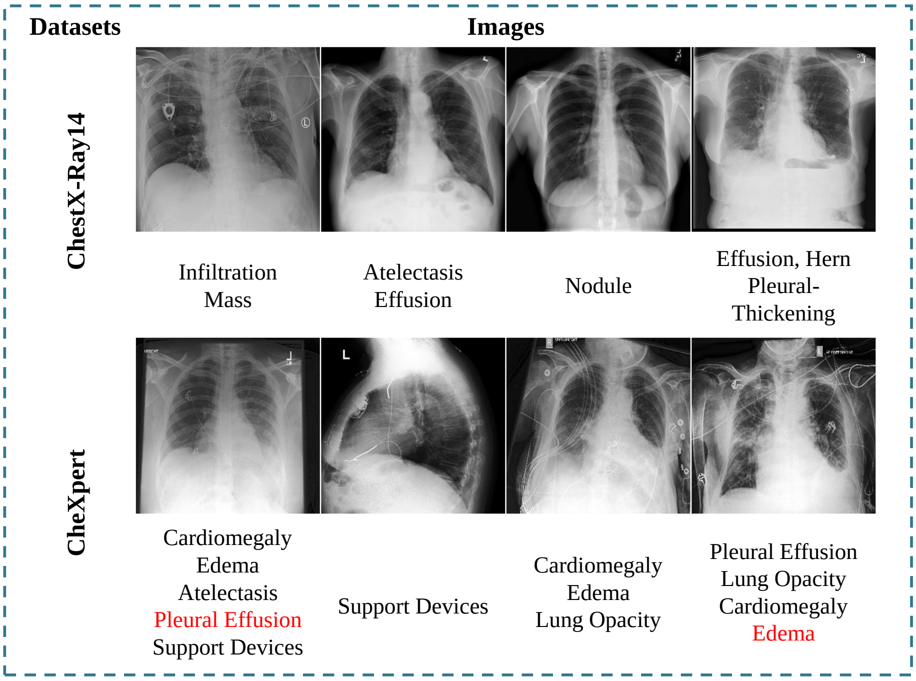


Figure 4. Example images and the corresponding labels in the ChestX-Ray14 and CheXpert datasets. Each image is labeled with one or more pathologies. In CheXpert, the uncertain pathology is marked in red.

As described earlier, the proposed PGL module involves the global modeling of all pathologies on the basis of cooccurrence pairs, the results of which are the identification of potential pathologies present in each image. As shown in Figure 5, many pathology pairs with cooccurrence relationships were obtained by counting the occurrences of all pathologies in both datasets separately. For example, lung disease is frequently associated with pleural effusion, and atelectasis is frequently associated with infiltration. This phenomenon serves as a basis for constructing pathology correlation matrix A_{pc} and provides initial evidence of the feasibility of the proposed PGL module.

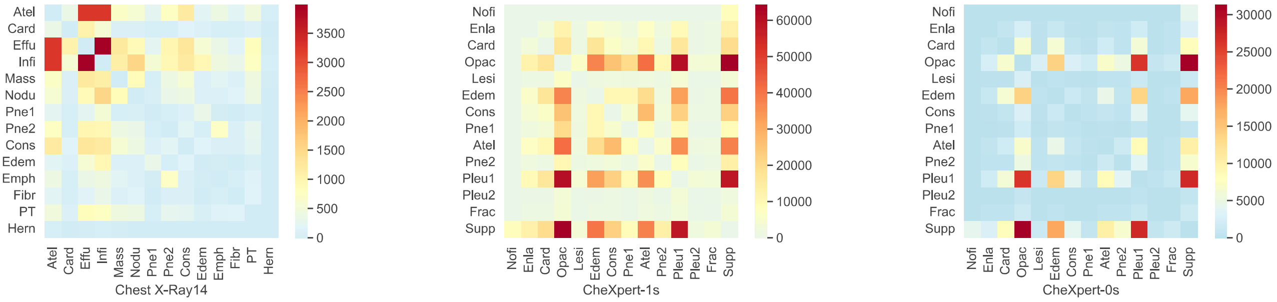


Figure 5. Graph representations of the pathology correlation extracted from the ChestX-Ray14, CheXpert_1s and CheXpert_0s datasets.

4.2. Implementation details

All experiments were run on an Intel 8268 CPU and NVIDIA Tesla V100 32 GB GPU. Moreover, it was implemented based on the PyTorch framework. First, we resize all images to 256 × 256 and normalize via the mean and standard deviation of the ImageNet dataset. Then, random cropping to make images 224 × 224, random horizontal flip, and random rotation were applied, as some images may have been flipped or rotated within the dataset. The output characteristic dimension D_1 of the backbone was 1664. In the PGL module, we designed a graph learning block consisting of 1-1 symmetrically structured GCN layers stacked with 2(2) graph attention layers (the number of attention heads within the layer). The number of GCN output channels was 1024 and 1024, respectively. We used a 2-layer GAT model, with the first layer using $K = 2$ attention heads, each head computing 512 features (1024 features in total), followed by exponential linear unit (ELU) [46] nonlinearity. The second layer did the same, averaging these features, followed by logistic sigmoid activation. In addition, we considered LeakyReLU with a negative slope of 0.2 as the nonlinear activation function used in the PGL module. The input pathology label word embedding was a 300-dimensional vector generated by the GloVe model pretrained on the Wikipedia dataset. When multiple words represented the pathology labels, we used the average vector of all words as the pathology label word embedding. In the MBM, we set D_3 = 14,336 to bridge the vectors of the two modes. Furthermore, we set $G = 1024$ with $g = 14$ to complete the GroupSum method. The ratios of dropout1 and dropout2 were 0.3 and 0.1, respectively. The whole network was updated by AdamW with a momentum of (0.9, 0.999) and a weight decay of 1e-4. The initial learning rate of the whole model was 0.001, which decreased 10 times every 10 epochs.

In our experiments, we used the AUC value [38] (the area under the receiver operating characteristic (ROC) curve [38]) for each pathology and the mean AUC value across all cases to measure the performance of MRChexNet. There was no data overlap between the training and testing subsets. The true label of each image was labeled with $L = [L_1, L_2, \ldots, L_C]$. In the dataset of two CXR label numbers $C = 14$, each element L_C indicated the presence or absence of the C-th pathology, i.e., 1 indicated presence and 0 indicated absence. For each image, the label was predicted as positive if the confidence level of the label was greater than 0.5.

4.3. Comparison with existing methods

In this section, we conduct experiments on ChestX-Ray14 and CheXpert to compare the performance of MRChexNet with existing methods.

Results from ChestX-Ray14 and discussion: We compared MRChexNet with a variety of existing methods including U-DCNN [2], LSTM-Net [1], CheXNet [10], DNet [39], AGCL [19], DR-DNN [12], CRAL [15], DualCheXN [25] and CheXGCN [6]. We present the results of the comparison on ChestX-Ray14 in Table 1 including the evaluation metrics for the entire dataset of 14 pathology labels. MRChexNet outperformed all candidate methods on most pathology-labeled metrics. Figure 6 illustrates the ROC curves of our model over the 14 pathologies on ChestX-Ray14. Specifically, MRChexNet outperformed these previous methods in mean AUC score, especially for U-DCNN (0.745) and LSTM-Net (0.798), with improvements of 10.5% and 3.7%, respectively. Moreover, it outperformed DualCheXNet (0.823) and improved the AUC score of detecting consolidation (0.819

vs. 0.746) and pneumonia (0.783 vs. 0.727) by more than 6.0%. Notably, the mean AUC score of MRChexNet improved by 2.4% over CheXGCN (0.826). The AUC scores of some pathologies labeled with MRChexNet obviously improved, e.g., cardiomegaly (0.923 vs. 0.893), consolidation (0.819 vs. 0.751), edema (0.904 vs. 0.850) and atelecta (0.824 vs. 0.786). It must be mentioned that our proposed model performed somewhat poorly on the nodule and fibrosis labels. Note that the pathogenesis of these diseases is systemic, and we generated word embeddings of their pathological labels using only their noun meanings without adding additional semantics to explain their sites of pathogenesis. This issue led to the unsatisfactory performance of MRChexNet on these pathologies. Overall, the proposed MRChexNet improved the multi-label recognition performance of ChestX-Ray14 and outperformed existing methods.

Table 1. AUC comparisons of MRChexNet with existing methods on ChestX-Ray14.

Method	ChestX-Ray14														Mean AUC
	AUC														
	atel	card	effu	infi	mass	nodu	pne1	pne2	cons	edem	emph	fibr	pt	hern	
U-DCNN [2]	0.700	0.810	0.759	0.661	0.693	0.669	0.658	0.799	0.703	0.805	0.833	0.786	0.684	0.872	0.745
LSTM-Net [1]	0.772	0.904	0.859	0.695	0.792	0.717	0.713	0.841	0.788	0.882	0.829	0.767	0.765	0.914	0.798
DR-DNN [12]	0.766	0.801	0.797	**0.751**	0.760	0.741	0.778	0.800	0.787	0.820	0.773	0.765	0.759	0.748	0.775
AGCL [19]	0.756	0.887	0.819	0.689	0.814	0.755	0.729	0.850	0.728	0.848	0.906	0.818	0.765	0.875	0.803
CheXNet [10]	0.769	0.885	0.825	0.694	0.824	0.759	0.715	0.852	0.745	0.842	0.906	0.821	0.766	0.901	0.807
DNet [39]	0.767	0.883	0.828	0.709	0.821	0.758	0.731	0.846	0.745	0.835	0.895	0.818	0.761	0.896	0.807
CRAL [15]	0.781	0.880	0.829	0.702	0.834	0.773	0.729	0.857	0.754	0.850	0.908	0.830	0.778	0.917	0.816
DualCheXN [25]	0.784	0.888	0.831	0.705	0.838	0.796	0.727	0.876	0.746	0.852	0.942	**0.837**	0.796	0.912	0.823
CheXGCN [6]	0.786	0.893	0.832	0.699	0.840	**0.800**	0.739	0.876	0.751	0.850	**0.944**	0.834	0.795	0.929	0.826
MRChexNet (Ours)	**0.824**	**0.923**	**0.894**	0.719	**0.857**	0.779	**0.783**	**0.888**	**0.819**	**0.904**	0.920	0.835	**0.808**	**0.946**	**0.850**

Note: The 14 pathologies in Chest X-Ray14 are atelectasis (atel), cardiomegaly (card), effusion (effu), infiltration (infi), mass, nodule (nodu), pneumonia (pne1), pneumothorax (pne2), consolidation (cons), edema (edem), emphysema (emph), fibrosis (fibr), pleural thickening (pt) and hernia (hern).

Table 2. AUC comparisons of MRChexNet with previous baseline on CheXpert_1s.

Method	CheXpert_1s														Mean AUC
	AUC														
	nofi	enla	card	opac	lesi	edem	cons	pne1	atel	pne2	pleu1	pleu2	frac	supp	
ML-GCN [22]	0.879	0.630	0.841	0.723	0.773	0.856	0.692	0.740	0.713	0.829	0.873	0.802	0.762	0.868	0.784
U_Ones [23]	0.890	0.659	0.856	0.735	0.778	0.847	0.701	0.756	0.722	0.855	0.871	0.798	0.789	0.878	0.795
DenseNet-169 [11]	0.916	0.717	0.895	0.770	0.783	0.882	**0.710**	**0.774**	0.728	0.871	0.916	0.817	0.805	0.909	0.821
MRChexNet_1s (Ours)	**0.976**	**0.738**	**0.900**	**0.887**	**0.940**	**0.884**	0.701	0.719	**0.759**	**0.925**	**0.924**	**0.852**	**0.958**	**0.944**	**0.865**

Note: The 14 pathologies in CheXpert are no Finding (nofi), enlarged cardiomediastinum (enla), cardiomegaly (card), lung opacity (opac), lung lesion (lesi), edema (edem), consolidation (cons), pneumonia (pne1), atelectasis (atel), pneumothorax (pne2), pleural effusion (pleu1), pleural other (pleu2), fracture (frac) and support devices (supp).

Table 3. AUC comparisons of MRChexNet with the previous baseline on CheXpert_0s.

Method	CheXpert_0s														Mean AUC
	AUC														
	nofi	enla	card	opac	lesi	edem	cons	pne1	atel	pne2	pleu1	pleu2	frac	supp	
ML-GCN [22]	0.864	0.673	0.831	0.681	0.802	0.770	0.713	0.758	0.654	0.845	0.841	0.764	0.754	0.838	0.771
U_Zeros [23]	0.885	0.678	0.865	0.730	0.760	0.853	0.735	0.740	0.700	0.872	0.880	0.775	0.743	0.877	0.792
DenseNet-169 [11]	0.912	0.715	0.884	0.738	0.780	**0.861**	0.753	0.770	0.711	0.860	**0.904**	0.830	0.758	**0.878**	0.811
MRChexNet_0s (Ours)	**0.914**	**0.808**	**0.894**	**0.748**	**0.913**	0.827	**0.801**	**0.868**	**0.744**	**0.928**	0.876	**0.909**	**0.915**	0.859	**0.858**

Results from CheXpert and discussion: To our limited knowledge, the test set of CheXpert has yet to be publicly available and can only be redivided by itself. Fewer state-of-the-art methods are available for comparison. Based on that, we further evaluated the comparison of our model with the uncertainty labeling treatments mentioned in the original dataset (U_Ones and U_Zeros). As shown in

Table 2, MRChexNet_1s obtained higher mean AUC scores on 14 pathological labels for CheXpert_1s, which were 1.5% higher than the techniques in the original paper U_Ones. Additionally, compared to the vanilla DenseNet-169, the improvement is 3.8%. As shown in Table 3, MRChexNet_0s obtained higher mean AUC scores on 14 pathological labels for CheXpert_0s, which were 2.1% higher than the techniques U_Zeros in the original paper. The mean AUC score of MRChexNet is 3.1% higher than that of vanilla DenseNet-169. These results prove that our two proposed modules can work better when reinforcing each other. Overall, the AUC score of MRChexNet_1s was better than that of MRChexNet_0s by 0.3%, especially for lung lesions by 3.5% (0.788→0.823), atelectasis by 2.5% (0.707→0.732) and fracture by 2.7% (0.793→0.820). This is because the true value of these uncertainty labels on the image is likely to be negative. The converse is also true. Figure 6 illustrates the ROC curves of MRChexNet on ChestX-ray14, CheXpert_1s and CheXpert_0s for the 14 pathologies.

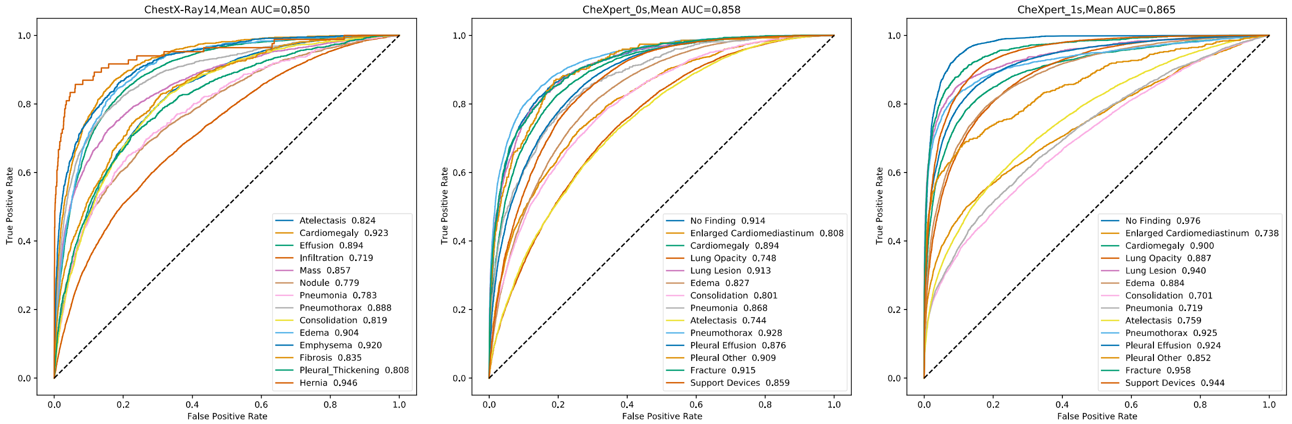


Figure 6. ROC curves of MRChexNet on the ChestXRay14 and CheXpert, respectively. The corresponding AUC scores are given in Tables 1−3.

Table 4. Comparison of AUC of MRChexNet with its different components on ChestX-Ray14.

Method	Chest X-Ray14														Mean AUC
	AUC														
	atel	card	effu	infi	mass	nodu	pneu1	pneu2	cons	edem	emph	fibr	pt	hern	
Baseline : DenseNet-169 [11]	0.775	0.879	0.826	0.685	0.766	0.689	0.725	0.823	0.788	0.841	0.838	0.767	0.742	0.811	0.782
Baseline + MBM	0.800	0.892	0.860	0.707	0.856	0.760	0.741	0.859	0.810	0.870	0.883	0.711	0.781	0.796	0.809
Baseline + PGL	0.820	0.920	0.888	0.710	0.784	0.769	0.756	0.873	0.808	0.896	0.874	0.744	0.799	0.804	0.818
MRChexNet (Ours)	**0.824**	**0.923**	**0.894**	**0.719**	**0.857**	**0.779**	**0.783**	**0.888**	**0.819**	**0.904**	**0.920**	**0.835**	**0.808**	**0.946**	**0.850**

4.4. Ablation experiments and discussion

MRChexNet with its different components on ChestX-Ray14: We experimented with the performance of the components of the MRChexNet; the results are shown in Table 4. In baseline + PGL, we use a simple summation of elements instead of MBM to fuse the visual feature vectors of pathology and the semantic word vectors of pathology. The obtained simple fusion vectors are used as the node features of the graph learning block. Compared to the baseline DenseNet-169, the mean AUC score of baseline + PGL was significantly higher by 3.6% (0.782 → 0.818), especially in atelectasis (0.775 → 0.820), cardiomegaly (0.879 → 0.920), effusion (0.826 → 0.888) and nodule (0.689 → 0.769), exceeding the vanilla DenseNet-169 by an average of 5.7% in those pathology labels. The experimental results showed that the proposed PGL module is crucial in mining the global cooccurrence between pathologies. Note that in the baseline + MBM model, the fixed direct $input_2$ to the MBM module is a vector of 14 pathology-annotated words with initial semantic information. We learn the output of the resulting cross-modal fusion vectors from one FC layer by aligning the visual features of pathology with the semantic word vectors of pathology. Compared to the DenseNet-169 baseline, the mean AUC score of baseline + MBM was significantly higher by 2.7% (0.782 → 0.809), especially in atelectasis (0.775 → 0.800), effusion (0.826 → 0.860), pneumothorax (0.823 → 0.859), and mass (0.766 → 0.856) on pathology, exceeding the vanilla DenseNet-169 by an average of 4.6% in those pathology labels. With the addition of the MBM and PGL modules, MRChexNet significantly improved the mean AUC score by 6.8%. In particular, the AUC score improvement was significant for atelectasis (0.775 → 0.824), pneumothorax (0.823 → 0.888), and emphysema (0.838 → 0.920). This phenomenon indicates that the MBM and PGL modules in our framework can reinforce and complement each other to make MRChexNet perform at its best.

Table 5. Comparison of the test time of MRChexNet with its different components.

Method	Test time (1 image)
Baseline : DenseNet−169	2.5×10^{-6} s
(Baseline + MBM) − Baseline	12.1×10^{-6} s
(Baseline + PGL) − Baseline	20.3×10^{-6} s
MRChexNet (Ours)	33.7×10^{-6} s

Testing time for different components in MRChexNet: We experimented with the inference time for each component of MRChexNet, and the results are shown in the Table 5. We have set the inference time in seconds and the inference duration as the time to infer 1 image. Then, we first tested an image using Baseline and the obtained time as a base. After testing an image using Baseline + MBM and Baseline + PGL to get the duration, the base inference duration of the previous baseline is subtracted to get the exact inference duration of each module. According to the results, it can be seen that MBM and PGL increase the reasoning time of the model by 20.3×10^{-6} and 33.7×10^{-6} s, respectively. It is worth mentioning that the interaction of the two achieves a satisfactory recognition performance, which is an acceptable result compared to the manual reasoning time of the radiologist.

MRChexNet under different types of word embeddings: We default to using GloVe [40] as the token representation as input to the multi-modal bridge module (MBM). In this section, we evaluate the performance of MRChexNet under other types of popular word representations. Specifically, we investigate four different word embedding methods, including GloVe [40], FastText [41], and simple single-hot word embedding. Figure 7 shows the results using different word embeddings on ChestX-Ray14 and CheXpert. As shown, we can see that thoracic disease recognition accuracy is not significantly affected when using different word embeddings as inputs to the MBM. Furthermore, the observations (especially the results of one-hot) demonstrate that the accuracy improvement achieved by our approach does not come entirely from the semantics produced by the word embeddings. Furthermore, using powerful word embeddings led to better performance. One possible reason may be that the word embeddings learned from a large text corpus maintain some semantic topology. That is, semantic-related concept embeddings are close in the embedding space. Our model can employ these implicit dependencies and further benefit thoracic disease recognition.

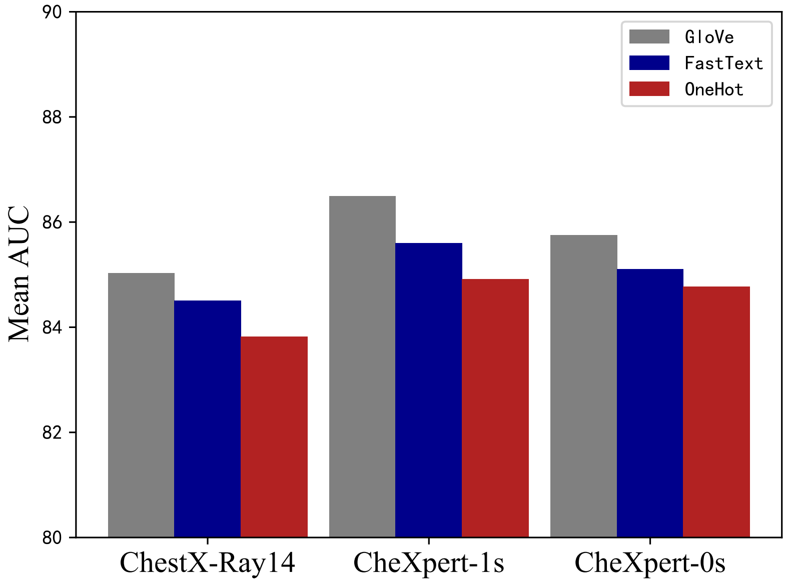


Figure 7. Effects of different pathology word embedding approaches. It is clear that different pathology word embeddings have little effect on accuracy. This shows that our improvements are not necessarily due to the semantic meanings derived from the pathology word embeddings but rather to our MRChexNet.

Groups G and elements g in GroupSum: In this section, we evaluate the performance of the MBM in MRChexNet by using a different number of groups G and the number of elements g within a group. With the GroupSum in the MBM, each D_3-dimensional vector will be converted into a G-dimensional vector. We have a set of G-g $\in \{(2048, 7), (1024, 14), (512, 28), (256, 56), (128, 112)\}$ to generate a low-dimensional bridging vector. As shown in Figure 8, MRChexNet obtains better performance on ChestX-Ray14 when $G = 1024$ and $g = 14$ are chosen, while the change in the mean AUC is very slight on CheXpert. We believe that the original semantic information between the pathology word embeddings can be better expressed by $G = 1024$ and $g = 14$. Other values of G-g bring similar results, which do not affect the model too much.

Different numbers of GCN layers and GAT layers of the graph learning block in PGL: Since the front end of the graph learning block we have designed is a GCN with a dual-branch symmetric structure, the main discussion is about the number of GCN layers on each branch. We set the graph attention layer at the end of the graph learning block. To maintain the symmetry of the graph learning block structure, we kept the number of layers the same as the number of attention heads within the layer. We show the performance results for different GCN layers of our model in Table 6. For the 1-1 layer model to GCN, in each branch, the output dimensions of the sequential layers are 1024. For the 2-2 layer model to GCN, in each branch, the output dimensions of the sequential layers are 1024 and 1024. For the 3-3 layer model to GCN, in each branch, the output dimensions of the sequential layers are 1024. We aligned the number of graph attention layers with the number of attention heads. Specifically, for the 1-layer GAT model, with the layer using $K = 1$ attention heads, the head computes 1024 features (1024 features in total). For the 2-layer GAT model, with the first layer using $K = 2$ attention heads, each head computes 512 features (1024 features in total), and the second layer does the same. As shown in the table, the pathology recognition performance on both datasets decreased when the number of GCN layers and the number of GAT layers increased. The performance degradation was due to the accumulation of information transfer between nodes when more GCN and GAT layers were used, leading to oversmoothing.

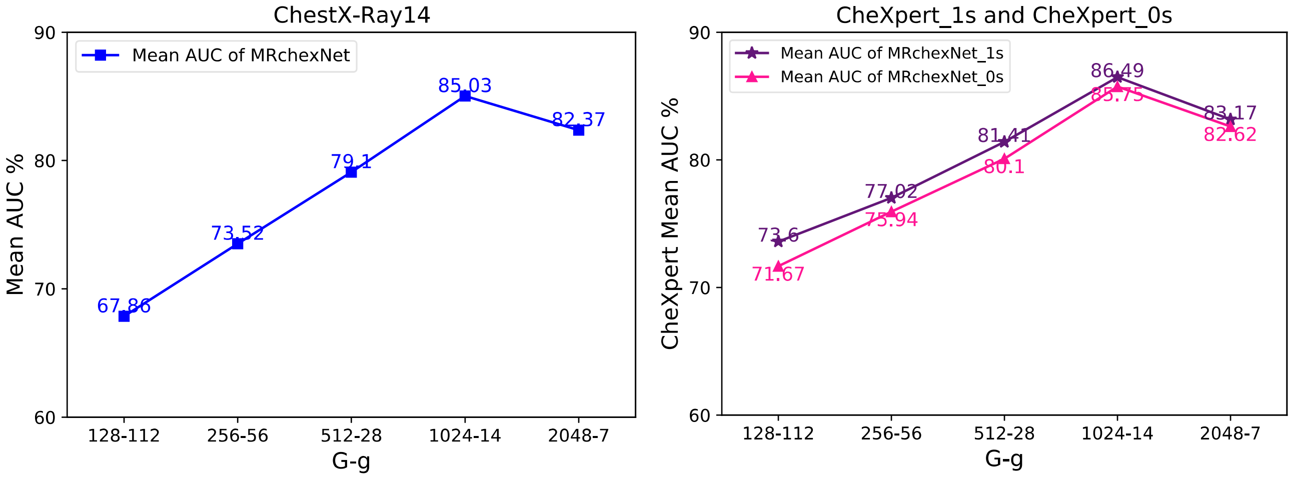


Figure 8. The change of mean AUC using different values of G-g.

Table 6. The different number of GCN layers and GAT layers of the graph learning block in PGL.

Mean AUC				
#Layer		Dataset		
Dual-branch GCN	GAT (heads)	ChestX-Ray14	CheXpert-0s	CheXpert-1s
1-1	1(1)	0.8417	0.8493	0.8366
	2(2)	**0.8503**	**0.8649**	**0.8575**
2-2	1(1)	0.8342	0.8402	0.8309
	2(2)	0.8251	0.8323	0.8187
3-3	1(1)	0.8187	0.8238	0.8194
	2(2)	0.8063	0.8109	0.8057

4.5. Visualization of lesion areas for qualitative assessment

In Figure 9, we visualize the original images and the corresponding label-specific activation maps obtained by our proposed MRChexNet. It is clear that MRChexNet can capture the discriminative semantic regions of the images for the different chest diseases. Figure 10 illustrates a visual represen-

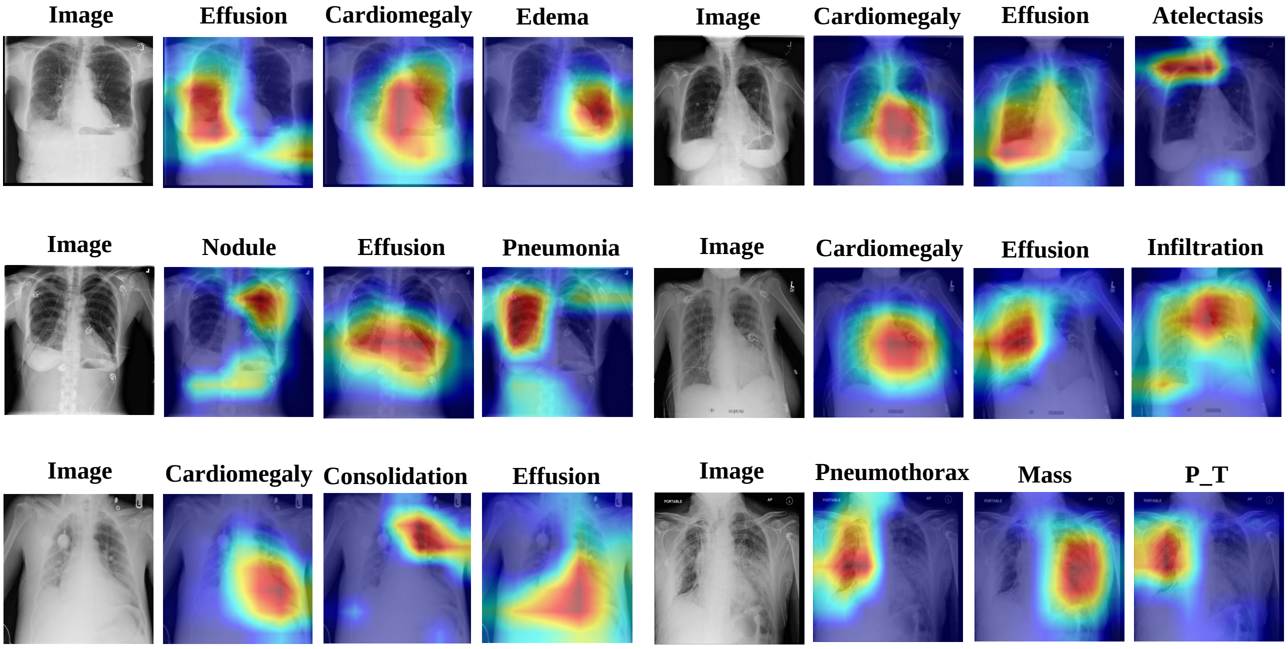
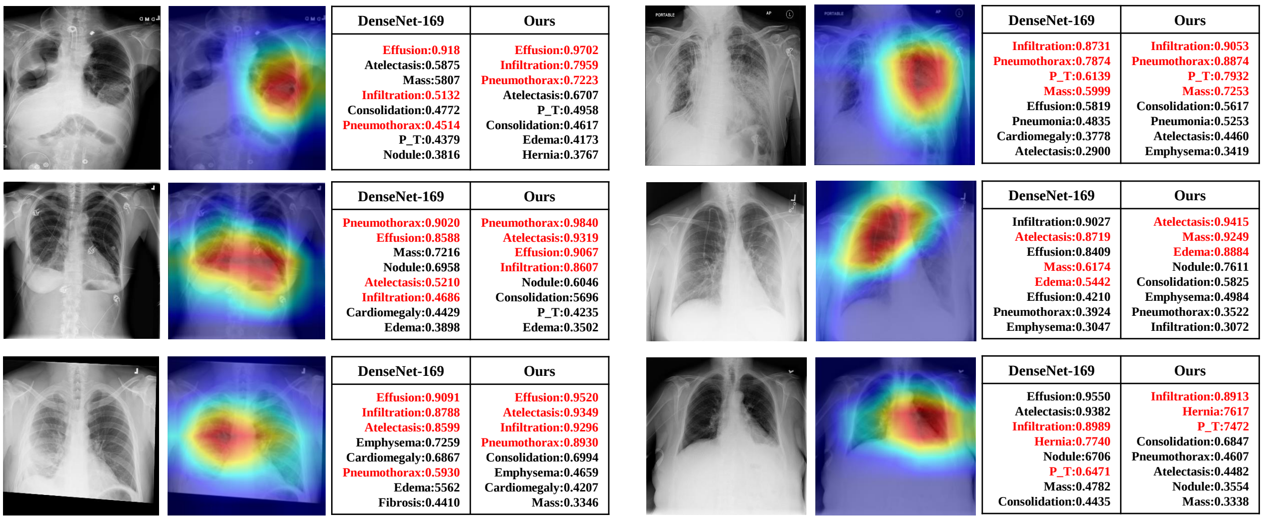


Figure 9. Visualization results of pathology correlation activation maps on ChestX-Ray14 dataset. The three columns on the right are three samples with different diseases and their corresponding activation maps.



DenseNet-169	Ours
Effusion:0.918	Effusion:0.9702
Atelectasis:0.5875	Infiltration:0.7959
Mass:5807	Pneumothorax:0.7223
Infiltration:0.5132	Atelectasis:0.6707
Consolidation:0.4772	P_T:0.4958
Pneumothorax:0.4514	Consolidation:0.4617
P_T:0.4379	Edema:0.4173
Nodule:0.3816	Hernia:0.3767

DenseNet-169	Ours
Infiltration:0.8731	Infiltration:0.9053
Pneumothorax:0.7874	Pneumothorax:0.8874
P_T:0.6139	P_T:0.7932
Mass:0.5999	Mass:0.7253
Effusion:0.5819	Consolidation:0.5617
Pneumonia:0.4835	Pneumonia:0.5253
Cardiomegaly:0.3778	Atelectasis:0.4460
Atelectasis:0.2900	Emphysema:0.3419

DenseNet-169	Ours
Pneumothorax:0.9020	Pneumothorax:0.9840
Effusion:0.8588	Atelectasis:0.9319
Mass:0.7216	Effusion:0.9067
Nodule:0.6958	Infiltration:0.8607
Atelectasis:0.5210	Nodule:0.6046
Infiltration:0.4686	Consolidation:5696
Cardiomegaly:0.4429	P_T:0.4235
Edema:0.3898	Edema:0.3502

DenseNet-169	Ours
Infiltration:0.9027	Atelectasis:0.9415
Atelectasis:0.8719	Mass:0.9249
Effusion:0.8409	Edema:0.8884
Mass:0.6174	Nodule:0.7611
Edema:0.5442	Consolidation:0.5825
Effusion:0.4210	Emphysema:0.4984
Pneumothorax:0.3924	Pneumothorax:0.3522
Emphysema:0.3047	Infiltration:0.3072

DenseNet-169	Ours
Effusion:0.9091	Effusion:0.9520
Infiltration:0.8788	Atelectasis:0.9349
Atelectasis:0.8599	Infiltration:0.9296
Emphysema:0.7259	Pneumothorax:0.8930
Cardiomegaly:0.6867	Consolidation:0.6994
Pneumothorax:0.5930	Emphysema:0.4659
Edema:5562	Cardiomegaly:0.4207
Fibrosis:0.4410	Mass:0.3346

DenseNet-169	Ours
Effusion:0.9550	Infiltration:0.8913
Atelectasis:0.9382	Hernia:7617
Infiltration:0.8989	P_T:7472
Hernia:0.7740	Consolidation:0.6847
Nodule:6706	Pneumothorax:0.4607
P_T:0.6471	Atelectasis:0.4482
Mass:0.4782	Nodule:0.3554
Consolidation:0.4435	Mass:0.3338

Figure 10. Visualization results of our model scoring the highest pathology on the images to be tested in the ChestX-Ray14 dataset. We present the top-eight predicted pathology labels and the corresponding probability scores. The ground truth labels are highlighted in red.

tation of multi-label CXR recognition. The top-eight predicted scores for each test subject are given and sorted top-down by the magnitude of the predicted score values. As shown in Figure 10, compared with the vanilla DenseNet-169 model, the proposed MRChexNet enhances the performance of multi-label CXR recognition. Our MRChexNet can effectively improve associated pathology confidence scores and suppress nonassociated pathology scores with fully considered and modeled global label relationships. For example, in column 1, row 2, MRChexNet fully considers the pathological relationship between effusion and atelectasis. In the presence of effusion, the corresponding confidence score for atelectasis was (0.5210 → 0.9319); compared to vanilla DenseNet-169 performance, the confidence score improved by approximately 0.4109. For the weakly correlated labels, effusion ranked first in column 2, row 3 regarding the DenseNet-169 score. While MRChexNet fully considers the global interlabel relationships, its confidence score does not reach the top 8. To some extent, this demonstrates the ability of our model to suppress the confidence scores of nonrelevant pathologies.

5. Conclusions

Improving the performance of multi-label CXR recognition algorithms in clinical environments by considering the correspondence between pathology labels in different modalities and capturing the correlation relationship between related pathologies is vital, as is aligning pathology-relationship representations in different modalities and learning the relationship information of pathologies within each modality. In this paper, we propose a multi-modal bridge and relational learning method named MRChexNet to align pathological representations in different modalities and learn information about the relationship of pathology within each modality. Specifically, our model first extracts pathology-specific feature representations in the imaging modality by designing a practical RLM. Then, an efficient MBM is designed to align pathological word embeddings and image-level pathology-specific feature representations. Finally, a novel PGL is intended to comprehensively learn the correlation of pathologies within each modality. Extensive experimental results on ChestX-Ray14 and CheXpert show that the proposed MBM and PGL can effectively enhance each other, thus significantly improving the model's multi-label CXR recognition performance with satisfactory results. In the future, we will introduce the relation weight parameter in pathology relation modeling to learn more accurate pathology relations to help further improve the multi-label CXR recognition performance.

In the future, we will extend the applicability of the proposed method to other imaging modalities, such as optical coherence tomography (OCT). Among them, OCT is a noninvasive optical imaging modality that provides histopathology images with microscopic resolution [42–45]. Our next research direction is extending the proposed method for OCT-based pathology image analysis. In addition, exploring the interpretability and readability of models has been a hot research topic in making deep learning techniques applicable to clinical diagnosis. Our next research direction is also how to make our model more friendly and credible for clinicians' understanding.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the National Nature Science Foundation of China (No. 61872225), Introduction and Cultivation Program for Young Creative Talents in Colleges and Universities of Shandong Province (No. 2019-173), the Natural Science Foundation of Shandong Province (No. ZR2020KF013, No. ZR2020ZD44, No. ZR2019ZD04, No. ZR2020QF043, No. ZR2023QF094) and the Special fund of Qilu Health and Health Leading Talents Training Project.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, K. Lyman, Learning to diagnose from scratch by exploiting dependencies among labels, preprint, arXiv:1710.10501.
2. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 2097–2106. https://doi.org/10.1109/cvpr.2017.369
3. C. Galleguillos, A. Rabinovich, S. Belongie, Object categorization using co-occurrence, location and appearance, in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2008), 1–8. https://doi.org/10.1109/cvpr.2008.4587799
4. L. Luo, D. Xu, H. Chen, T. T. Wong, P. A. Heng, Pseudo bias-balanced learning for debiased chest X-ray classification, in *Medical Image Computing and Computer Assisted Intervention*, (2022), 621–631. https://doi.org/10.1007/978-3-031-16452-1_59
5. G. Karwande, A. B. Mbakwe, J. T. Wu, L. A. Celi, M. Moradi, I. Lourentzou, Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest X-rays, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2022), 581–591. https://doi.org/10.1007/978-3-031-16431-6_55
6. B. Chen, J. Li, G. Lu, H. Yu, D. Zhang, Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification, *IEEE J. Biomed. Health Inf.*, **24** (2020), 2292–2302. https://doi.org/10.1109/jbhi.2020.2967084
7. L. Luo, H. Chen, Y. Zhou, H. Lin, P. A. Heng, Oxnet: deep omni-supervised thoracic disease detection from chest X-rays, in *Medical Image Computing and Computer Assisted Intervention*, (2021), 537–548. https://doi.org/10.1007/978-3-030-87196-3_50
8. B. Hou, G. Kaissis, R. M. Summers, B. Kainz, Ratchet: Medical transformer for chest X-ray diagnosis and reporting, in *Medical Image Computing and Computer Assisted Intervention*, (2021), 293–303. https://doi.org/10.1007/978-3-030-87234-2_28

9. W. Liao, H. Xiong, Q. Wang, Y. Mo, X. Li, Y. Liu, et al., Muscle: Multi-task self-supervised continual learning to pre-train deep models for X-ray images of multiple body parts, in *Medical Image Computing and Computer Assisted Intervention*, (2022), 151–161. https://doi.org/10.1007/978-3-031-16452-1_15

10. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, et al., Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning, preprint, arXiv:1711.05225.

11. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 4700–4708. https://doi.org/10.1109/cvpr.2017.243

12. Y. Shen, M. Gao, Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization, in *International Workshop on Machine Learning in Medical Imaging*, Springer, (2018), 389–397. https://doi.org/10.1007/978-3-030-00919-9_45

13. F. Zhu, H. Li, W. Ouyang, N. Yu, X. Wang, Learning spatial regularization with image-level supervisions for multi-label image classification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 5513–5522. https://doi.org/10.1109/cvpr.2017.219

14. Z. Wang, T. Chen, G. Li, R. Xu, L. Lin, Multi-label image recognition by recurrently discovering attentional regions, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 464–472. https://doi.org/10.1109/iccv.2017.58

15. Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, Y. Yang, Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification, preprint, arXiv:1801.09927.

16. J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, CNN-RNN: A unified framework for multi-label image classification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 2285–2294. https://doi.org/10.1109/cvpr.2016.251

17. P. P. Ypsilantis, G. Montana, Learning what to look in chest X-rays with a recurrent visual attention model, preprint, arXiv:1701.06452.

18. X. Wang, Y. Peng, L. Lu, Z. Lu, R. M. Summers, Tienet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 9049–9058. https://doi.org/10.1109/cvpr.2018.00943

19. Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, R. M. Summers, Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs, in *International Workshop on Machine Learning in Medical Imaging*, Springer, (2018), 249–258. https://doi.org/10.1007/978-3-030-00919-9_29

20. C. W. Lee, W. Fang, C. K. Yeh, Y. C. F. Wang, Multi-label zero-shot learning with structured knowledge graphs, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 1576–1585. https://doi.org/10.1109/cvpr.2018.00170

21. J. Yu, Y. Lu, Z. Qin, W. Zhang, Y. Liu, J. Tan, et al., Modeling text with graph convolutional network for cross-modal information retrieval, in *Pacific Rim Conference on Multimedia*, Springer, (2018), 223–234. https://doi.org/10.1007/978-3-030-00776-8_21

22. Z. M. Chen, X. S. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 5177–5186. https://doi.org/10.1109/cvpr.2019.00532

23. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 590–597. https://doi.org/10.1609/aaai.v33i01.3301590

24. Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L. J. Li, et al., Thoracic disease identification and localization with limited supervision, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 8290–8299. https://doi.org/10.1109/cvpr.2018.00865

25. B. Chen, J. Li, X. Guo, G. Lu, Dualchexnet: dual asymmetric feature learning for thoracic disease classification in chest X-rays, *Biomed. Signal Process. Control*, **53** (2019), 101554. https://doi.org/10.1016/j.bspc.2019.04.031

26. H. Sak, A. Senior, F. Beaufays, Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition, preprint, arXiv:1402.1128.

27. Q. Li, X. Peng, Y. Qiao, Q. Peng, Learning category correlations for multi-label image recognition with graph networks, preprint, arXiv:1909.13005.

28. A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, preprint, arXiv:1606.01847.

29. Y. Hu, K. Liu, K. Ho, D. Riviello, J. Brown, A. R. Chang, et al., A simpler machine learning model for acute kidney injury risk stratification in hospitalized patients, *J. Clin. Med.*, **11** (2022), 5688. https://doi.org/10.3390/jcm11195688

30. R. Xu, F. Shen, H. Wu, J. Zhu, H. Zeng, Dual modal meta metric learning for attribute-image person re-identification, in *2021 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, IEEE, **1** (2021), 1–6. https://doi.org/10.1109/icnsc52481.2021.9702261

31. J. H. Kim, K. W. On, W. Lim, J. Kim, J. W. Ha, B. T. Zhang, Hadamard product for low-rank bilinear pooling, preprint, arXiv:1610.04325.

32. Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering, *IEEE Trans. Neural Networks Learn. Syst.*, **29** (2018), 5947–5959. https://doi.org/10.1109/tnnls.2018.2817340

33. Y. Wang, Y. Xie, Y. Liu, K. Zhou, X. Li, Fast graph convolution network based multi-label image recognition via cross-modal fusion, in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, (2020), 1575–1584. https://doi.org/10.1145/3340531.3411880

34. M. Lin, Q. Chen, S. Yan, Network in network, preprint, arXiv:1312.4400.

35. T. Y. Lin, A. RoyChowdhury, S. Maji, Bilinear cnn models for fine-grained visual recognition, in *Proceedings of the IEEE International Conference on Computer Vision*, (2015), 1449–1457. https://doi.org/10.1109/iccv.2015.170

36. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, preprint, arXiv:1710.10903.

37. Z. Cao, T. Qin, T. Y. Liu, M. F. Tsai, H. Li, Learning to rank: from pairwise approach to listwise approach, in *Proceedings of the 24th International Conference on Machine Learning*, (2007), 129–136. https://doi.org/10.1145/1273496.1273513

38. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, et al., pROC: an open-source package for R and S+ to analyze and compare ROC curves, *BMC Bioinf.*, **12** (2011), 1–8. https://doi.org/10.1186/1471-2105-12-77

39. S. Guendel, S. Grbic, B. Georgescu, S. Liu, A. Maier, D. Comaniciu, Learning to recognize abnormalities in chest X-rays with location-aware dense networks, in *Iberoamerican Congress on Pattern Recognition*, Springer, (2018), 757–765. https://doi.org/10.1007/978-3-030-13469-3_88

40. J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2014), 1532–1543, https://doi.org/10.3115/v1/d14-1162

41. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, preprint, arXiv:1301.3781.

42. R. K. Meleppat, K. E. Ronning, S. J. Karlen, M. E. Burns, E. N. Pugh Jr, R. J. Zawadzki, *In vivo* multimodal retinal imaging of disease-related pigmentary changes in retinal pigment epithelium, *Sci. Rep.*, **11** (2021), 16252. https://doi.org/10.1038/s41598-021-95320-z

43. R. Meleppat, M. Matham, L. Seah, An efficient phase analysis-based wavenumber linearization scheme for swept source optical coherence tomography systems, *Laser Phys. Lett.*, **12** (2015), 055601. https://doi.org/10.1088/1612-2011/12/5/055601.

44. R. Meleppat, C. Fortenbach, Y. Jian, E. Martinez, K. Wagner, B. Modjtahedi, et al., *In vivo* imaging of retinal and choroidal morphology and vascular plexuses of vertebrates using swept-source optical coherence tomography, *Transl. Vision Sci. Technol.*, **11** (2022), 11. https://doi.org/10.1167/tvst.11.8.11

45. K. Ratheesh, L. Seah, V. Murukeshan, Spectral phase-based automatic calibration scheme for swept source-based optical coherence tomography systems, *Phys. Med. Biol.*, **61** (2016), 7652. https://doi.org/10.1088/0031-9155/61/21/7652

46. D. A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (elus), preprint, arXiv:1511.07289.