Original Article

Modeling long-range dependencies for weakly supervised disease classification and localization on chest X-ray

Fangyun Li[1#], Lingxiao Zhou[2#], Yunpeng Wang[1#], Chuan Chen[3], Shuyi Yang[4,5], Fei Shan[3], Lei Liu[1,6]

[1]Institute of Biomedical Sciences, Fudan University, Shanghai, China; [2]Institute of Microscale Optoelectronics, Shenzhen University, Shenzhen, China; [3]Department of Radiology, Shanghai Public Health Clinical Center, Fudan University, Shanghai, China; [4]Department of Radiology, Zhongshan Hospital, Fudan University, Shanghai, China; [5]Shanghai Institute of Medical Imaging, Shanghai, China; [6]School of Basic Medical Sciences, Fudan University, Shanghai, China

Contributions: (I) Conception and design: All authors; (II) Administrative support: All authors; (III) Provision of study materials or patients: C Chen, S Yang; (IV) Collection and assembly of data: F Li; (V) Data analysis and interpretation: F Li; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work and should be considered as co-first authors.

Correspondence to: Lei Liu, PhD. Institute of Biomedical Sciences and the School of Basic Medical Sciences, Fudan University, 138 Yixueyuan Rd, Shanghai 200032, China. Email: liulei_sibs@163.com.

Background: Computer-aided diagnosis based on chest X-ray (CXR) is an exponentially growing field of research owing to the development of deep learning, especially convolutional neural networks (CNNs). However, due to the intrinsic locality of convolution operations, CNNs cannot model long-range dependencies. Although vision transformers (ViTs) have recently been proposed to alleviate this limitation, those trained on patches cannot learn any dependencies for inter-patch pixels and thus, are insufficient for medical image detection. To address this problem, in this paper, we propose a CXR detection method which integrates CNN with a ViT for modeling patch-wise and inter-patch dependencies.

Methods: We experimented on the ChestX-ray14 dataset and followed the official training-test set split. Because the training data only had global annotations, the detection network was weakly supervised. A DenseNet with a feature pyramid structure was designed and integrated with an adaptive ViT to model inter-patch and patch-wise long-range dependencies and obtain fine-grained feature maps. We compared the performance using our method with that of other disease detection methods.

Results: For disease classification, our method achieved the best result among all the disease detection methods, with a mean area under the curve (AUC) of 0.829. For lesion localization, our method achieved significantly higher intersection of the union (IoU) scores on the test images with bounding box annotations than did the other detection methods. The visualized results showed that our predictions were more accurate and detailed. Furthermore, evaluation of our method in an external validation dataset demonstrated its generalization ability.

Conclusions: Our proposed method achieves the new state of the art for thoracic disease classification and weakly supervised localization. It has potential to assist in clinical decision-making.

Keywords: Long-range dependencies; vision transformer (ViT); chest X-rays (CXRs); disease classification; localization

Submitted Nov 18, 2021. Accepted for publication Mar 14, 2022.
doi: 10.21037/qims-21-1117
View this article at: https://dx.doi.org/10.21037/qims-21-1117

Introduction

Chest X-ray (CXR) imaging is one of the most commonly available and widely applied radiological examinations at present, and is essential for screening and clinical diagnosis (1,2). In clinical practice, the reading of CXRs heavily depends on manual observation by radiologists with professional knowledge and experience. However, due to the complex pathologies of different lung lesions, even radiologists with long-term clinical training and professional guidance can make mistakes (3). The emergence and rapid progress of deep learning can improve disease diagnosis based on CXRs and lessen the chances of mistakes being made (4-6). As a fundamental task in medical image analysis, image detection allows for disease classification and lesion localization, which helps radiologists to work more efficiently and accurately. Therefore, it is crucial that a computer-aided diagnosis system for thoracic disease classification and localization on CXR images is developed (7).

The development of deep learning methods for disease detection is heavily dependent on supervised learning with high-quality disease annotations, such as pixel-level labels, which can be costly. Large public chest radiography datasets such as ChestX-ray14 (8) and CheXpert (9), which do not provide pixel-wise annotations or coarse bounding boxes for most CXR images, are not suitable for training supervised disease detection models. Consequently, several unsupervised and weakly supervised learning methods have been proposed to handle tasks without pixel-level annotations (8,10,11). However, the established workflow of unsupervised methods is complex, challenging, and inefficient in clinical-aided analysis (12). Therefore, the development of automatic detection models for thoracic diseases must rely on weakly supervised methods with image-level labels and a minimal amount of bounding box annotations (3,8).

Recently, convolutional neural networks (CNNs) have been widely adopted for weakly supervised thoracic disease detection (8,13-15). In CNNs, each kernel slides over the image and detects the same local pattern in each field. Convolution layers learn image features from small input patches and preserve the spatial relations between pixels. A recent study showed that CNNs can learn object detectors effectively, even when they are trained as classifiers using only global labels (16). The localization ability of CNNs is achieved by identifying the regions used for classification in the last convolutional layer of the network. However, CNN-based approaches generally carry limitations for the modeling of long-range dependencies that are present in an image; such limitations have been noted in computer vision tasks (17,18). In brain ventricle segmentation, for example, Valanarasu *et al.* (18) found that CNNs misclassified the background as a mask but long-range dependency learning could help to prevent the segmentation network from making this mistake. Due to the intrinsic locality of convolution operations, each CNN convolutional kernel pays attention to only several local pixels in the whole image, which forces the CNN to concentrate on local characteristics rather than learning the global patterns. *Figure 1* visualizes an example of lesion localization results of a CNN-based method and our method in the ChestX-ray14 dataset. In *Figure 1A*, the CNN misidentifies the lower lobe of the left lung as effusion. For a network to provide efficient localization, it should activate the class-specific image regions. However, due to the limitations in the modeling of long-range dependencies, the "lesions" identified by the CNN are only local abnormalities and not true lesions. Our proposed method does not make this mistake because it learns the long-range dependencies of the global context of the image, which helps the network to discriminate true abnormal regions and reduce false positives.

To address the problem of long-range dependencies, previous studies have proposed combining CNNs with, for example, attention mechanisms and bag-of-visual-words model, to further input features from CNNs into the attention module or use such features to define visual dictionaries for extracting the semantic context (19-21). However, these studies focused on the semantic context and spatial relationship between regions of interest (ROIs) and the features of CNNs rather than global attention on full-size images. Moreover, no previous study has focused on modeling long-range dependencies to improve the performance of medical image localization tasks. Using a transformer derived from natural language processing (NLP) for sequence-to-sequence prediction is a possible solution with an innate global self-attention mechanism (22). Following its success in NLP, the transformer has very recently been applied to computer vision applications (17,23). Dosovitskiy *et al.* (23) proposed the vision transformer (ViT), which achieved the state of the art in ImageNet (24) classification. However, a ViT trained on patches alone limits the network's ability to learn any dependencies for inter-patch pixels, which is insufficient for medical image detection (18).

Some existing research shows that the localization ability of a network is improved when the feature maps before

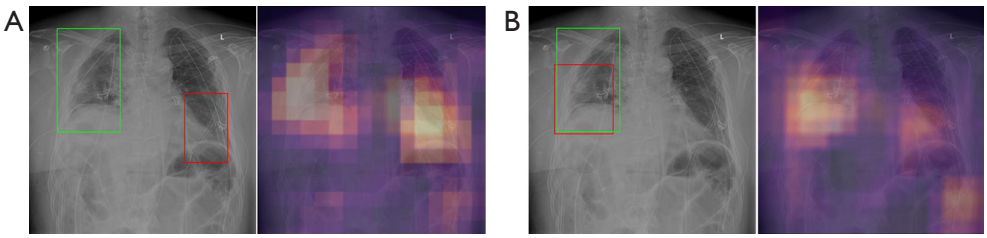

Figure 1 Visual localization results of a CXR from a case diagnosed with "effusion". The bounding boxes in green represent the published ground truth, and the red represents the predicted results of our proposed method. (A) The lesion localized by DenseNet121 (a CNN-based method). (B) The lesion localized by our proposed method. CXR, chest X-ray; CNN, convolutional neural network.

the global pooling layer have high spatial resolution (25). However, most previous weakly supervised approaches have typically applied deep network structures. The feature maps generated by these approaches usually lack sufficient spatial resolution, which prevents accurate lesion localization. Therefore, CNN-based deep networks generally have weak performances, especially when target structures in pathological images exhibit significant inter-patient variation in texture, shape, and size.

To address the above limitations, this work proposes a weakly supervised deep learning model with the following merits. First, to obtain fine-grained saliency maps in which both large and small lesions can be accurately delineated, we used a pyramid structure to upsample and combine the low-resolution features from multiple layers of DenseNet, a CNN-based network (26). Second, we adjusted the ViT to generate features that capture the global context of the whole image. To integrate the features from the pyramid DenseNet and the adaptive ViT, we designed a fusion module, which forced the network to focus on patch-wise and inter-patch dependencies simultaneously. Our model jointly generates disease classification and lesion localization by relying on saliency map detectors (8,27) and Log-Sum-Exp (LSE) pooling functions. Based on the official training and test set split, we experimented with our proposed method on the ChestX-ray14 dataset. Our proposed method achieved state-of-the-art disease classification and localization results in both quantitative and qualitative visual assessments. We also evaluated the generalization performance of our method in an external CXR dataset from the validation set of CheXpert. Our results suggest that weakly supervised object detection benefits from the transformer's ability to model long-range dependencies. We present the following article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-21-1117/rc).

Methods

Datasets

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

To train and evaluate our proposed model in disease classification and lesion localization, we performed our experiments on the large public CXR dataset in the United States, ChestX-ray14 (8). The ChestX-ray14 dataset contains 112,120 frontal-view CXRimages of 30,805 patients collected from the clinical picture archiving and communication system (PACS) database at the National Institutes of Health Clinical Center between 1992 and 2015. The dataset has 14 disease image labels that were text-mined from the associated radiological reports using NLP (where each image can have multi-labels). The 14 most common thoracic pathologies in the dataset are atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia. In the test set of this dataset, 880 images are annotated with bounding boxes relating to the first eight diseases mentioned above.

To externally evaluate the generalization performance of our method, we collected 123 CXR images from the validation set of another public CXR dataset, CheXpert (9). The CheXpert dataset consists of 224,316 chest radiographs of 65,240 patients that were collected from Stanford Hospital between October 2002 and July 2017. The dataset has 14 disease image labels which differ slightly from those of ChestX-ray14, including no finding, atelectasis, cardiomegaly, pleural effusion, pneumonia, pneumothorax, consolidation, edema, enlarged cardiomediastinum, lung opacity, lung lesion, pleural other, fracture, and support devices. The images selected for our study needed to contain at least one positive result for five diseases:

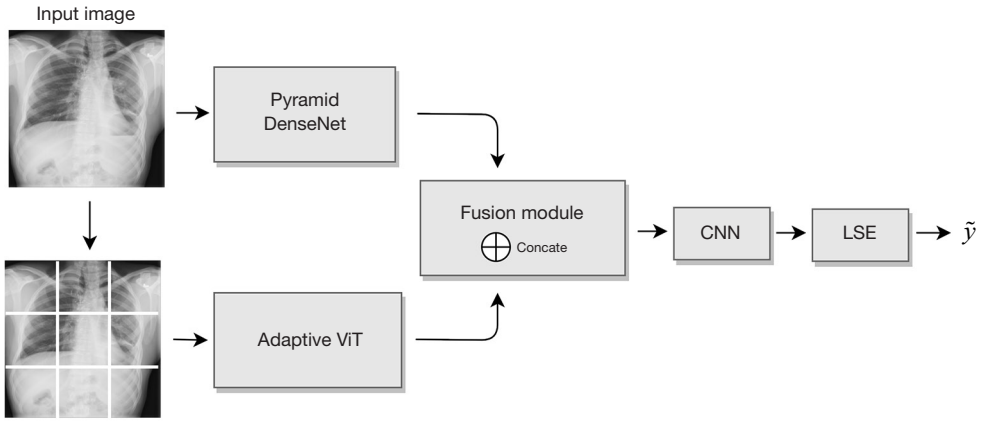

Figure 2 An overview of our proposed method. Our model comprises two branches and a fusion module. The classification score was acquired from the LSE layer, and the localization was obtained from the saliency maps before the LSE layer. ViT, vision transformer; CNN, convolutional neural network; LSE, Log-Sum-Exp.

atelectasis, cardiomegaly, pleural effusion, pneumonia, and pneumothorax, which overlapped with the first eight diseases for the ChestX-ray14 dataset. Also, based on the ground truth of the cases from the CheXpert validation set, two board-certified chest radiologists by Shanghai Municipal Health Commission [with 8 and 2 years of experience in computed tomography (CT) scan interpretation] performed a double-blind review to obtain bounding box annotations using the MarkMan software (http://www.getmarkman.com/).

Method overview

This work aimed to establish a weakly supervised detection model for thoracic disease classification and lesion localization which performs localization by drawing bounding boxes on CXRs. The model was trained on CXRs with only image-wise labels of 14 diseases. Our proposed approach was shaped by three ideas. First, to solve the problem of insufficient resolution caused by the deep network, we designed a DenseNet with a feature pyramid structure to generate fine-resolution feature maps. The reasons for choosing DenseNet as our CNN-based network branch were as follows: (I) due to feature reuse, DenseNet has a small number of parameters and efficient calculation, which was crucial to the parallel training of our proposed two network branches (28); (II) DenseNet improves the information and gradient propagation with dense connections, which makes the training of very deep networks tractable (28). Second, we adjusted the ViT to encode the global context features of the whole image and combined them with the features from the pyramid DenseNet for training together. Third, we encoded the features from previous stages into class-specific saliency maps and used an LSE pooling structure to pool each saliency map into a single disease probability score. Disease localization was achieved by calculating saliency maps in the forward propagation of the trained model. Therefore, our proposed method generated the features with fine resolution, modelled the long-range and inter-patch pixels dependencies, and combined the context of two branches for training. The network structure is illustrated in *Figure 2*.

Pyramid DenseNet

Huang *et al.* (28) proposed the DenseNet, which introduces direct connections from any preceding layers to all subsequent layers with the same feature map size. For each layer, the inputs consist of feature maps from all previous convolution blocks, and the outputs generated by this layer are passed to all subsequent layers. However, the repeated downsampling processes in dense blocks diminish the spatial resolution of feature maps. After several dense blocks in the basic structure have encoded, the output feature maps significantly decrease in size compared to the input image, losing the detailed semantic information.

To improve the localization ability of the network, we designed a pyramid DenseNet to generate fine-resolution feature maps. We retained the structure of DenseNet before the fully connected layer as our basic structure and

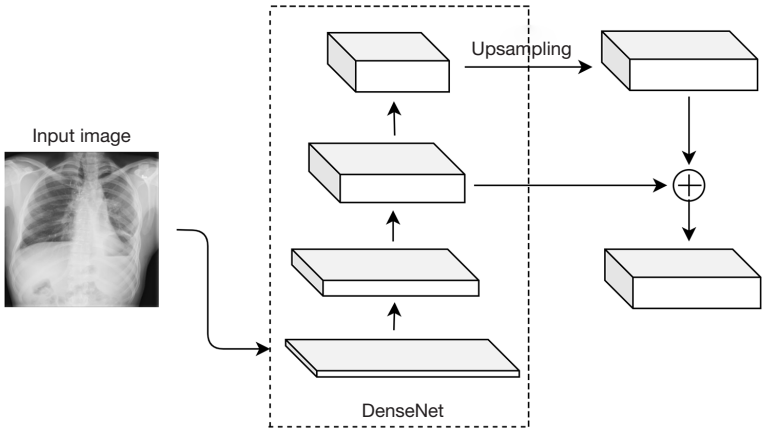

Figure 3 The architecture of the pyramid DenseNet. The network receives the input image and passes it via the main structure, which consists of four dense blocks interspersed with three transition blocks. After being generated by the last dense block, low-resolution feature maps were upsampled and fused with the outputs of the previous third dense block to obtain the fine resolution feature maps.

introduced the feature pyramid structure. Specifically, we upsampled the feature maps of the fourth dense block of DenseNet and fused them with the outputs of the third dense block. When the input image size was 512×512, the pyramid DenseNet output the feature maps with the spatial size of 64×64, which was much larger than the feature maps obtained after the final dense block of DenseNet (16×16). *Figure 3* shows the pyramid DenseNet in detail. We used DenseNet121 as our primary structure in this study. Note that we did not continue to upsample the fused feature maps of previous outputs, because a size increase in the feature maps of the ViT—another branch of our model introduced in the next section—could lead to huge resource consumption. With the pyramid DenseNet, we obtained fine-resolution feature maps.

Adaptive ViT

ViTs use a pure transformer model for image classification by embedding image patches that take on the role of words and extracting deep features of images (23). However, the patch-wise training of ViTs captures only the semantic information between patches and loses the information or dependencies for inter-patch pixels.

To model the patch-wise and inter-patch dependencies of the whole image in the network, we designed an adaptive ViT to capture global features and a fusion module to integrate the local features from the pyramid DenseNet and the global features from the adaptive ViT. Details of our adaptive ViT and fusion module are shown in *Figures 2*,*4*.

According to the work of Dosovitskiy *et al.* (23), several variants of ViT have been defined, and we used ViT-Base in our work. ViT-Base consists of a patch embedding layer, 12 transformer encoder layers comprising an attention layer and a feed-forward layer, and a classification head implemented by a multilayer perceptron (MLP) with a single hidden layer (23). The adaptive ViT maintains the ViT-Base structure before the classification head as the basic structure to obtain the embedded patch sequence. For the purpose of localization, our adaptive ViT recovers the spatial order of the patch sequence by reshaping it into two-dimension feature maps.

In the first stage, the adaptive ViT divides the image into non-overlapping square patches of size p^2 pixels, where p refers to the side length of the square patch and is designated as 16 in our work. Then, the adaptive ViT flattens the patches into vectors and linearly projects them to embedded patches. Along with the position embedding, the embedded patches are passed through a transformer encoder with 12 layers and obtained an embedded patch sequence. Finally, the patch sequence is reshaped into two-dimensional feature maps. Given the input image of size $H\times W\times 3$, our adaptive ViT obtains the feature maps of size $H/16\times W/16\times C$, where C refers to the number of feature map channels.

Our designed fusion module receives the feature maps from the pyramid DenseNet and adaptive ViT as inputs. As shown in *Figure 2*, the fused features are passed through a convolutional block that produces K saliency maps (one map for each class). The global branch adaptive ViT focuses on deep context information, and the local branch pyramid

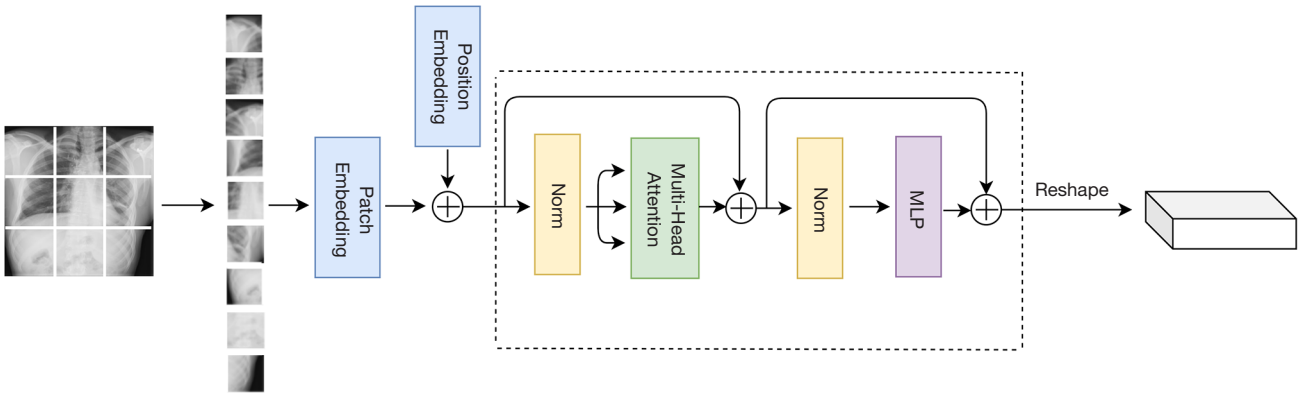

Figure 4 The architecture of the adaptive ViT. Norm, normalization; MLP, multilayer perceptron; ViT, vision transformer.

DenseNet focuses on fine-grained information. Our proposed method models patch-wise and inter-patch dependencies and combines the features of two branches for training.

LSE pooling

A specific weakly supervised learning technique was introduced into our method to accomplish accurate disease classification and lesion localization. The presence or absence of each pathology is denoted by y. As shown in *Figure 2*, we applied LSE pooling to pool each of the saliency maps into a single disease probability score, as denoted by $\tilde{y}$. We trained our network with binary cross-entropy loss between the prediction and the ground truth. The label distribution of our dataset was unbalanced, and there were many cases with no pathologies; therefore, we used positive and negative weight factors $W_P, W_N \in R^K$ to balance the label distribution for each disease. The loss function for each sample i is:

$$\ell(\tilde{\boldsymbol{y}}_i, \boldsymbol{y}_i) = -W_P \boldsymbol{y}_i \log(\tilde{\boldsymbol{y}}_i) - W_N (1-\boldsymbol{y}_i) \log(1-\tilde{\boldsymbol{y}}_i) \quad [1]$$

Where $W_P(c)=1-P_c/|G|$, $W_N(c)=P_c/|G|$, P_c indicates the number of positive cases for class c and $|G|$ indicates the training set size.

To obtain the bounding box of the disease, a thresholding technique was applied to segment the class-specific saliency maps generated by the CNN module in *Figure 2*. Based on ground-truth bounding box annotations, the thresholds used for segmentation were empirically determined and followed the criterion of previous approaches in this field (8,15). Therefore, our method is proposed for disease classification and lesion localization and requires only CXRs and coarse labeling of image-level annotations for network training.

Experiment

Dataset partition and experimental set-up

To maintain a fair comparison, we tested our proposed method in the ChestX-ray14 dataset and followed the official training-test set partitioning. For the selection of model parameters, we selected 10% of the training set as the validation set and repeated the random split 10 times to improve the reliability of the results. For localization assessment, all 880 CXRs and their ground-truth boundary boxes were used. The label distribution of the experimental data used for training and evaluation is shown in *Table 1*.

At the training stage, we used a learning rate of 0.00001 and the optimization of Adam (29), with a momentum of 0.9, weight decay of 0.001, and a minibatch size of 16. To improve the computational performance, inputs of 1,024×1,024 were downsampled to 512×512 and normalized using the ImageNet mean and standard deviation (24). During training, we applied simple image augmentation, in which each image was zoomed between 0 and 0.1, translated in four directions between −50 and 50 pixels, and rotated between −10 and 10 degrees.

Evaluation metrics

To evaluate the classification performance of our method, we calculated the accuracy, precision, sensitivity, and F1-score using the following equations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [2]$$

Table 1 The label distribution of training and evaluation data

Disease	Training set	Test set for classification	Test set for localization	External dataset
Atelectasis	7,194	3,255	180	74
Cardiomegaly	1,499	1,065	146	66
Effusion	7,479	4,648	153	64
Infiltration	11,920	6,088	123	–
Mass	3,546	1,712	85	–
Nodule	4,064	1,615	79	–
Pneumonia	749	477	120	8
Pneumothorax	2,324	2,661	98	6
Consolidation	2,412	1,815	–	–
Edema	1,172	925	–	–
Emphysema	1,236	1,093	–	–
Fibrosis	1,097	435	–	–
Pleural thickening	1,964	1,143	–	–
Hernia	130	86	–	–
Total	46,786	27,018	984	218

$$Precision = \frac{TP}{TP + FP} \quad [3]$$

$$Sensitivity = \frac{TP}{TP + FN} \quad [4]$$

$$F1score = 2 \times \frac{Presision \times Sensitivity}{Presision + Sensitivity} \quad [5]$$

We also calculated the area under the curve (AUC) for each class and the average of all classes to compare our classification results with those of other state-of-the-art methods (8,30,31). To evaluate the localization accuracy, we computed the intersection of the union (IoU) between the predicted bounding boxes of our proposed approach and the provided ground truth.

Results

Table 2 shows the classification performance of our proposed method for 14 diseases. *Table 3* compares the AUCs for classification between our approach and several baselines. For all 14 diseases, our model achieved better classification results than the model of Wang *et al.* (8). Our model also performed significantly better than that of Ma *et al.* (31) in 11 out of 14 diseases, and it outperformed Yao *et al.*'s (30) model in 9 out of 14 diseases. Furthermore, the AUCs of our model were more than 2% higher than those of the Yao *et al.* (30) and Ma *et al.* (31) models for "atelectasis", "infiltration", "nodule", "pneumonia", "pneumothorax", "emphysema", and "pleural thickening".

Table 4 compares the localization performance of our method with that of previously published methods under different IoU thresholds (8,32). Our method achieved better localization results than that of Wang *et al.* (8) in 6 of 8 diseases, which showed that the saliency maps were well fitted with the ground truth bounding boxes. For "atelectasis" and "nodule", our model was only slightly inferior to the existing CNN-based methods, which indicated that focusing on more local information may activate more accurate image regions for those two diseases. Furthermore, when compared to the model of Zhou *et al.* (32), which is based on a CNN and consists of a customized pooling structure and an adaptive DenseNet front-end, our model showed significant improvements in the localization of all 14 diseases. The localization results of the above methods and our proposed method are visualized in *Figure 5* (8,32). The "abnormalities" identified by Wang *et al.*'s method were usually multiple and covered large areas, and the results of Zhou *et al.* were small but incorrect. With CNNs as their backbone, these

Table 2 The classification performance of our proposed method for 14 diseases

Disease	Accuracy	Precision	Sensitivity	F1 score
Atelectasis	0.722±0.004	0.613±0.007	0.733±0.008	0.669±0.004
Cardiomegaly	0.865±0.003	0.727±0.014	0.810±0.013	0.766±0.011
Effusion	0.796±0.010	0.695±0.009	0.835±0.016	0.758±0.010
Infiltration	0.593±0.006	0.551±0.009	0.787±0.007	0.648±0.007
Mass	0.858±0.008	0.471±0.010	0.694±0.014	0.561±0.011
Nodule	0.850±0.005	0.622±0.012	0.664±0.012	0.642±0.010
Pneumonia	0.798±0.013	0.271±0.014	0.490±0.009	0.349±0.008
Pneumothorax	0.867±0.012	0.638±0.007	0.766±0.009	0.696±0.005
Consolidation	0.581±0.009	0.234±0.019	0.776±0.010	0.360±0.012
Edema	0.933±0.010	0.229±0.018	0.467±0.010	0.307±0.012
Emphysema	0.926±0.004	0.312±0.012	0.960±0.007	0.471±0.008
Fibrosis	0.886±0.005	0.153±0.012	0.367±0.018	0.216±0.016
Pleural thickening	0.867±0.007	0.275±0.008	0.630±0.014	0.383±0.012
Hernia	0.849±0.010	0.119±0.001	0.970±0.001	0.212±0.001
Mean	0.814±0.003	0.422±0.010	0.714±0.008	0.503±0.006

The results of mean ± standard deviation are reported by randomly splitting the training and validation sets for ten times.

Table 3 The AUC classification performance of our method (labeled as Ours) and several baselines for 14 diseases

Disease	Wang *et al*. (8)	Yao *et al*. (30)	Ma *et al*. (31)	Ours
Atelectasis	0.700	0.772	0.777	0.797±0.005*
Cardiomegaly	0.810	0.904*	0.894	0.872±0.016
Effusion	0.759	0.859*	0.829	0.852±0.004
Infiltration	0.661	0.695	0.696	0.711±0.008*
Mass	0.693	0.792	0.838	0.843±0.013*
Nodule	0.669	0.717	0.771	0.795±0.012*
Pneumonia	0.658	0.713	0.722	0.742±0.014*
Pneumothorax	0.799	0.841	0.862	0.894±0.006*
Consolidation	0.703	0.788*	0.750	0.779±0.008
Edema	0.805	0.882*	0.846	0.858±0.004
Emphysema	0.833	0.829	0.908	0.935±0.014*
Fibrosis	0.786	0.767	0.827*	0.825±0.014
Pleural thickening	0.684	0.765	0.779	0.800±0.007*
Hernia	0.872	0.914	0.934*	0.907±0.033
Mean	0.745	0.803	0.817	0.829±0.005*

The results of our method are reported with mean ± standard deviation by randomly splitting the training and validation sets for ten times. *, represents the best results. AUC, area under the curve.

Table 4 Comparison of disease localization accuracy using T(IoU), which measures the proportion of test images with different IoU thresholds: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7

T(IoU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
0.1	Wang	0.689*	0.938	0.660	0.707	0.400	0.139*	0.633	0.378
	Zhou	0.411	0.966	0.595	0.813*	0.529	0.076	0.742	0.327
	Ours	0.635	1.000*	0.748*	0.788	0.694*	0.070	0.786*	0.394*
0.2	Wang	0.472*	0.685	0.451	0.480	0.259	0.051*	0.350	0.235
	Zhou	0.239	0.945	0.340	0.675	0.306	0.013	0.575	0.204
	Ours	0.404	1.000*	0.664*	0.737*	0.429*	0.014	0.691*	0.277*
0.3	Wang	0.244*	0.459	0.301	0.277	0.153	0.038*	0.167	0.133
	Zhou	0.144	0.925	0.144	0.520	0.153	0.000	0.442	0.143
	Ours	0.205	1.000*	0.441*	0.525*	0.265*	0.000	0.548*	0.192*
0.4	Wang	0.094	0.281	0.203	0.122	0.071	0.013*	0.075	0.072
	Zhou	0.089	0.877	0.059	0.358	0.118	0.000	0.258	0.112
	Ours	0.103*	0.979*	0.273*	0.465*	0.184*	0.000	0.381*	0.128*
0.5	Wang	0.050*	0.178	0.111	0.065	0.012	0.013*	0.033	0.031
	Zhou	0.044	0.781	0.013	0.252	0.059	0.000	0.142	0.051
	Ours	0.045	0.873*	0.133*	0.343*	0.123*	0.000	0.333*	0.096*
0.6	Wang	0.022*	0.075	0.046	0.024	0.000	0.013*	0.017	0.031
	Zhou	0.011	0.521	0.000	0.155	0.024	0.000	0.042	0.041
	Ours	0.013	0.599*	0.063*	0.232*	0.061*	0.000	0.167*	0.043*
0.7	Wang	0.006	0.027	0.020	0.000	0.000	0.000	0.008	0.020
	Zhou	0.006	0.233	0.000	0.073	0.012	0.000	0.017	0.010
	Ours	0.006	0.261*	0.021*	0.101*	0.020*	0.000	0.095*	0.021*

T(IoU), different IoU thresholds; Wang, the results of Wang *et al*.'s (8); Zhou, the results of Zhou *et al*.'s (32); Ours, the results of our proposed method. *, best results among Wang *et al*.'s (8), Zhou *et al*.'s (32), and Ours.

methods focus on local patterns, which caused multiple of incorrect abnormalities to be identified. Both quantitative and qualitative experimental results showed that our method could effectively capture the lesions due to its learning of inter-patch and patch-wise dependencies and activation of the imaging regions used for classification globally. The localization results of our proposed method for some examples are visualized in *Figure 6*.

Discussion

In clinical procedures, it is insufficient that computer-aided chest radiography diagnostic systems only predict disease classification. Spatial localization of disease lesions provides visual evidence for the classification result, which is indispensable for auxiliary diagnosis. Disease localization potentially decreases the number of false positives and improves the diagnostic efficiency. Our weakly supervised disease detection method was trained only on image-level annotations but can generate disease classifications and lesion spatial localization simultaneously. In contrast to the previous state of the art, our method can identify global abnormalities in whole images by integrating ViT and DenseNet121 with a pyramid structure, which improves the localization performance.

To evaluate the generalization performance of our method, we calculated the AUCs for disease classification and IoU localization results in 123 external test cases

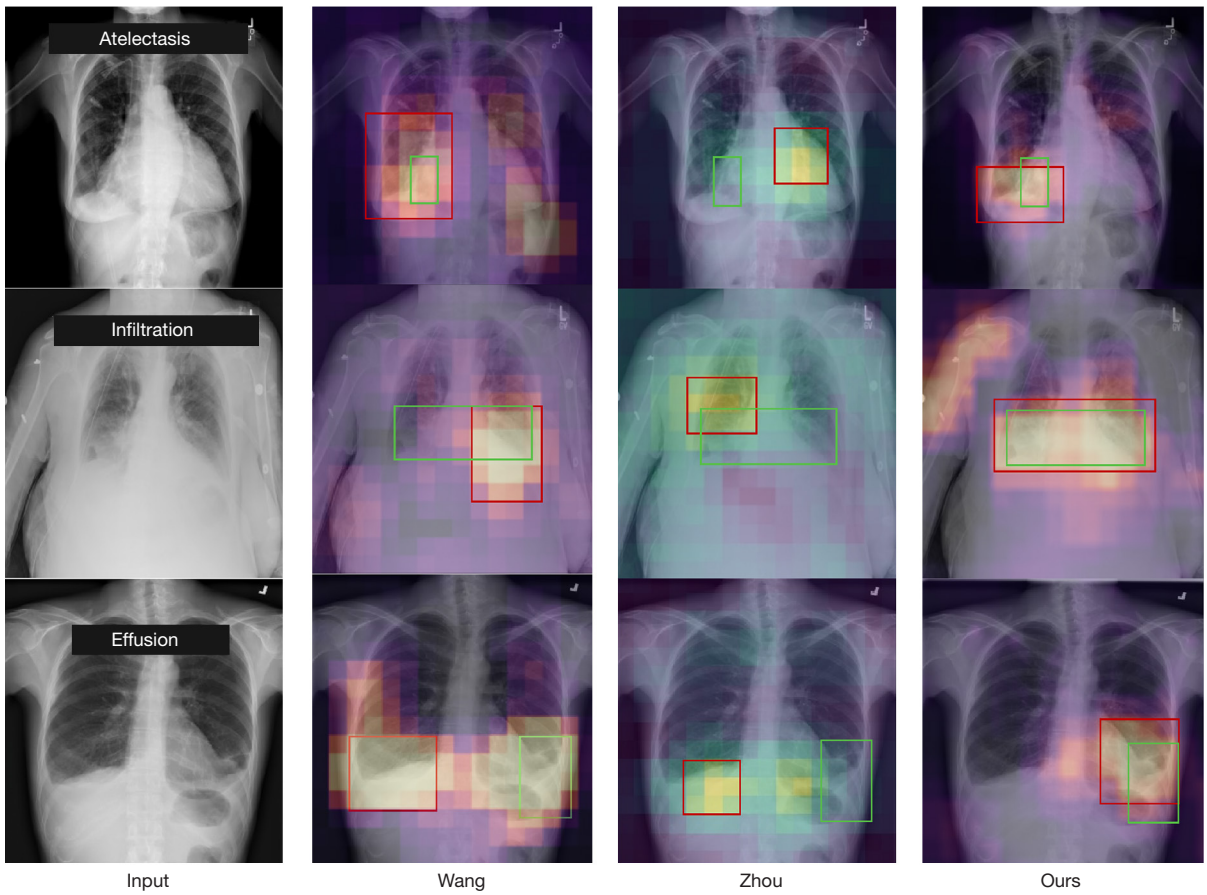

Figure 5 Several visualized localization results of Wang *et al.* (8), Zhou *et al.* (32), and our proposed method. The bounding boxes in green represent the published ground truth, and the red represents the predicted results using different methods.

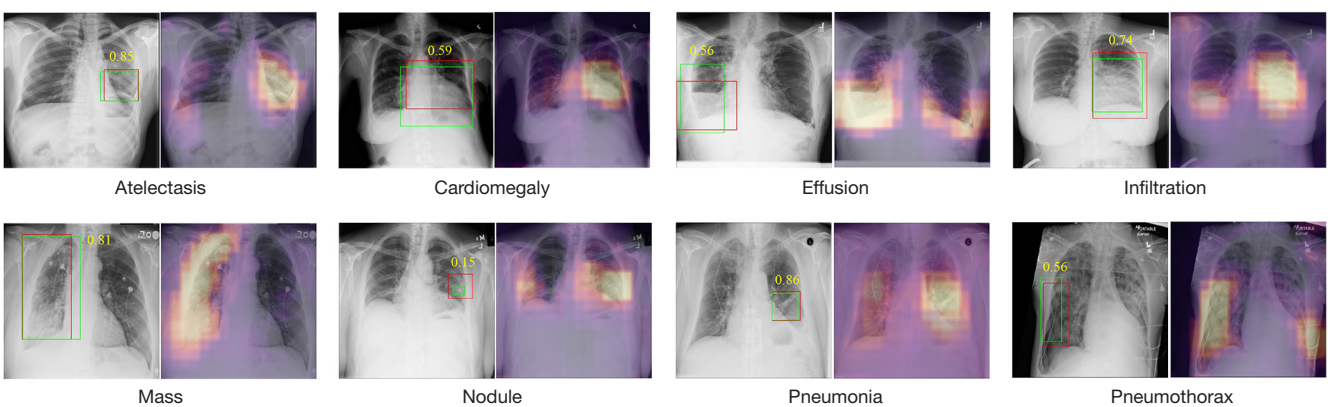

Figure 6 Examples of visual localization outcomes for eight diseases in the ChestX-ray14 dataset. The images on the left are the input CXRs and are labeled with two bounding boxes: the green shows the published ground truth, and the red is the predicted result obtained by a simple threshold method on the class-specific saliency maps, which are shown on the right. The IoU results calculated by ground truth and predicted bounding boxes are shown on the left. CXR, chest X-ray; IoU, intersection of the union.

Table 5 The AUC classification and IoU localization results of our method on external 123 test cases

T(IoU)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AUC
Atelectasis	0.554	0.486	0.419	0.324	0.230	0.081	0.014	0.739
Cardiomegaly	0.909	0.848	0.758	0.470	0.288	0.151	0.076	0.818
Effusion	0.672	0.500	0.250	0.110	0.016	0.000	0.000	0.802
Pneumonia	0.750	0.375	0.375	0.250	0.250	0.250	0.000	0.571
Pneumothorax	0.167	0.167	0.167	0.167	0.167	0.000	0.000	0.742

AUC, area under the curve; T(IoU), different IoU thresholds.

Table 6 The ablation study on three different backbones: DenseNet121, ViT, and our network (DenseNet121 + ViT)

T(IoU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
0.1	DenseNet121	0.596	0.993	0.678	0.707	0.674*	0.113	0.786	0.418*
	ViT	0.365	0.937	0.392	0.586	0.306	0.000	0.429	0.202
	DenseNet121+ViT	0.673*	1.000*	0.713*	0.808*	0.633	0.127*	0.833*	0.330
0.2	DenseNet121	0.321	0.923	0.566	0.636	0.469	0.000	0.691	0.259*
	ViT	0.179	0.887	0.147	0.374	0.102	0.000	0.286	0.149
	DenseNet121+ViT	0.372*	1.000*	0.615*	0.737*	0.510*	0.000	0.738*	0.245
0.3	DenseNet121	0.160*	0.606	0.364	0.414	0.326	0.000	0.571	0.145
	ViT	0.100	0.740	0.063	0.242	0.061	0.000	0.071	0.096
	DenseNet121+ViT	0.160*	0.958*	0.420*	0.616*	0.347*	0.000	0.643*	0.160*
0.4	DenseNet121	0.064*	0.232	0.217	0.303	0.225*	0.000	0.500	0.103
	ViT	0.058	0.486	0.014	0.172	0.020	0.000	0.071	0.032
	DenseNet121+ViT	0.064*	0.704*	0.259*	0.535*	0.204	0.000	0.524*	0.106*
0.5	DenseNet121	0.013	0.070	0.105	0.222	0.123	0.000	0.309	0.053
	ViT	0.032*	0.225	0.007	0.091	0.020	0.000	0.024	0.021
	DenseNet121+ViT	0.006	0.409*	0.112*	0.424*	0.163*	0.000	0.381*	0.085*
0.6	DenseNet121	0.000	0.021	0.028	0.111	0.032	0.000	0.095	0.043*
	ViT	0.006*	0.070	0.007	0.040	0.000	0.000	0.000	0.000
	DenseNet121+ViT	0.006*	0.190*	0.042*	0.232*	0.041*	0.000	0.262*	0.043*
0.7	DenseNet121	0.000	0.007	0.007	0.061	0.020*	0.000	0.024	0.011
	ViT	0.006*	0.014	0.006	0.010	0.000	0.000	0.000	0.000
	DenseNet121+ViT	0.006*	0.049*	0.028*	0.162*	0.020*	0.000	0.119*	0.021*

*, the best results of our experiments. ViT, vision transformer; DenseNet121+ViT, our proposed method which includes two network branches, DenseNet121 and ViT; T(IoU), different IoU thresholds.

(*Table 5*). The overall results showed that our method performed well in the external dataset, indicating its significant detection performance and excellent generalization ability. Interestingly, "pneumonia" only achieved an AUC of 0.571, which can be explained by our external dataset having only eight positive pneumonia cases; this makes this result biased and unreliable.

Table 6 shows the ablation study for the quantitative

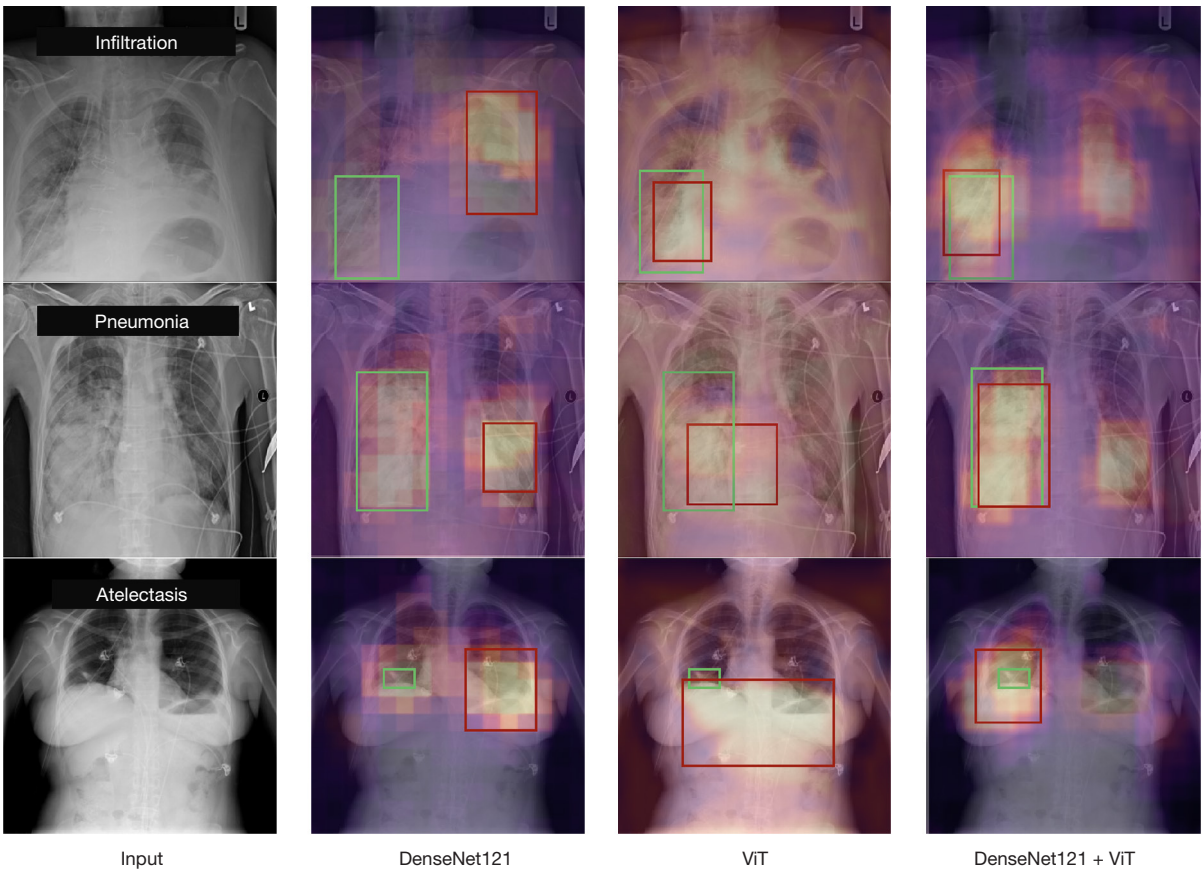

Figure 7 Several qualitative results of different feature encoder backbones in the ChestX-ray14 dataset: DenseNet121, ViT, and our network (DenseNet121+ViT). The bounding boxes in green represent the published ground truth, and the red represents the predicted results. ViT, vision transformer; DenseNet121+ViT, our proposed method which includes two network branches, DenseNet121 and ViT.

localization results of different feature encoders with DenseNet121, ViT, and our network (DenseNet121+ViT) as backbones. Qualitative evaluation of these three networks is shown in *Figure* 7. The predicted bounding boxes show that our network captures long-range dependencies and generates fine-grained feature maps extremely well. For example, as shown in the first row of *Figure* 7, DenseNet121 cannot detect real lesions because it only identifies locally suspected "abnormalities" and is confused about real lesions. However, as the ViT backbone encodes long-range dependencies, it learns to successfully localize in the whole image. As shown in the second and third rows, the ViT backbone network captures the lesion and its coarse location as the patch-wise training restricts the network in learning inter-patch and fine-grained information. As shown in the last column of *Figure* 7, our method takes long-range and inter-patch dependencies into account by integrating ViT and DenseNet121 with a pyramid structure, which makes its predictions accurate and detailed.

However, our method has some notable shortcomings. First, the image-level annotations in ChestX-ray14 were obtained using NLP technology rather than through manual observation by radiologists. The annotations used in our method training and validation are not completely correct. Therefore, we will explore advanced text mining methods to improve the accuracy of automatic tagging in our subsequent work (9). Second, due its integration of two network branches, our method requires much computing resources and exhibits low prediction efficiency. Therefore, another future improvement of our method is to develop a cascade network to encode patch-wise and inter-patch dependencies. Fewer parameters and simpler networks can improve the diagnosis efficiency of computer-aided diagnosis systems. Third, only a small external dataset with

123 CXR images was used for model validation, and the generalization ability of our method could not be effectively evaluated. Therefore, in our future work, we will collect more CXRs with annotations from clinical scenarios to expand the dataset for model validation and to evaluate the generalization of the model comprehensively and effectively.

Conclusions

Both quantitative and qualitative visual results showed that our proposed method achieves the new state-of-the-art for disease classification and weakly supervised localization. Compared to previous methods, our method produced superior results by using ViT to learn long-range dependencies between patches of the whole images, activating image regions used for classification globally, and avoiding mislocalization caused by the intrinsic locality of convolution operations. Furthermore, because ViTs can only learn inter-patch dependencies, we applied DenseNet121 along with a feature pyramid structure to encode high-resolution spatial information and learn inter-patch dependencies in our network. As an alternative framework to the dominant CNN-based approaches for medical image detection, our method achieves a superior performance.

Acknowledgments

We would like to thank the Shanghai Double First-Class University Construction and Development of High-Level Local Universities: Intelligent Medicine Emerging Interdisciplinary Cultivation Project and the Medical Research Data Center of Fudan University for providing financial assistance and technical advice.
Funding: This work was supported by the National Natural Science Foundation of China (No. 91846302) and the S&T Program of Hebei (No. 21377734D).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-21-1117/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-21-1117/coif). FL has a pending patent (No. 202110640995.8). The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The datasets used in this study are all public CXRs. No ethical approval and written informed consent were required.

References

1. Körner M, Weber CH, Wirth S, Pfeifer KJ, Reiser MF, Treitl M. Advances in digital radiography: physical principles and system overview. Radiographics 2007;27:675-86.
2. Hui TCH, Khoo HW, Young BE, Haja Mohideen SM, Lee YS, Lim CJ, Leo YS, Kaw GJL, Lye DC, Tan CH. Clinical utility of chest radiography for severe COVID-19. Quant Imaging Med Surg 2020;10:1540-50.
3. Yan C, Yao J, Li R, Xu Z, Huang J. Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays. Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; Washington, DC, USA: Association for Computing Machinery; 2018:103-4.
4. Kim HG, Lee KM, Kim EJ, Lee JS. Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus X-ray using multiple deep learning models. Quant Imaging Med Surg 2019;9:942-51.
5. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.
6. Setio AAA, Traverso A, de Bel T, Berens MSN, Bogaard CVD, Cerello P, et al. Validation, comparison, and

combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 2017;42:1-13.
7. Qin C, Yao D, Shi Y, Song Z. Computer-aided detection in chest radiography based on artificial intelligence: a survey. Biomed Eng Online 2018;17:113.
8. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. editors. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017:3462-71.
9. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proceedings of the AAAI Conference on Artificial Intelligence; 2019:590-7.
10. Li P, Sun M, Wang Z, Chai B. OPTICS-based Unsupervised Method for Flaking Degree Evaluation on the Murals in Mogao Grottoes. Sci Rep 2018;8:15954.
11. Shahid KT, Schizas I. Unsupervised Mitral Valve Tracking for Disease Detection in Echocardiogram Videos. J Imaging 2020;6:93.
12. Zadeh Shirazi A, Seyyed Mahdavi Chabok SJ, Mohammadi Z. A novel and reliable computational intelligence system for breast cancer detection. Med Biol Eng Comput 2018;56:721-32.
13. Kumar P, Grewal M, Srivastava MM, editors. Boosted Cascaded Convnets for Multilabel Classification of Thoracic Diseases in Chest Radiographs. In: International Conference Image Analysis and Recognition; Springer; 2018:546-52.
14. Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification. arXiv preprint arXiv:1801.09927v1 2018.
15. Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Thorax disease classification with attention guided convolutional neural network. Pattern Recognition Letters 2020;131:38-45.
16. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Object Detectors Emerge in Deep Scene CNNs. Proceedings of the 2015 International Conference on Learning Representations; 2015.
17. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306v1 2021.
18. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM, editors. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer; 2021:36-46.
19. Sitaula C, Hossain MB. Attention-based VGG-16 model for COVID-19 chest X-ray image classification. Appl Intell (Dordr) 2021;51:2850-63.
20. Sitaula C, Aryal S. New bag of deep visual words based features to classify chest x-ray images for COVID-19 diagnosis. Health Inf Sci Syst 2021;9:24.
21. Huang Z, Wang X, Wei Y, Huang L, Shi H, Liu W, Huang TS. CCNet: Criss-Cross Attention for Semantic Segmentation. IEEE Trans Pattern Anal Mach Intell 2020. [Epub ahead of print]. doi: 10.1109/TPAMI.2020.3007032.
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017:30.
23. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929v2 2020.
24. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60:84-90.
25. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. editors. Learning Deep Features for Discriminative Localization. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016:2921-9.
26. Lin T, Dollár P, Girshick R, He K, Hariharan B, Belongie S, editors. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017:2117-25.
27. Yao L, Prosky J, Poblenz E, Covington B, Lyman K. Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions. arXiv preprint arXiv:1803.07703v1 2018.
28. Huang G, Liu Z, Maaten LVD, Weinberger KQ, editors. Densely Connected Convolutional Networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017:4700-8.
29. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980v5 2014.
30. Yao L, Poblenz E, Dagunts D, Covington B, Bernard

D, Lyman K. Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501v2 2017.
31. Ma C, Wang H, Hoi SCH, editors. Multi-label Thoracic Disease Image Classification with Cross-Attention Networks. arXiv preprint arXiv:2007.10859v1 2020.
32. Zhou B, Li Y, Wang J. A Weakly Supervised Adaptive DenseNet for Classifying Thoracic Diseases and Identifying Abnormalities. arXiv preprint arXiv:1807.01257v2 2018.

Cite this article as: Li F, Zhou L, Wang Y, Chen C, Yang S, Shan F, Liu L. Modeling long-range dependencies for weakly supervised disease classification and localization on chest X-ray. Quant Imaging Med Surg 2022;12(6):3364-3378. doi: 10.21037/qims-21-1117

