Learning to Recognize Thoracic Disease in Chest X-Rays With Knowledge-Guided Deep Zoom Neural Networks

KUN WANG, XIAOHONG ZHANG, SHENG HUANG, (Member, IEEE), FEIYU CHEN, XIANGBO ZHANG, AND LUWEN HUANGFU

1Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing 400030, China
2School of Big Data and Software Engineering, Chongqing University, Chongqing 400030, China
3Fowler College of Business, San Diego State University, San Diego, CA 92182-8230, USA

Corresponding author: Xiaohong Zhang (xhongz@cqu.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2019CDYGYB014, in part by the National Key Research and Development Project of China under Grant 2018YFB2101200, in part by the Chongqing Major Theme Projects under Grant csc2018jszx-cyztzxX00017, and in part by the National Natural Science Foundation of China under Grant 61772093.

ABSTRACT Automatic and accurate thorax disease diagnosis in Chest X-ray (CXR) image plays an essential role in clinical assist analysis. However, due to its imaging noise regions and the similarity of visual features between diseases and their surroundings, the precise analysis of thoracic disease becomes a challenging problem. In this study, we propose a novel knowledge-guided deep zoom neural network (KGZNet) which is a data-driven model. Our approach leverage prior medical knowledge to guide its training process, due to thoracic diseases typically limit within the lung regions. Also, we utilized weakly-supervised learning (WSL) to search for finer regions without using annotated samples. Learning on each scale consists of a classification sub-network. The KGZNet starts from global images, and iteratively generates discriminative part from coarse to fine; while a finer scale sub-network takes as input an amplified attended discriminative region from previous scales in a recurrent way. Specifically, we first train a robust modified U-Net model of lung segmentation and capture the lung area from the original CXR image through the Lung Region Generator. Then, guided by the attention heatmap, we obtain a finer discriminative lesion region from the lung region images by the Lesion Region Generator. Lastly, the most discriminative features knowledge is fused, and the complementary features information is learned for final disease prediction. Extensive experiments demonstrate that our method can effectively leverage discriminative region information, and significantly outperforms the other state-of-the-art methods in the thoracic disease recognition task. Furthermore, the proposed KGZNet can gradually learn the discriminative region from coarse to fine in a mutually reinforced way. The code is will available at: https://github.com/ISSE-AILab/KGZNet.

INDEX TERMS Chest X-ray, thoracic disease recognition, weakly-supervised learning, discriminative feature learning, biomedical imaging.

I. INTRODUCTION CHEST X-ray (CXR) is one of the most common types of medical imaging examinations globally, effectively assisting the clinical diagnosis and treatment for a series of thoracic diseases, including lung cancer, tuberculosis, and pneumonia, etc. The CXR images are typically interpreted by expert radiologists in clinical practice. However, this task is prone to subjective diagnostic errors and time-consuming [1]. More importantly, false positives can cause patient anxiety, unnecessary follow-up, and invasive diagnostic procedures. Diseases that are missed during the screening process may not be determined until the condition is aggravated and unsuitable for treatment [2]. Therefore, a computer-aided diagnosis (CAD) system for multi-label thoracic disease classification can improve patient care by providing better information on disease progression and helping the radiologist reduce a lot of time. In recent years, deep learning...
technology has achieved remarkable success in computer vision because it provides a unified feature extraction classification framework, liberating users from tedious manual feature extraction [3]–[5]. The success achievement widely draws many investigators’ attention to apply deep convolutional neural networks (CNNs) in medical image analysis. For example, disease classification [6]–[9], lesion segmentation or detection [10]–[13], image registration [14], [15], and so on. In this paper, we explore the classification task of thoracic disease in CXR images using deep learning.

Several existing state-of-the-art works have been proposed to diagnose the thoracic diseases in CXR images automatically. In 2017, Wang et al. [16] first present the largest publicly available chest X-ray dataset, namely “Chest X-ray 14”, which contained 14 common thoracic diseases. Many research works [17]–[20] are based on this large-scale dataset. In general, most of the previous works used the original CXR images as input for features extractor network learning. But, this training strategy is limited on the following points. On the one hand, the presence of irrelevant objects and the poor alignment of some CXR images hinder the network performance. On the other hand, a common method to alleviate the computational burden is resizing original CXR pictures to a low resolution in training CNNs. However, these operations will result in loss of CXR images details, which may be crucial for diagnosing pathologies, especially small ones (e.g., “Atelectasis,” “Nodule,” and “Mass”). Furthermore, some researchers [21]–[23] focus on determining the localization of the lesion region with image-level or limited supervision. Li et al. [19] proposed a unified model for disease classification and localization in a combined model with limited lesion area labeling information. Yao et al. [24] combined multi-resolution and multi-instance features extractor learning with a customized pooling function to provide more accurate diagnosis and higher resolution pathologically significant maps. In the latest study, Zhang et al. [25] proposed another method to learn distinguishing features in triple images and perform cyclic training on regional features to verify whether the region of interest contains indications of disease. Although promising performances have been reported, further improvement suffers from the following limitations. First, these research methods did not make the most of the features information from different discriminative regions that can be mutually reinforced and do not use prior domain knowledge. Second, subtle disease visual features that existed in the local region are still challenging to learn.

To deal with the above issues, we propose a novel knowledge-guided deep zoom neural networks (KGZNet) framework. This model consists of three scale scope features extractor and a joint fusion learning branch. First of all, it is worth noting that accurate discriminative region localization can effectively improve region-based feature learning to help recognize, and vice versa. This issue requires the assistance and guidance of prior medical knowledge. We believe that deep learning is essentially an algebraic computing system, not the most effective way to acquire highly complex human knowledge. In fact, thoracic diseases are usually limited to the lung fields. For example, “Nodule,” “Effusion,” “Pneumothorax,” and so on. Therefore, if lung field can provide accurate localization, the lung region image can be used to train the network to exclude noisy areas in the original image (i.e., poor alignment and irrelevant objects). Then, we expanded the focus to the lesion region of the specific disease. Besides, since some features of the disease may require structural information related to the overall CXR image, the whole CXR image is also considered. Therefore, as the main core of our framework, it shows that the information of pathological features in different scopes regions are fully effectively fused and thus get more productive knowledge about the samples. The discriminative region detection and finer feature learning are mutually correlated and thus can reinforce each other. The proposed KGZNet can be guided by knowledge from coarse to fine to zoom in to the most discriminative regions gradually (e.g., from the global region to the lesion region). Our method working mechanism is similar to a radiologist’s visual attention and comprehensive analysis during diagnosis. It is well accepted that radiologists need to consider all discriminative regions information in clinical practice is illustrated in Figure 1. Note that all the regions of interest (ROI) are obtained by deep learning instead of manual. To the best of our knowledge, this work represents the first attempt to propose a multi-scale visual enhancement diagnosis network based on the perspective of prior clinical experience. And our KGZNet can achieve superior classification performance as state-of-the-art methods on the Chest X-ray 14 dataset: the average AUC scores for 14 pathologies is 0.878.

In summary, the main contributions of this paper can be summarized as follows:
(1) We address the challenge of thoracic disease recognition by proposing a novel knowledge-guided deep zoom neural network (KGZNet) that can gradually zoom into the most discriminative regions from coarse to fine, which can get rich deep features knowledge about the sample to obtain a more robust solution.

(2) With multi-stream feature representation and discriminative region feature integration. Our proposed network can reduce many false positives. The proposed fusion model can further leverage the advantage of the relation of discriminative region learning, global and local features from multiple scales are deeply fused to diagnose disease.

(3) The proposed Lung Region Generator is captured the lung regions from global CXR images, guided by the prior domain knowledge. Then, we augmented focus on the more fine-grained lesion area in the lung regions, which our proposed Lesion Region Generator did. Furthermore, we find that all the discriminative region detection are mutually correlated and reinforce each other.

(4) Extensive experiment results evaluate our method on the large-scale public NIH Chest X-ray 14 dataset. Our method outperforms other state-of-the-art multi-label thoracic disease recognition methods.

A preliminary version of this work was presented in BIBM 2019 [26]. In this paper, we have substantially revised and extended the original paper. The main extension includes: (1) the ablation study of different scale training strategies, (2) discussion of various parameters of the network in detail, (3) network training strategies and algorithm implementation, (4) the performance of lung segmentation and lesion localization, and (5) discussion of the limitation of the proposed method, and verify the robustness of the method on another Chest X-rays dataset.

The rest of the paper is organized as follows. A brief review of the related work in Section II and Section III gives a detailed description of the proposed method. Then, we explain the datasets, experiments, and results in Section IV. In Section V, we make a full ablation study. In Section VI, we discuss the limitation of the proposed method, and verify the method’s robustness on another Chest X-rays dataset. Finally, we conclude our proposed research work in Section VII.

II. RELATED WORK
A. DEEP LEARNING FOR MULTI-LABEL CHEST X-RAYS RECOGNITION
Deep learning have shown promise in the field of medical image analysis [6]–[8], [14], [28], [29], especially in multi-label chest X-rays recognition [17], [18], [21], [30]. Yao et al. [18] proposed a method combining Long-short Term Memory Network (LSTM) and DenseNet network [31] to predict thoracic disease through label correlation. Kumar et al. [32] implemented a cascade network that relies on label-correlation to improve the performance of disease classification. Rajpurkar et al. [20] developed an algorithm called CheXNet, which trained a fine-tuning 121-layer DenseNet [33] architecture pre-trained on ImageNet [34] dataset and got a good performance in pneumonia detection. Wang et al. [35] proposed TieNet, which could improve the performance of thoracic disease classification with additional radiology report embedding information. This article is different from previous work. According to the radiologist reading the CXR image program, we use more visual and semantic information in the identification area to improve the recognition performance of thoracic diseases, without using auxiliary diagnostic reports and lesion location information.

B. DISCRIMINATIVE FEATURE LEARNING
Learning discriminative features is crucial for multi-label CXR image classification to distinguish different pathologies. Previous works mainly focus on the extra annotations of the bounding box and part annotations to localize the lesion region. However, in clinical practice, data annotations are usually expensive and time-consuming to be obtained in the medical domain. Furthermore, some researchers [17], [22], [23], [36] attempted to use various attention mechanisms to capture the discriminative features from the informative lesion areas. For example, Ypsilantis and Montana [22] proposed a recurrent attention model (RAM) that is capable of sampling the whole CXRs sequentially and focus on the most informative areas. They considered only one disease “enlarged heart” in their work. More recently, Tang et al. [23] presented an iterative attention-guided refinement framework further to improve the weakly-supervised localization and disease classification performance. Guan et al. [21] proposed an attention guided CNN (AG-CNN) to combine the global and local cues for disease classification. This paper is inspired by those thoracic diseases that are generally limited to the lung regions and interdependence between lesion localization. We propose to learn the most discriminative areas guided by the prior clinical knowledge and finer region-based feature representation on the multi-scale branches. Compared with the previous work which relies on bounding box annotations to localize the discriminative regions, our method only utilizes the medical domain knowledge-guided and image-level labels. This way can effectively reduce human resources and improve diagnostic efficiency in practical application.

III. THE PROPOSED METHODS
A. KGZNet ARCHITECTURE
The proposed KGZNet algorithm consists of three major steps: (1) extracting multi-scale optimal zoom-in view, (2) extracting the global, lung, and lesion regions on two-dimension (2D) CXR images, (3) constructing three KGZ submodels and training each of them using the most
discriminative regions, creating and training the MS-KGZ model for thoracic disease classification. Note that all the regions of interest (ROI) are obtained by deep learning instead of manual. The detailed process is described in III-C.

A chart that summarizes this algorithm is shown in Figure 2.

The overview of the proposed method is shown in Figure 3. We analyze the network with three scales as an example, and finer-scale regions can be stacked in a similar way. The framework for thoracic recognition, which gradually zooms into the most discriminative region from coarse to fine. Specifically, the algorithm goes through roughly four parts. We first train the globalNet branch using global CXR images to learn the overall visual features and structural information. Second, guided by the prior medical knowledge that those pathologies are usually limited to the lung regions. U-Net is used to segment lung field from the global region images and the lung region images, which were cropped and resized by the Lung Region Generator (LRG-1). Then, guided by the attention heatmap, the Lesion Region Generator (LRG-2) was used to obtain the medical discriminative lesion region from the lung area images. Finally, the feature fusion module concatenates the global average pooling layers of three different scope feature extractors and is fine-tuned for the final thoracic disease classification.

![Figure 2](image1.png)

**FIGURE 2.** (a) Captures of multi-scale optimal zoom-in view (Global, Lung, and Lesion). (I) Radiologist’s clinical screening process. (II) The automatic system which mimics the diagnostic process of a radiologist. Note that all the regions of interest (ROI) are obtained by deep learning instead of manual. (b) Workflow of the proposed discriminative region extraction. (c) Multi-scale KGZ sub-models fusion learning.

![Figure 3](image2.png)

**FIGURE 3.** The framework of our proposed KGZNet. It consists of three scale-feature branches and one fusion branch, i.e., the classifier of global branch $C_{gb}$, lung branch $C_{lg}$, lesion branch $C_{ln}$, and fusion branch $C_{fu}$. The inputs are from coarse global images $F_1$ to finer lesion region $F_3$ (from top to bottom). The final prediction lesion area is visualized using CAM [27] technology.
B. MULTI-LABEL SETUP

In our work, we first define each CXR image with a $T$-dimension vector $L = [l_1, l_2, \ldots, l_T]$, where $T = 14$ and $l_T \in \{0, 1\}$, $l_T$ represents whether there is any pathology, (i.e., 0 for absence and 1 for presence.) When the $L$ is a vector of all-zeros, it means that none of the diseases is found within the range of any of 14 disease categories as listed. This definition transforms the multi-label classification problem into a regression-like loss setting. The label of “No Finding” is not considered in this work. Furthermore, it is noted that each CXR image in Chest X-ray 14 database is labeled with one or more pathologies.

In this paper, the feed to the globalNet branch is the whole input global image $x$ of $[7 \times 224 \times 224 \times 3]$ to the proposed KGZ sub-model, which contains three pre-trained ResNet-50 networks (See Figure 4). The subnet used for this work includes 50 learnable network layers. It includes a $7 \times 7$ convolutional layer, generating 64 feature maps, a $3 \times 3$ max-pooling layer, four down-sampling bottlenecks architectures, followed by a global average pooling (GAP) layer and a neural network with 14 neurons’ fully connected (FC) layer. Note that each bottleneck architecture includes three convolutional layers with kernel sizes of $1 \times 1$, $3 \times 3$, and $1 \times 1$, respectively (See Table 1).

1) GlobalNet BRANCH

In this paper, the feed to the globalNet branch is the whole CXR images take down-sampled low-resolution. Given an input global image $F_1 \in \mathbb{R}^{224 \times 224 \times 3}$, in the globalNet branch, we use a modified of ResNet-50 [37] as the backbone of our GlobalNet. As mentioned before, the sub-model consists of five down-sampling blocks, followed by a GAP layer ($\mathbb{R}^{1 \times 1 \times 2048}$) and a 14-dimension FC layer for thoracic disease classification. The output vector $p_{gb}(y_i|x_i)$ of the FC layer was normalized to $[0,1]$ using a sigmoid function, which is defined as follows:

$$p_{gb}(y_i|x_i) = 1/[1 + \exp(-f(x_i))]$$

(1)

where $\hat{p}_g(y_i|x_i)$ is the predicted confidence score of $i$-th category, $i \in \{1, 2, 3, \ldots, T\}$. $f(x_i)$ is the $i$-th element at the fully-connected layer.

In addition, due to the unbalanced distribution between samples, we introduce a weighted binary cross-entropy (W-BCE) loss function $L(y, \hat{y})_{gb}$ to alleviate the problem of pathologies imbalance, is as follows:

$$L^*_{(y, \hat{y})} = -w_P \sum_{y_i=1}^{T} \left[ \hat{y}_i \log(p_{gb}(y_i|x_i)) \right]$$

$$-w_N \sum_{y_i=0}^{T} \left(1 - \hat{y}_i \right) \log(1 - p_{gb}(y_i|x_i))$$

(2)

where $T$ is the number of thoracic diseases (i.e., $T = 14$), $y_i$ refer as the ground truth label of the $i$-th class. $w_P$ is set to $|P|/|P|+|N|+1$ while $w_N$ is set to $|N|+1/|P|+|N|+1$. $|P|$ and $|N|$ are the total numbers of “Positives(1)” and “Negatives(0)” in a batch of image labels.

2) LungNet BRANCH

To capture the prominent locality of the lung regions, we also adopt the same convolutional network structure with the globalNet branch as the backbone of our LungNet. Due to the fact that those thoracic diseases limited within lung regions. The input of the lungNet branch using the lung region images $F_2 \in \mathbb{R}^{224 \times 224 \times 3}$, which are located and cropped via the proposed Lung Region Generator in Section III-D. Specially, we conduct the same normalization and optimization as the globalNet branch. We denote the predicted probability of lungNet branch as $p_{lg}(y_i|x_i)$, and the loss function of the lung region branch can be defined as $L(y, \hat{y})_{lg}$.

3) LesionNet BRANCH

In this part, the lesionNet branch focuses on the specific lesion area and is expected to pay more attention to the characteristics of the smaller pathological lesions (i.e., “Mass,” “Nodule”). In more details, we first inference a heatmap to guide attention and crop a discriminative region from the lung region image via the proposed Lesion Region Generator in Section III-E. Then, the lesion region images $F_3 \in \mathbb{R}^{224 \times 224 \times 3}$ is generated as input to train the lesionNet branch. Note that the network possesses the same structure as the lungNet branch. We represent the predicted probability of lesionNet branch as $p_{ln}(y_i|x_i)$ and the loss function defined as $L(y, \hat{y})_{ln}$.

4) FusionNet BRANCH

In particular, during the classification process, we can get a multi-scale representation from full-scale global images to finer disease region attention. The image $F$ can be represented a set of multi-scale $\{F_1, F_2, \ldots, F_m\}$ representations, and $m$ is the total number of region scales. Specifically, we first concatenate the global average pooling outputs of the globalNet, lungNet, and lesionNet branches to obtain the pooled feature maps of dimension $\mathbb{R}^{1 \times 1 \times 2048 \times 3}$ on the fusionNet branch. Then, the concatenated layer is connected to a 14-dimensional FC layer for final thoracic disease classification. The predicted probability of fusion branch is
denoted as $\bar{p}_{fu}(y_i|x_i)$. The loss function of the fusion branch is $L(y, \bar{y}_{fu})$. Overall, our fundamental objective to train the whole framework can be defined as:

$$
\min_{\Theta, W} L_{obj} = \psi_1 L_{gb} + \psi_2 L_{ig} + \psi_3 L_{in} + \psi_4 L_{fu}
$$

where $\psi_1, \psi_2, \psi_3$, and $\psi_4$ are weights for each part of the loss. The definition of $L_{gb}, L_{ig}, L_{in}$, and $L_{fu}$ denotes the loss function of the three-scale branch, and fusionNet branch, respectively.

Algorithm 1 FusionNet Branch Joint Learning Process
1: Input image $F_i$, the label vector of disease $L_i$
2: Set the hyperparameters $\psi_i, \rho, \theta_i$, and the maximum number of epochs $N$
3: Initialize the weights of global, lung and lesion network $W = \{W_1, W_2, W_3\}$ and $\Theta$;
4: concatenating $Pool_{gb}, Pool_{ig}$ and $Pool_{in}$ learning $W$
5: for $t = 1 : N$ do
6: Compute the output $\hat{X}$ via forward propagation
7: Compute the predicted assignment $P$
8: Update $W$ and $\Theta$ via backward propagation
9: end for

D. LUNG REGION GENERATOR

1) U-NET FOR LUNG SEGMENTATION

In this section, the purpose of lung segmentation is to segment the lung field and remove other unrelated features for better disease classification. In this way, the impact of the poor alignment and noise in non-disease regions can be alleviated. Specifically, we randomly selected 1000 CXR images from the public Chest X-ray 14 dataset of the NIH (National Institutes of Health). With the help of a radiologist at Chongqing University Cancer Hospital, the lung masks were roughly labeled with 1,000 training CXR images. The total images are randomly split into 80% for training and 20% images for validation. We use the U-Net model, a fully convolutional network (FCN), for fast and precise image segmentation. The architecture details for the modified U-Net is shown in Figure 5. Note that the network structure is slightly different from the original article [38]. Compared with the original U-Net decoder, some minor modifications have been made. Each block in the decoder performs $2 \times 2$ bilinear upsampling on the input features. In the training phase, the input original CXR images and lung masks were downscaled to the size of $512 \times 512$ pixels images and then fed into the U-Net. Finally, a batch of probability matrix with dimensions of $\mathbb{R}^{512 \times 512 \times 1}$ is output to denote the final lung segmentation result. Moreover, the lung segmentation model is optimized using Adam optimizer and a mini-batch size of eight. The initial learning rate is set to 1e-3 and weight decay 1e-4.

2) LUNG REGION GENERATION

As can be seen in Figure 2(b), we first use the modified U-Net model to obtain the lung field prediction mask $M_f$ from global region CXR images. The result of the lung segmentation mask can be regarded as a binary classification. Then, we calculate a maximum connected region based on the predict lung field mask, which covers the discriminative points in $M_f$. The maximum connected region is represented by the minimum and maximum coordinates on the horizontal and vertical axis $[w_{min}, h_{min}, w_{max}, h_{max}]$. Finally, according to the boundary of the connected region as a reference point, the bounding boxes of the lung field were cropped from the

![FIGURE 5. The architecture for lung segmentation.](image-url)
global region CXR images and zoom in to $224 \times 224$ pixels used as the lung region images for the lungNet branch of the proposed KGZNet.

### E. LESION REGION GENERATOR

To better focus on the suspicious lesions and see clearly, especially in potentially small lesion areas. As shown in Figure 2(b), given the prior work [39], we use the method of visualizing the disease-specific class activation map (CAM, or heatmap). In this way, disease localization is performed through classification-trained CNNs and weakly-supervised learning (WSL) without using any bounding box annotation. Specifically, we first input the lung region image into the lungNet branch of the proposed KGZNet, and then learn and extract the feature maps of the last convolution layer through classification-trained CNNs and weakly-supervised learning (WSL) without using any bounding box annotation.

Specifically, we first input the lung region image into the lungNet branch of the proposed KGZNet, and then learn and extract the feature maps of the last convolution layer through the lung branch network. Let $f^d_l(x, y)$ denotes the $d$-th feature map at spatial position $(x, y)$ at the last convolution layer, where $d \in \{1, \ldots, D\}$, $D = 2048$ in ResNet-50 (or $D = 1024$ in DenseNet-121). We define $l$ represents the lungNet branch, $D$ is the number of feature maps. The attention heatmap $Z_l$ is produced by computing the maximum values along with feature maps, shown as,

$$Z_l(x, y) = \max_d (f^d_l(x, y)), \quad d \in \{1, \ldots, D\} \tag{4}$$

where $Z_l$ directly denotes the activation of attention in the spatial grid $(x, y)$ for thoracic disease classification. The crop mask $S_l(x, y)$ is obtained from $Z_l$ by setting element $Z_l(x, y)$, which is larger than threshold parameter $\theta_c \in [0, 1]$ to 1, and setting others to 0. Specifically,

$$S_l(x, y) = \begin{cases} 
1, & \text{if } Z_l(x, y) > \theta_c \\
0, & \text{otherwise} 
\end{cases} \tag{5}$$

where $\theta_c$ is the parameter that adjusts the size of the finer lesion region. Larger $\theta_c$ will result in smaller lesion regions, and vice versa.

By using the crop mask $S_l$, we calculate a maximum connected area, which can cover all the discriminative points in $S_l$. The maximum connected area represents as the minimum and maximum coordinates on the horizontal and vertical axis $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$. Finally, the medical lesion discriminative region image crop from the previous-scale lung region image and resize to $224 \times 224$ resolution image as input of the lesionNet branch. Specifically, assume that the upper left corner of the original image is the origin of the pixel coordinate system, and its $x$-axis and $y$-axis are defined as left-to-right and top-to-bottom, respectively. We can use parameterization from the top left (represented as $tl$) and bottom right (represented as $br$) points of the attention region as following:

$$x_{\min}(tl) = t_x - t_l, \quad y_{\min}(tl) = t_y - t_l$$

$$x_{\max}(br) = t_x + t_l, \quad y_{\max}(br) = t_y + t_l \tag{6}$$

where $t_x, t_y$ represents the center coordinate of the square between the $x$ and $y$ axes, respectively, and $t_l$ represents half of the square’s side length. As illustrated in Figure 6, the bounding box obtained by weakly supervision method that can cover the entire selected positive finer lesion region, which means the subtle pathologies can be seen closer to diagnose better. Although the lesion area has been localized, it is sometimes difficult to extract effective features from high-localized parts. Therefore, each boundary of the bounding box enlarges the domain to a large size through adaptive zooming.

### F. TRAINING STRATEGY OF KGZNet

To better optimize discriminative localization and thoracic disease classification in a mutually reinforced way, our training strategy is as follows:
Step I: We use the same pre-trained ResNet-50 network initialization (global, lung, and lesion) convolution/classification layer from ImageNet.

Step II: The proposed Lung Region Generator (LRG-1) is used to acquire lung region images and feed them into the LungNet branch for the fine-tuning process. \( \tilde{p}_l(y|x) \) is normalized by Eq. 1. when we fine-tune the LungNet branch, the weights in the globalNet branch are fixed.

Step III: Once the lesion region images are obtained by Lesion Region Generator (LRG-2). We feed them into the LesionNet branch for fine-tuning learning. \( \tilde{p}_n(y|x) \) is also normalized by Eq. 1. Note that when we fine-tune the weight of the lesionNet branch, the weights of the globalNet branch and lungNet branch are fixed.

Step IV: To take full advantage of discriminative features of global images, lung region images, and lesion region images. Let \( \text{Pool}_{gb} \), \( \text{Pool}_{lg} \), and \( \text{Pool}_{ln} \) represent the Pooling-5 layer outputs of globalNet, lungNet and lesionNet branches, respectively. Then, we concatenate them for a fine-tuning fusion module, while fixing the weights of the previous three branches. The same hyperparameters and optimization methods by Eq. 1, 2 are used.

![Illustration of the distribution of the number of multiple pathology labels in Chest X-ray 14 dataset.](image)

FIGURE 7. Illustration of the distribution of the number of multiple pathology labels in Chest X-ray 14 dataset.

### IV. EXPERIMENTS AND RESULTS

#### A. DATASET AND IMPLEMENTATION DETAILS

1) DATASET

In our study, the proposed method was validated on the NIH Chest X-14\(^1\) dataset [16], which is a widely used touchstone for multi-label thoracic disease classification in CXR images. This is a large-scale publicly available dataset that contains 112,200 frontal-view CXR images from 30,805 unique patients. The size of images is 1024 × 1024 resolution with labeled up to 14 common thoracic diseases. These labels are accessed by analyzing the associated radiology reports. As shown in Figure 7, a wide range of are distributed in the Chest X-ray 14 dataset, but the number of thoracic disease are highly unbalanced. For example, 19,894 samples are labeled with “infiltration,” while only 227 samples are labeled with “Hernia.” Moreover, the dataset also contains 984 ground truth bounding boxes annotated by board-certified radiologists for 880 CXR images.

2) IMPLEMENTATION DETAILS

Our feature extractor sub-network is implemented with PyTorch. We initialized the DenseNet-121 (or ResNet-50 [37]) with the weights pre-trained on the ImageNet [34]. The original CXR images were resized to 224 × 224 low-resolution and augmented by random horizontal flipping to reduce the network computation cost. In our experiments, the entire dataset was randomly split to training (70%), validation (10%), and testing (20%). We used the binary cross-entropy loss (W-BCE), and the training of each stage was done Stochastic Gradient Descent (SGD) and with a momentum of 0.9. For the globalNet, lungNet, lesionNet and fusionNet branch, the mini-batch was set to 64, 64, 32, and 32, respectively. The each branch sub-network was trained for 80, 80, 50, and 50 epochs, respectively. The initial learning rate is \( \rho = 10^{-3} \) and divide by 20 each time whenever the validation loss reaches a plateau after an epoch. Moreover, we use some tricks to set the weight balance between the four loss functions, according to experimental experience, the weights are set to \( \psi_1 = 1, \psi_{2,3,4} = 0.25 \), respectively.

#### B. EVALUATION METRICS

In our work, the segmentation model of U-Net attempts to segment the lung field from the global CXR images. The intersection over union (IoU) is used to evaluate lung segmentation performance, which is defined as:

\[
\text{IoU}(G_{\text{tru}}, P_{\text{pre}}) = \frac{G_{\text{tru}} \cap P_{\text{pre}}}{G_{\text{tru}} \cup P_{\text{pre}}} \tag{7}
\]

where \( G_{\text{tru}} \) is the ground truth, and \( P_{\text{pre}} \) is the predicted results. The dice similarity coefficient (DSC) is used to measure the overlap between the predicted results and ground truth, which can be denoted as:

\[
\text{DSC}(G_{\text{tru}}, P_{\text{pre}}) = \frac{2 \times |G_{\text{tru}} \cap P_{\text{pre}}|}{|G_{\text{tru}}| + |P_{\text{pre}}|} \tag{8}
\]

In addition, we also use the sensitivity and the positive predictive value as auxiliary evaluation metrics. To evaluate the performance of our KGZNet method, the False Positive Rate (FPR) is the proportion of predicted errors in samples with the actual label “0”. The True Positive Rate (TPR) is the proportion of samples that are predicted to be “1” that are correctly predicted. The formula is as follows:

\[
\begin{align*}
\text{TPR} &= \frac{TP}{TP + FN}, \\
\text{FPR} &= \frac{FP}{FP + TN} \tag{9}
\end{align*}
\]

The task for the classification of thoracic disease can be treated as a multi-label classification (14 classes). We utilize the average Area under the Receiver Operating Characteristic (AUC) for evaluation.

\(\text{}\)

\(^1\)https://nihcc.app.box.com/v/ChestXray-NIHCC
curve (AUC) of 14 classes as our evaluation metric, which is the AUC of sensitivities and 1-specificities through changing the disease classification threshold from 0 to 1.0. Moreover, it notes that the more significant score of AUC, the performance of the diagnostic classifier is better.

**TABLE 2.** Segmentation accuracy over the test dataset. The values are indicated in the “mean ± standard deviation” format.

| Method       | DSC (%) | Sensitivity(%) | Positive predictive value(%) |
|--------------|---------|----------------|-------------------------------|
| Modified U-Net | 89.83±1.26 | 92.54±3.46 | 90.23±2.64 |

**C. LUNG SEGMENTATION PERFORMANCE OF U-Net**

We visualized the segmentation results in Figure 8. As can be seen, the U-Net can effectively localize in lung regions, and the difference between the lung segmentation results and annotated pairs is small. It shows U-Net is a better network for lung segmentation tasks in the proposed KGZNet method. The lung region images localized and cropped by the Lung region Generator largely reduced the noise regions in original CXR images such as the irregular regions and irrelevant objects (e.g., tags, medical devices). In Figure 9, we showed that the batch loss on the training set and validation set, respectively. The lung segmentation model, with an average IoU of 90.32%. In Table 2, the lung segmentation model yields a mean dice similarity coefficient (DSC) of 89.83%, a sensitivity of 92.56%, and a positive predictive value of 90.23%. Obviously, the accuracy of lung segmentation is positively correlated with the performance of the sub-network classification.

**D. COMPARISON WITH STATE-OF-THE-ART METHODS**

In Table 3, we compared the results acquired by our proposed method with the state-of-the-art results [16], [18]–[20], [32], [36] on the public NIH Chest X-ray 14 dataset. The score of mean AUC across the 14 thoracic disease classes achieved 0.869 (KGZNet-1) and 0.878 (KGZNet-2). Specifically, KGZNet outperforms these previous baselines, especially for Wang et al. [16] and Li et al. [19] with improvements of 14% and 12.3%. In addition, our method is still 0.7% better than the current best method [36]. Moreover, we compared the proposed algorithm with our baseline model (globalNet branch). The proposed method achieved a better result of AUC (0.841 vs 0.869) using ResNet-50 as the backbone. The reason for the excellent performance is that our approach makes full use of the visual features from different discriminative areas, thus gaining a richer understanding of the sample. Compared with fusion lung and lesion regions, the AUC scores of some diseases with KGZNet are obviously improved, e.g. Effusion (0.911 vs 0.879), pneumothorax (0.928 vs 0.872) and Nodule (0.818 vs 0.775). It is worth explaining that the lungNet (or lesionNet) branch yields lower accuracy than the globalNet branch, because we used the global CXR images as our validation dataset to train our sub-model. This way will result in a lack of structural information about the disease during testing.
FIGURE 9. Batch loss on the training set and validation set, and the average IoU on the validation set, respectively.

TABLE 3. Comparison with previous state-of-the-art baselines on Chest X-ray 14 dataset.

| Methods            | CNN | Atelis | Cardiety | Effusion | Infilt | Mass | Nodule | Pneum | Pneumothorax | Consol | Edema | Emphysema | Fibros | Pleural Thickening | PT | Hern | Mean |
|--------------------|-----|--------|----------|----------|--------|------|--------|-------|--------------|--------|--------|------------|--------|-------------------|----|------|------|
| Wang et al. [16]   | R-50| 0.716  | 0.807    | 0.784    | 0.809  | 0.706| 0.671  | 0.633 | 0.806   | 0.708   | 0.835  | 0.815     | 0.769  | 0.708             | 0.767 | 0.738 |
| Li et al. [19]     | R-50| 0.727  | 0.836    | 0.789    | 0.672  | 0.776| 0.696  | 0.649 | 0.808   | 0.720   | 0.806  | 0.888     | 0.771  | 0.737             | 0.693  | 0.755 |
| Yao et al. [18]    | $-$ | 0.772  | 0.904    | 0.859    | 0.695  | 0.792| 0.717  | 0.715 | 0.841   | 0.788   | 0.882  | 0.829     | 0.767  | 0.765             | 0.914  | 0.803 |
| Guan et al. [17]   | D-121| 0.781 | 0.883    | 0.831    | 0.679  | 0.830| 0.764  | 0.725 | 0.866   | 0.758   | 0.853  | 0.911     | 0.826  | 0.780             | 0.918  | 0.816 |
| Rajpurohit et al. [20] | D-121| 0.809 | 0.925    | 0.864    | 0.735  | 0.868| 0.780  | 0.768 | 0.889   | 0.790   | 0.888  | 0.937     | 0.805  | 0.806             | 0.916  | 0.841 |
| Kumar et al. [31]  | D-161| 0.762 | 0.913    | 0.864    | 0.692  | 0.750| 0.666  | 0.715 | 0.859   | 0.784   | 0.888  | 0.898     | 0.756  | 0.774             | 0.802  | 0.799 |
| Guan et al. [36]   | D-121| 0.853 | 0.999    | 0.903    | 0.764  | 0.902| 0.828  | 0.774 | 0.921   | 0.842   | 0.924  | 0.864     | 0.837  | 0.921             | 0.871  | 0.871 |
| GlobalNet branch (Scale 1) | D-121| 0.835 | 0.910    | 0.885    | 0.710  | 0.857| 0.784  | 0.763 | 0.840   | 0.815   | 0.849  | 0.924     | 0.837  | 0.786             | 0.912  | 0.841 |
| LungNet branch (Scale 2) | R-50| 0.798 | 0.913    | 0.878    | 0.697  | 0.835| 0.732  | 0.746 | 0.855   | 0.807   | 0.840  | 0.888     | 0.815  | 0.754             | 0.868  | 0.820 |
| LungNet branch (Scale 3) | R-50| 0.789 | 0.933    | 0.868    | 0.691  | 0.774| 0.690  | 0.734 | 0.833   | 0.795   | 0.879  | 0.872     | 0.747  | 0.735             | 0.866  | 0.804 |
| KGZNet-1 (Scale 1x2x3) | R-50| 0.841 | 0.938    | 0.901    | 0.760  | 0.894| 0.830  | 0.778 | 0.916   | 0.843   | 0.918  | 0.945     | 0.853  | 0.846             | 0.913  | 0.869 |
| GlobalNet branch (Scale 1) | D-121| 0.815 | 0.901    | 0.879    | 0.702  | 0.833| 0.775  | 0.760 | 0.872   | 0.806   | 0.890  | 0.917     | 0.837  | 0.781             | 0.951  | 0.840 |
| LungNet branch (Scale 2) | D-121| 0.817 | 0.899    | 0.878    | 0.677  | 0.834| 0.760  | 0.774 | 0.863   | 0.811   | 0.905  | 0.922     | 0.832  | 0.769             | 0.940  | 0.831 |
| LungNet branch (Scale 3) | D-121| 0.798 | 0.898    | 0.866    | 0.667  | 0.790| 0.688  | 0.741 | 0.835   | 0.800   | 0.899  | 0.846     | 0.786  | 0.735             | 0.833  | 0.801 |
| KGZNet-2 (Scale 1x2x3) | R-50| 0.853 | 0.948    | 0.911    | 0.758  | 0.913| 0.818  | 0.804 | 0.928   | 0.849   | 0.931  | 0.941     | 0.872  | 0.837             | 0.966  | 0.878 |

We calculate the AUC score for each class and the average AUC scores for 14 thoracic diseases. The fine-tuned ResNet-50 (R-50) and DenseNet-121 (D-121) as the backbone in our approach, respectively. $/$ represents the use of a combination of DenseNet and ResNet. The 14 thoracic diseases in Chest X-ray 14 are Atelectasis (Atelis), Cardiomegaly (Cardiety), Effusion (Effusion), Infiltration (Infilt), Mass, Nodule, Pneumonia (Pneum), Pneumothorax (Pneumothorax), Consolidation (Consol), Edema, Emphysema (Emphy), Fibrosis (Fibros), Pleural Thickening (PT), and Hernia (Hern). For each column, the best result is highlighted in black bold.

However, the training strategies are more conducive to learning discriminate region’s knowledge about the objective and are necessary for clinical practice. Most importantly, we find all the discriminative region detection are mutually correlated and reinforce each other. The ROC curves of our algorithm KGZNet with two fine-tuned backbones, as shown in Figure 10. As can be seen, the performance of KGZNet-1 (0.878) is slightly better than KGZNet-2 (0.869).
E. QUALITATIVE RESULTS
We visualized some thoracic disease classification results, as shown in Figure 11, and presented the top-8 probability scores for each pathology. The ground truth of disease labels is a highlight in red. For the multi-label thoracic disease classification task in CXR images, it can be seen that the probability scores of ground truth diseases are relatively higher. In addition, the score gap between the ground truth thoracic diseases and other thoracic diseases are relatively large, e.g., the predicted score of ground truth pathology “Effusion” (row 1, column 2) is 0.832359, which far exceeds other unrelated pathologies (“Mass” is 0.035287, “PT” is 0.025008 and “Nodule” is 0.023229). Compared with the pre-trained model of VGG-16, the mean AUC score based on ResNet-50 (or DenseNet-121) can be approximately improved 7%. Therefore, we use ResNet-50 and DenseNet-121 as our baseline backbone to extract the discriminative region features at each scaled image.

V. ABLATION STUDIES
A. DIFFERENT OF PRE-TRAINED MODELS ANALYSIS
To choose the best network model as the backbone of our algorithm. In Table 4, we initialize the DCNN framework by comparing five different pre-trained models (AlexNet, GoogLeNet, VGG-16, ResNet-50, and DenseNet-121), and obtain eight pathologies multi-label thoracic classification AUC values. It can be seen, the quantitative performance results vary greatly, and pre-trained models based on ResNet-50 and DenseNet-121 can achieve better results. For example, compare with the pre-trained model of VGG-16, the mean AUC score based on ResNet-50 (or DenseNet-121) can be approximately improved 7%. Therefore, we use ResNet-50 and DenseNet-121 as our baseline backbone to extract the discriminative region features at each scaled image.
TABLE 4. Verify initial classification results of different pre-trained CNN models.

| Backbone          | Atelectasis | Cardiomegaly | Effusion | Infiltration | Mass | Nodule | Pneumonia | Pneumothorax | Mean   |
|-------------------|-------------|--------------|----------|--------------|------|--------|-----------|--------------|--------|
| AlexNet [4]       | 0.6458      | 0.6925       | 0.6642   | 0.6041       | 0.5644 | 0.6487 | 0.5493    | 0.7425       | 0.6389 |
| GoogLeNet [40]    | 0.6307      | 0.7056       | 0.6876   | 0.6088       | 0.5363 | 0.5579 | 0.5990    | 0.7824       | 0.6385 |
| VGGNet-16 [41]    | 0.6281      | 0.7084       | 0.6502   | 0.5896       | 0.5103 | 0.6556 | 0.5100    | 0.7516       | 0.6255 |
| ResNet-50 [37]    | 0.7069      | 0.8141       | 0.7362   | 0.6128       | 0.5699 | 0.7164 | 0.6333    | 0.7891       | 0.6962 |
| DenseNet-121 [33] | 0.7032      | 0.8253       | 0.7421   | 0.6024       | 0.5321 | 0.7238 | 0.6482    | 0.7763       | 0.6942 |

TABLE 5. Compare the accuracy of disease localization.

| T(IoU) | Model | Atelectasis | Cardiomegaly | Effusion | Infiltration | Mass | Nodule | Pneumonia | Pneumothorax | Mean   |
|--------|-------|-------------|--------------|----------|--------------|------|--------|-----------|--------------|--------|
| 0.1    | [42]  | 0.6977      | 0.8333       | 0.6234   | 0.6350       | 0.4324 | 0.1234 | 0.6973    | 0.4687       | 0.5639 |
|        | Ours  | 0.7124      | 0.7136       | 0.6835   | 0.6850       | 0.5843 | 0.6264 | 0.5850    | 0.6515       | 0.6527 |
| 0.25   | [42]  | 0.4534      | 0.8277       | 0.4840   | 0.5734       | 0.3866 | 0.0023 | 0.5342    | 0.3512       | 0.4516 |
|        | Ours  | 0.4625      | 0.9024       | 0.5438   | 0.5935       | 0.3924 | 0.3482 | 0.3283    | 0.3563       | 0.4909 |
| 0.5    | [42]  | 0.2198      | 0.5231       | 0.2473   | 0.2412       | 0.1854 | 0.0019 | 0.3693    | 0.0716       | 0.2325 |
|        | Ours  | 0.3249      | 0.4135       | 0.1658   | 0.1943       | 0.1938 | 0.1497 | 0.1257    | 0.3268       | 0.2368 |

FIGURE 13. Average AUCs for different threshold values of parameter $\theta_c$ on the test dataset.

is close to the lung branch. Comparing with other range threshold values, when the value is in the range $\theta_c \in [0.5, 0.7]$, the relative AUCs score gains significantly. This improvement mainly comes from accurate discriminative attention localization. KGZNet is pretty stable, with the charge of the threshold value of $\theta_c$. It is worth mentioning that our method’s effectiveness is from the fusing the local cues (lung, lesion regions) with the global information. Moreover, the performance of KGZNet has an improvement of at least 1.8% at $\theta_c = 0.3$ by leveraging the power of the discriminative feature ensemble.

Finally, we capture the discriminative regions visual features for each disease by Grad-CAMs [27] to show the visual explanation of how KGZNet recognizes CXR images. As shown in Figure 14, different thoracic diseases have different shapes, textures, and sizes. The nodule and mass are often small in size. The CXR images finding of pneumonia are airspace opacity, interstitial opacities, and focus on the lung field. For Infiltration, the ground truth bounding box includes all the lung region, while the heatmaps focus on the lesion region on a large scale, which can also recognize Infiltration effectively. Obviously, thoracic diseases are generally limited within the lung area of CXR images, and the proposed method acquired satisfied detection results. In Table 5, we compare our localization disease region with the ground truth bounding-box (GT-Bbox). To verify the accuracy of our localization disease region and GT-Bbox, we use the varying $T(IoU) \in \{0.1, 0.25, 0.5\}$ for measurement, where $T(*)$ is the threshold. Comparing the method model of [42], our model generally performs better without using any annotated images. For example, when evaluated by $T(IoU) = 0.25$, our “Nodule” accuracy is 34.52%, while the reference model obtains only 0.2%; when evaluated by $T(IoU) = 0.5$, our “Pneumothorax” achieves 32.68%, while the reference model only is 7.16%. Please note that the type of disease “pneumonia,” the reference model is better than our result. The main reason is that our model treats discrete regions as prediction regions, so the area obtained using GT-Bbox is much larger than ground-truth.

TABLE 6. The ablation study of different training strategies.

| Method                | Global | Lung | Lesion | AUC   |
|-----------------------|--------|------|--------|-------|
| KGZNet-1(scale 3)     | √      |      |        | 0.804 |
| KGZNet-1(scale 2)     | √      |      |        | 0.820 |
| KGZNet-1(scale 1)     |         |      |        | 0.841 |
| KGZNet-1(scale 1+2)   | √      |      |        | 0.856 |
| KGZNet-1(scale 2+3)   | √      |      |        | 0.832 |
| KGZNet-1(scale 1+2+3) | √      |      |        | 0.869 |
| KGZNet-2(scale 3)     | √      |      |        | 0.801 |
| KGZNet-2(scale 2)     | √      |      |        | 0.831 |
| KGZNet-2(scale 1)     | √      |      |        | 0.840 |
| KGZNet-2(scale 1+2)   | √      |      |        | 0.863 |
| KGZNet-2(scale 2+3)   | √      |      |        | 0.842 |
| KGZNet-2(scale 1+2+3) | √      |      |        | 0.875 |

VOLUME 8, 2020 159801
Furthermore, although the lesion branch (scale 3) achieves 80.4%, the performance can increase to 83.5% after combining the feature from the lung region (scale 2). This performance shows that the discriminative areas are mutually reinforced each other. Note that KGZNet-2 (scale 2) can also achieve better performance with 83.1%. Besides, (scale 2) slightly drop than KGZNet-2 (scale 1), due to the lack of structural disease information existed in global CXR images. Note that the excellent result benefits from the complementary advantage from multiple discriminative regions feature learning. By combing features at three scales via concatenating the global average pooling, we achieve the best mean AUC = 87.8%. In addition, we extend KGZNet to more scales visual learning, but the performance saturates as discriminative visual features information has encoded into the previous scales. In particular, the improvement benefits from the most discriminative regions visual features learning instead of adding the number of network parameters.

VI. DISCUSSIONS

A. THE LIMITATION OF COMPUTATION OF KGZNet

To verify our proposed model’s average computation time. As shown in Table 7, we show the comparison of the computation time between KGZNet and previous state-of-the-art baselines. Our experiment implement works conduct on NVIDIA Tesla V100 GPU, and we recorded the test time cost of 22,424 test images from the Chest X-ray 14 test dataset and averaged it at the end of the test run. Compared with other state-of-the-art models, our training time and test time are higher than them, such as the method of Kumar et al. [32] and KGZNet-2 are (64min vs. 98min). The main reason is that the complexity of our proposed model and training strategies, which increase the cost of training time. This drawback is indeed the limitation of our model. However, the test time cost of KGZNet-1 (0.0506) and KGZNet-2 (0.0552) increases slightly compared with other baseline models, which can be ignored in clinical diagnosis. Moreover, the proposed of our KGZNet model achieves an average AUCs score that is about 3% higher than the state-of-the-art CheXNet [20]. Therefore, this limitation can not obscure the superior performance of our algorithm.

![Figure 14](image-url)
TABLE 8. Abnormality classification on the KRPD dataset.

| Methods    | Youden | Sensitivity | Specificity | Acc   |
|------------|--------|-------------|-------------|-------|
| ResNet-50  | 0.6606 | 0.8937      | 0.7669      | 0.9021|
| ResNet-101 | 0.6618 | 0.9036      | 0.7852      | 0.9037|
| ResNet-152 | 0.6760 | 0.8905      | 0.7585      | 0.9102|
| DenseNet-121 | 0.6701 | 0.9028      | 0.7673      | 0.9038|
| KGZNet-1(ours) | 0.7164 | 0.9128      | 0.8036      | 0.9328|
| KGZNet-2(ours) | 0.7432 | 0.9274      | 0.8158      | 0.9410|

B. APPLICATION IN ANOTHER MEDICAL CHEST X-RAY IMAGES

To verify the robustness of our method, we also use another medical Chest X-ray images from the Kaggles RSNA Pneumonia Detection Challenge 2018 (KRPDc) [43]. The dataset with two categories (Pneumonia/Normal), which consists of approximately 26,000 X-ray images for training and 3,000 X-ray images for testing. In our experiment, we randomly select 1,000 x-rays from the training data set as the validation set to prevent the model from overfitting. The size of the original images are downsampled from 1024 × 1024 to 224 × 224 pixels. We also augment the dataset by rotating, panning, zooming, and flipping horizontally. The classification performance is evaluated by youden, sensitivity, specificity, and overall accuracy (Acc). As shown in Table 8, the proposed KGZNet-1 and KGZNet-2 perform better than the corresponding ResNet-50 and DenseNet-121 on the KRPDc dataset, and the mean Acc scores are improved 3.07% and 3.72% (0.9328 vs. 0.9021) & (0.9410 vs. 0.9038), respectively. Furthermore, it is obvious to see that the youden scores of KGZNets are far higher than others. It because of our model can jointly learn features from multiple views to improve the classification performance.

VII. CONCLUSION

In this study, we present a novel knowledge-guided deep zoom neural network (KGZNet) framework to demonstrate the effectiveness of thoracic disease recognition. Specifically, the proposed framework gradually learns the discriminative region guided by the prior medical knowledge and finer region-based feature representation on the multiple scale branches. Extensive experimental results prove the superiority of this method, which can learn the most medical discriminative visual features from coarse to fine and through comprehensive fusion learning. This strategy can minimize the loss of detailed information during the network downsampling process. The ablation studies also demonstrate that the visual features information of both lung and lesion regions can boost the performance of thoracic disease classification and reinforce each other. Furthermore, our diagnostic network can also provide heatmaps to help radiologists determine the location of the lesion, which can effectively reduce the workforce and improve diagnosis efficiency in practical applications. In future work, we will perform a thorough study using an attentional map instead of the current model architecture and lightweight our model.

REFERENCES

[1] A. Brady, R. Laoide, P. McCarthy, and R. McDermott, “Discrepancy and error in radiology: Concepts, causes and consequences,” The Ulster Med. J., vol. 81, no. 1, p. 3, 2012.
[2] S. Mayer, “International evaluation of an ai system for breast cancer screening,” Nature, vol. 577, no. 7788, pp. 89–94, 2020.
[3] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1285–1298, May 2016.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2012, pp. 1097–1105.
[5] J. Fu, H. Zheng, and T. Mei, “Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4438–4446.
[6] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, “Weakly supervised deep learning for thoracic disease classification and localization on chest X-rays,” in Proc. ACM Int. Conf. Bioinf., Comput. Biol., Health Informat., Aug. 2018, pp. 103–110.
[7] M. Anthimopoulos, S. Christoudelis, L. Ebner, A. Christe, and S. Mougiakakou, “Lung pattern classification for interstitial lung diseases using a deep convolutional neural network,” IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1207–1216, May 2016.
[8] X. Zhang, S. Wang, J. Liu, and C. Tao, “Computer-aided diagnosis of four common cutaneous diseases using deep learning algorithm,” in Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM), Nov. 2017, pp. 1304–1306.
[9] Q. Wang, D. Yang, Z. Li, X. Zhang, and C. Liu, “Deep regression via multi-channel multi-modal learning for pneumonia screening,” IEEE Access, vol. 8, pp. 78530–78541, 2020.
[10] E. Gibson, F. Giganti, Y. Hu, E. Bonnati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, “Automatic multi-organ segmentation on abdominal CT with dense V-Networks,” IEEE Trans. Med. Imag., vol. 37, no. 8, pp. 1822–1834, Aug. 2018.
[11] R. Wang, B. Chen, D. Meng, and L. Wang, “Weakly supervised lesion detection from fundus images,” IEEE Trans. Med. Imag., vol. 38, no. 6, pp. 1501–1512, Jun. 2019.
[12] C. Ye, W. Wang, S. Zhang, and K. Wang, “Multi-depth fusion network for whole-heart CT image segmentation,” IEEE Access, vol. 7, pp. 23421–23429, 2019.

FIGURE 15. The validation loss during training.

The validation loss during training is shown in Figure 15. It can be seen that the loss of our framework (KGZNet) decline faster than other networks, which means that the lung and lesion region feature information can accelerate the convergence speed.

2https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

FIGURE 15. The validation loss during training.
[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

[35] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, “TeNet: Text-image embedding network for common thorax disease classification and reporting in chest X-Rays,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 9049–9058.

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[37] Olaf Ronneberger, Philipp Fischer, and Tomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervent., 2015, pp. 234–241.

[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2921–2929.

[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.

[40] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[41] Yan Shen and Mingchen Gao, “Dynamic routing on deep neural network (DNN) architectures,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 8192–8199.

[42] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[43] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[44] R. Chen, Y. Yang, and Y. Yu, “Weakly supervised medical diagnosis and localization from deep neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 38–45. Mar. 2020.

[45] P.-P. Ypsilantis and G. Montana, “Learning what to look in chest X-rays with a recurrent visual attention model,” 2017, arXiv:1711.05225. [Online]. Available: http://arxiv.org/abs/1711.05225

[46] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[47] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[48] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[49] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[50] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[51] R. Chen, Y. Yang, and Y. Yu, “Weakly supervised medical diagnosis and localization from deep neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 38–45. Mar. 2020.

[52] P.-P. Ypsilantis and G. Montana, “Learning what to look in chest X-rays with a recurrent visual attention model,” 2017, arXiv:1711.05225. [Online]. Available: http://arxiv.org/abs/1711.05225

[53] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[54] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[55] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[56] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[57] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[58] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[59] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[60] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.

[61] Q. Guan and Y. Huang, “Thorax disease identification and localization with limited supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8209–8299.
SHENG HUANG (Member, IEEE) received the B.Eng. degree in software engineering and the Ph.D. degree in computer science from Chongqing University, Chongqing, in 2010 and 2015, respectively. From 2012 to 2014, he was a Visiting Student with the Department of Computer Science, Rutgers University, Piscataway, NJ, USA. He has authored several scientific articles in venues such as CVPR, BMVC, ICME, ISBI, and ICIP. His research interests include pattern recognition, computer vision, machine learning, and biometrics.

FEIYU CHEN received the B.Sc. degree from the School of Mathematics, Liaoning Normal University, China, in 2007, the M.S. degree from the School of Mathematics Sciences, Dalian University of Technology, China, in 2009, and the Ph.D. degree from the Department of Applied Mathematics, Delaware State University, USA, in 2013. He is currently a Lecturer with the School of Big Data and Software Engineering, Chongqing University, China. His current research interests include data clustering, machine learning, and optimization.

XIANGBO ZHANG received the B.Eng. degree in software engineering from Henan University, China, in 2019. He is currently pursuing the M.S. degree with the School of Big Data and Software Engineering, Chongqing University, China. His current research interests include the fields of medical image analysis, computer vision, and machine learning.

LUWEN HUANGFU received the B.S. degree in software engineering from Chongqing University, China, the M.S. degree in computer science from the Chinese Academy of Sciences, and the Ph.D. degree in management information systems from the University of Arizona. She is an Assistant Professor of management information systems with San Diego State University, USA. Her research interests include business analytics, text mining, data mining, artificial intelligence, and healthcare management.

* * *