Contralaterally Enhanced Networks for Thoracic Disease Detection

Gangming Zhao, Chaowei Fang, Guanbin Li, *Member, IEEE*,
Licheng Jiao, *Fellow, IEEE*, and Yizhou Yu, *Fellow, IEEE*

Abstract—Identifying and locating diseases in chest X-rays are very challenging, due to the low visual contrast between normal and abnormal regions, and distortions caused by other overlapping tissues. An interesting phenomenon is that there exist many similar structures in the left and right parts of the chest, such as ribs, lung fields and bronchial tubes. This kind of similarities can be used to identify diseases in chest X-rays, according to the experience of broad-certificated radiologists. Aimed at improving the performance of existing detection methods, we propose a deep end-to-end module to exploit the contralateral context information for enhancing feature representations of disease proposals. First of all, under the guidance of the spine line, the spatial transformer network is employed to extract local contralateral patches, which can provide valuable context information for disease proposals. Then, we build up a specific module, based on both additive and subtractive operations, to fuse the features of the disease proposal and the contralateral patch. Our method can be integrated into both fully and weakly supervised disease detection frameworks. It achieves 33.17 AP50 on a carefully annotated private chest X-ray dataset which contains 31,000 images. Experiments on the NIH chest X-ray dataset indicate that our method achieves state-of-the-art performance in weakly-supervised disease localization.

Index Terms—Chest X-ray, disease detection, contralateral context, deep learning.

Gangming Zhao is with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China, and also with the Department of Computer Science, The University of Hong Kong, Hong Kong (e-mail: gmzhao@connect.hku.hk).

Chaowei Fang and Licheng Jiao are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: cwfang@xidian.edu.cn; lchjiao@mail.xidian.edu.cn).

Guanbin Li is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: liguanbin@mail.sysu.edu.cn).

Yizhou Yu is with the Department of Computer Science, The University of Hong Kong, Hong Kong, and also with the Deepwise AI Laboratory, Beijing 100080, China (e-mail: yizhouy@acm.org).

Digital Object Identifier 10.1109/TMI.2021.3077913

I. INTRODUCTION

CHEST X-ray (CXR) is one of the most widely-used examination tools for the diagnosis of thoracic diseases such as lung nodules and pneumonia. Thanks to the development of deep learning technologies, stupendous progress has been achieved in automatic disease classification [1], [2] and localization [3]–[5] for chest X-rays. Considering there exist similar structures in the left and right parts of the chest, we focus on exploring the contralateral context information for both fully and weakly supervised disease detection.

The main challenges of disease identification in chest X-ray images include low visual contrast between lesion regions and other components, and distortions induced by other overlapping tissues. Sometimes it is difficult for medical specialists to recognize obscure diseases [6], [7]. Designing automatic artificial intelligence systems is beneficial for guaranteeing the diagnosis efficiency and accuracy. Previous deep learning methods for chest X-ray diagnosis mainly concentrated on disease classification [2], [7]–[9]. Recently, several literatures researched on detecting disease regions under weak/limited supervision. They can be grouped into two main categories: the first category of methods [1], [3] resort to convolutional neural networks (CNN) trained on the classification task and output disease localization results through calculating category activation maps [10]; the second kind of methods [4], [5] use the multiple instance learning to directly yield categoric probability maps which can be easily transformed into lesion detections. However, the performance of these methods is still far from practical clinical usage.

A mainstream pipeline of object detection is to screen out potential proposals followed by identifying the class of proposals [11], [12]. Through stacking piles of convolutional layers, CNN models are very advantageous at extracting surrounding context information. However, distant relationships are still hard to be exploited with convolutions which usually have small kernels. Particularly for chest X-rays, the left and right parts of the chest share lots of similar structures, such as ribs, lung fields and bronchial tubes. Although the two halves of the chest are not symmetrical (e.g., the left and right lung is composed of 2 and 3 lobes respectively, and the heart resides only in the left side of the chest), we are wondering whether the similarity information can exert a positive influence on disease detection. As shown in Fig. 1,
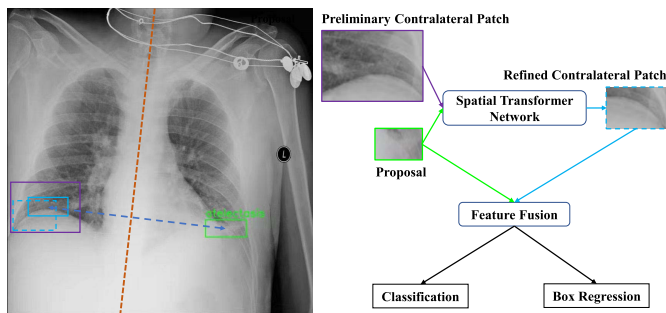
Fig. 1. For each disease proposal, we seek its initial contralateral patch under the guidance of the spine line. Then its pose is automatically adjusted to acquire a more appropriate patch to facilitate the final classification and localization of the disease proposal.

we devise a novel CNN-based module, named *Contralaterally Enhanced Networks*, to take advantage of such similarity information. First of all, we employ existing methods [5], [11], [13], [14] to acquire a number of disease proposals from the input image. For the sake of extracting the similarity information from the contralateral part of the chest, the spine line is used as the symmetrical axis to obtain a reference patch for every disease proposal. Then, with the help of the spatial transformer network [15], we sample an appropriate patch from a relatively large region enclosing the contralateral reference patch, to enhance the feature representation of the disease proposal. Finally, a fusion module is devised to aggregate the features of the disease proposal and its contralateral reference patch. Our proposed method can be plugged into both fully and weakly supervised disease detection frameworks. The main contributions of this paper can be summarized as follows.

- We build up a novel deep module, *Contralaterally Enhanced Networks*, for facilitating the disease detection in chest X-rays. We are the first to explicitly exploit the contralateral context information between the left and right parts of chest, to enhance the feature representations of disease proposals.
- An effective method is proposed to seek contralateral reference patches for disease proposals. A reference patch is extracted for each disease proposal under the guidance of the spine line and is further refined with the spatial transformer network. And A novel feature fusion module is devised to enhance the feature representation of a disease proposal with its contralateral reference patch.
- Our proposed method can improve existing object detection baselines [11], [13], [14], [16] with large margins on a chest X-ray dataset for fully supervised disease detection. It also achieves state-of-the-art performance on the NIH chest X-ray dataset [1] under the weakly supervised setting.

II. RELATED WORK

A. Object Detection

Object detection is a widely-studied topic in both natural and medical images. It aims at localizing object instances of interest such as faces, pedestrians and disease lesions. The most famous kind of deep learning approaches for object detection is the R-CNN [17] family. The primitive R-CNN extracts proposals through selective search [18], and then predicts object bounding boxes and categories from convolution features of these proposals. Fast R-CNN [19] adopts a shared backbone network to extract proposal features via RoI pooling. Faster R-CNN [11] automatically produces object proposals from top-level features with the help of pre-defined anchors. The above methods accomplish the detection procedure through two stages, including object proposal extraction, object recognition and localization. In [16], the feature pyramid network is exploited to further improve the detection performance of Fast R-CNN and Faster R-CNN with the help of multi-scale feature maps. The other pipeline for object detection implements object localization and identification in single stage through simultaneous bounding box regression and object classification, such as YOLO [20] and SSD [21]. The RetinaNet [13] is also built upon the feature pyramid network, and uses dense box predictions during the training stage. The focal loss is proposed to cope with the class imbalance problem. cornetNet [22] presents an anchor-free pipeline through detecting corners of bounding boxes. Despite of corners, the center point is also explored to guarantee the correctness of the obtained object boxes in CenterNet [14]. The detection task has also attracted a large amount of research interest in medical images, such as lesion detection in CT scans [12] and cell detection in malaria images [23]. This paper targets at detecting diseases in chest X-ray images. Practically, we propose a *Contralaterally Enhanced Networks* to exploit contralateral context information to enhance feature representations of disease proposals.

B. Disease Detection in Chest X-ray Images

Accurately recognizing and localizing diseases in chest X-Ray images is very challenging because of low textural contrast, large anatomic variation across patients, and organ overlapping. Previous works in this field mainly focus on disease classification [1], [2], [6], [24], [25]. Recently, the authors in [26] propose to transfer deep models pretrained on the ImageNet dataset [27] for recognizing pneumonia in chest X-ray images. In [28], the artificial ecosystem-based optimization algorithm is used to select the most relevant features for tuberculosis recognition. Based on the category activation map [10] which can be estimated with a disease recognition network, researchers attempt to localize disease in a weakly supervised manner [1], [3]. In [29], the triplet loss is used to facilitate the training of the disease classification model, and better performance is observed in class activation maps (CAM) estimated by the trained model. In [4], [5], multiple instance learning is employed to solve the disease localization problem. In [30], a novel weakly supervised disease detection model is devised on the basis of the DenseNet [31]. Two pooling layers including a class-wise pooling layer and a spatial pooling layer are used to transform 2-dimensional class attention maps into the final prediction scores. The performance of these methods is still far from practical usage in automatic diagnosis systems. In [32], a novel pipeline is proposed to identify and search
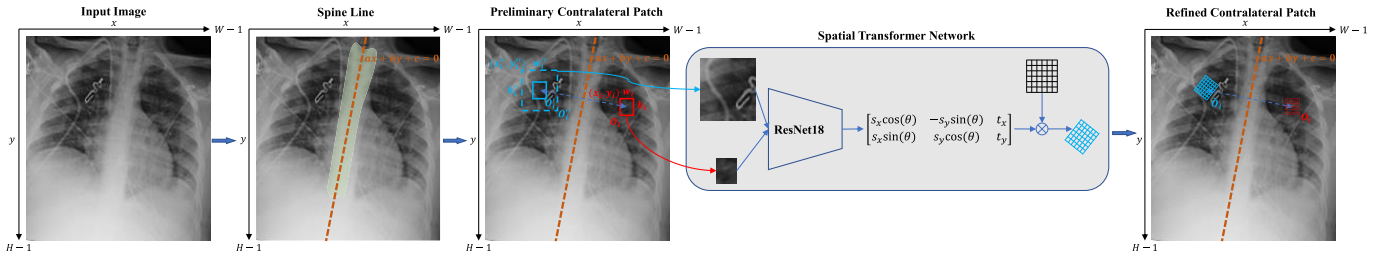
Fig. 2. Procedure of extracting the contralateral patch for each disease proposal. First the spine line is derived from the spine mask. For each disease proposal O_i, a preliminary contralateral patch O_i' is retrieved under the guidance of the spine line. Then a spatial transformer network is devised to acquire the final contralateral patch $\hat{O}_i$.

potential lung diseases with the help of heuristic algorithms, such as Moth-Flame and Ant Lion.

An interesting phenomenon is that there exist similar structures between the left and right halves of the chest, such as the left and right parts of lungs. Such contralateral context information can benefit the recognition of thoracic diseases according to the experience of broad-certificated radiologists. Reference [33] attempts to take advantage of the lung region symmetry when constructing hand-crafted features for thoracic disease or abnormality identification. In this paper, we devise a specific module to extract context information from the contralateral structures for strengthening the feature representations of disease proposals. The contralateral context information can effectively improve the performance of existing fully and weakly supervised disease detection methods.

III. CONTRALATERALLY ENHANCED NETWORKS

The target of this paper is to automatically locate various thoracic diseases in chest X-rays. There exists a certain degree of similarity between the left and right parts of the chest, from high-level structures of organs, such as lungs, bones and vessels, to low-level tissues. Based on this observation, we propose a feature enhancement module to exploit the contralateral context information for enhancing the feature representations of disease proposals. For every disease proposal extracted with an existing disease detection method, a reference patch from the contralateral location of the chest is acquired under the guidance of the spine line at first. Then a transformer network is devised to refine the pose of the reference patch, which is used to complement the representation of the disease proposal via an additive and subtractive feature fusion module. Our proposed module can be easily integrated into both fully and weakly supervised disease detection models. The technical details are illustrated as below.

A. Contralateral Patch Extraction

Given a CXR image I with size of $W \times H$, we can screen out n potential disease proposals $\{O_i \mid i = 1 \cdots n\}$ with the help of an existing disease/object detection method, such as the fully-supervised method [11] or the weakly-supervised method [5]. O_i is represented by a quad (x_i, y_i, w_i, h_i), indicating the horizontal and vertical coordinate of the top-left corner, width, and height, respectively. We denote the disease category of

O_i as l_i. Suppose the number of target disease categories is m. Hence, $l_i \in \{1, \cdots, m\}$. The pipeline of extracting the contralateral patch for each disease proposal is illustrated in Fig. 2.

1) Preliminary Contralateral Patch: Considering the spine is located at a relatively middle position of the chest, we exploit the spine line to fetch the preliminary contralateral patch for each disease proposal. The minimum circumscribed quadrilateral enclosing the spine mask can be obtained as in [34]. We regard the spine line bridged by the centers of two short edges as the symmetric axis, which can be expressed as $ax + by + c = 0$ (a, b and c are coefficients). For a disease proposal O_i, its preliminary contralateral patch $O_i'(= (x_i', y_i', w_i, h_i))$ is located through solving the following linear system,

$$\begin{cases} a\frac{x_i + x_i' + w_i - 1}{2} + b\frac{y_i + y_i' + h_i - 1}{2} + c = 0, \\ -b(x_i - x_i') + a(y_i - y_i') = 0. \end{cases} \quad (1)$$

The PSPNet proposed in [35] is chosen as our spine segmentation model. We use ResNet50 [36] as the backbone of PSPNet and modify the dimension of the final output into 1. The same settings as in [35] are adopted to optimize the network parameters. Quantitative and qualitative experimental results are reported in Section V-C.

2) Refined Contralateral Patch: To acquire a patch which is more suitable for enhancing the feature representation of the disease proposal, we further devise a spatial transformer network (abbr. STN) to refine the pose of O_i'. STN aims to find out a better contralateral patch for disease proposals through refining the location of the initial contralateral patch, it refines the localization of the initial contralateral patches and acts as a verifier of the disease localization for effective training. STN explicitly allows the spatial manipulation of data within the network. This differentiable module can be inserted into existing convolutional architectures, giving neural networks the ability to actively spatially transform feature maps, conditional on the feature map itself, without any extra training supervision or modification to the optimization process. In details, we set up STN on the basis of ResNet18 [36], through modifying the output dimension of the ultimate fully connected layer into 6. The ResNet18 model is pretrained on ImageNet dataset. And then it is trained with our framework in an end-to-end manner. We extend the borders of O_i' by Δ_x and Δ_y pixels along the horizontal and vertical axes respectively, resulting to a new patch $O_i'' = (x_i'', y_i'', w_i'', h_i'')$.

$x_i'' = x_i' - \Delta_x$, $y_i'' = y_i' - \Delta_y$, $w_i'' = w_i + 2\Delta_x$, and $h_i'' = h_i + 2\Delta_y$. Empirically, we set $\Delta_x = 0.25w_i$ and $\Delta_y = 0.25h_i$. Afterwards the original disease proposal O_i is padded to have the same size as O_i''. After being resized to a fixed size of $w^0 \times h^0$, the padded disease proposal and O_i'' are concatenated and fed into the STN, which gives rise to a tensor of 6 elements including two rescaling parameters (s_x and s_y), two transition parameters (t_x and t_y) and one rotation parameter (θ). These parameters can be used to locate the refined symmetrical patch $\hat{O}_i$. Here, w^0 and h^0 are both set as 64. For a point $(x^{\text{dst}}, y^{\text{dst}})$ in $\hat{O}_i$, we can obtain its corresponding point $(x^{\text{src}}, y^{\text{src}})$ in the original input image with a linear transformation operation, $[x^{\text{dst}}, y^{\text{dst}}]^T = \mathbf{T}_i [x^{\text{src}}, y^{\text{src}}, 1]^T$. $\mathbf{T}_i$ can be calculated as the following formulation,

$$\mathbf{T}_i = \begin{bmatrix} \frac{w_i''-1}{w_0-1} & 0 & x_i'' \\ 0 & \frac{h_i''-1}{h_0-1} & y_i'' \end{bmatrix} \begin{bmatrix} s_x \cos\theta & -s_y \sin\theta & t_x \\ s_x \sin\theta & s_y \cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix}. \quad (2)$$
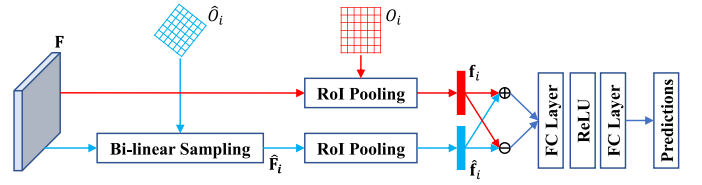
B. Feature Fusion Module

The architecture of the module for fusing features of a disease proposal and its contralateral patch is shown in Fig. 3. Given a feature map $\mathbf{F}$ extracted with a backbone network, the RoI pooling operation [19] is adopted to extract feature representation $\mathbf{f}_i$ for the disease proposal O_i. A feature map $\hat{\mathbf{F}}_i$ ($w \times h$) is extracted for the contralateral patch $\hat{O}_i$. For every point in $\hat{O}_i$, its coordinates are obtained according to transformation matrix in (2), and its feature vector is sampled from $\mathbf{F}$ via bi-linear interpolation. Again RoI pooling is used to aggregate $\hat{\mathbf{F}}_i$ into the feature representation $\hat{\mathbf{f}}_i$ for $\hat{O}_i$. The spatial sizes of $\mathbf{f}_i$ and $\hat{\mathbf{f}}_i$ are both set as 7×7. We directly flatten $\mathbf{f}_i$ and $\hat{\mathbf{f}}_i$ into 1-dimensional feature vectors. Every element in $\mathbf{f}_i$ or $\hat{\mathbf{f}}_i$ can be regarded as certain attribute factor of O_i or $\hat{O}_i$. The addition between $\mathbf{f}_i$ and $\hat{\mathbf{f}}_i$ can help highlight attributes which have large responses in both O_i and $\hat{O}_i$. This is beneficial to the identification of diseases striding across the left and right parts of the chest. On the other hand, the subtraction between $\mathbf{f}_i$ and $\hat{\mathbf{f}}_i$ can provide contrast information and suppress the responses of attributes which are irrelevant to disease recognition and localization. Considering the above issues, both addition and subtraction operations are used to merge $\mathbf{f}_i$ and $\hat{\mathbf{f}}_i$. Two fully connected layers are employed to transform the merged features into the final prediction. The output dimension of the first fully connected layer is 512, and that of the second fully connected layer depends on the length of the final prediction.

For fully supervised disease detection, the output of the contralaterally enhanced networks is composed of two groups including one m-dimensional category probability vector p_i, which identifies the disease type of O_i ($c_i = argmax_j p_i(j)$), and 4 m-dimensional bounding box offsets, t_i^x, t_i^y, t_i^w and t_i^h, which are used for adjusting the pose of O_i. The pose of the updated bounding box for O_i is as bellow,

$$x_i' = w_i * t_i^x(c_i) + x_i, \quad y_i' = h_i * t_i^y(c_i) + y_i, \quad (3)$$

$$w_i' = w_i * e^{t_i^w(c_i)}, \quad h_i' = h_i * e^{t_i^h(c_i)}. \quad (4)$$

Fig. 3. Feature fusion module. For the disease proposal $\mathbf{O}_i$, ROI pooling is directly used to extract the feature representation from the input feature map $\mathbf{F}$. For the contralateral patch $\hat{\mathbf{O}}_i$, a specific feature map is generated via bi-linear sampling, which is then used to produce a feature representation using ROI pooling as well. The feature representations of $\mathbf{O}_i$ and $\hat{\mathbf{O}}_i$ are fused, and then fed into two fully connected layers to produce the final predictions.

With the help of the above parameterizations, our method can determine the category, and resize the bounding box and translate the top-left corner of the disease proposal.

For weakly supervised disease detection, our method only rectifies the category prediction, and the final output is an m-dimensional category probability vector $\mathbf{p}_i$.

IV. DISEASE DETECTION FRAMEWORK

The pipeline for fully and weakly supervised disease detection frameworks is summarized in Fig. 4. Without specification, the residual network with 50 layers (ResNet50) is used as the backbone of disease detection models.

A. Fully Supervised Disease Detection

Our method can be incorporated into existing fully supervised disease detection frameworks, including both two-stage methods such as Faster R-CNN [11] and Faster R-CNN FPN [16], and one-stage methods such as RetinaNet [13] and CenterNet [14].

For two-stage methods, we directly replace the head for predicting the category and bounding box with our contralaterally enhanced networks. It is implemented through feeding the feature map produced by the backbone and class-agnostic disease proposals produced by the region proposal network into our proposed module. As shown in Fig.4, for detection models built with FPN, multi-scale feature maps are used to extract region proposals. Upon each scale of feature map, a set of proposals can be generated, and an independent contralaterally enhanced networks are employed to identify the category and rectify the pose of the proposals.

When incorporated with one-stage methods, we use the original detection models to produce disease proposals, ignoring the categorical information. Then, our module is adopted to predict the final disease class and bounding box for every disease proposal. The corresponding feature map which directly induces to the disease proposal is chosen to compute the feature representation for it and its contralateral patch.

For the primitive version of Faster R-CNN, the output of the 4-th convolution block (abbr. C4) is adopted as the input feature map of the contralaterally enhanced networks; for models using backbones with pyramid architectures (including Faster R-CNN FPN and RetinaNet), the exact feature map which gives rise to the disease proposal is chosen; for CenterNet,

TABLE I
CHEST X-RAY DATASET FOR FULLY SUPERVISED DISEASE DETECTION

| Disease	aorta widen	aorta calcification	papilla	multiple nodules	foreign matter	diffusive nodule
Images/Boxes	1,800/6,635	1,123/5,144	1,210/5,431	1,450/4,990	1,397/5,127	1,240/4,812
Disease	cardiomegaly	pneumothorax	hydropneumothorax	subcutaneous emphysema	cavity	fibrosis
Images/Boxes	1,130/3,125	1,120/6,147	1,370/6,000	1,700/7,101	1,100/6,113	1,700/8,100
Disease	widened mediastinal	nodule	rib abnormity	rib absence	shoulder abnormal	atelectasis
Images/Boxes	1,640/6,135	1,200/4,700	1,500/5,000	1,400/6,100	1,727/6,237	1,124/4,716
Disease	consolidation	emphysema	pulmonary tuberculosis	hilum increase	mass	effusion
Images/Boxes	1,175/5,764	1,034/4,100	1,204/6,257	1,027/6,131	810/3,102	910/4,105
Disease	pleural thickening	scoliosis	subphrenic air	diaphragm abnormity	calcification	rib fracture
Images/Boxes	2,130/8,170	814/3,204	805/4,017	827/3,015	723/2,113	1,240/7,154
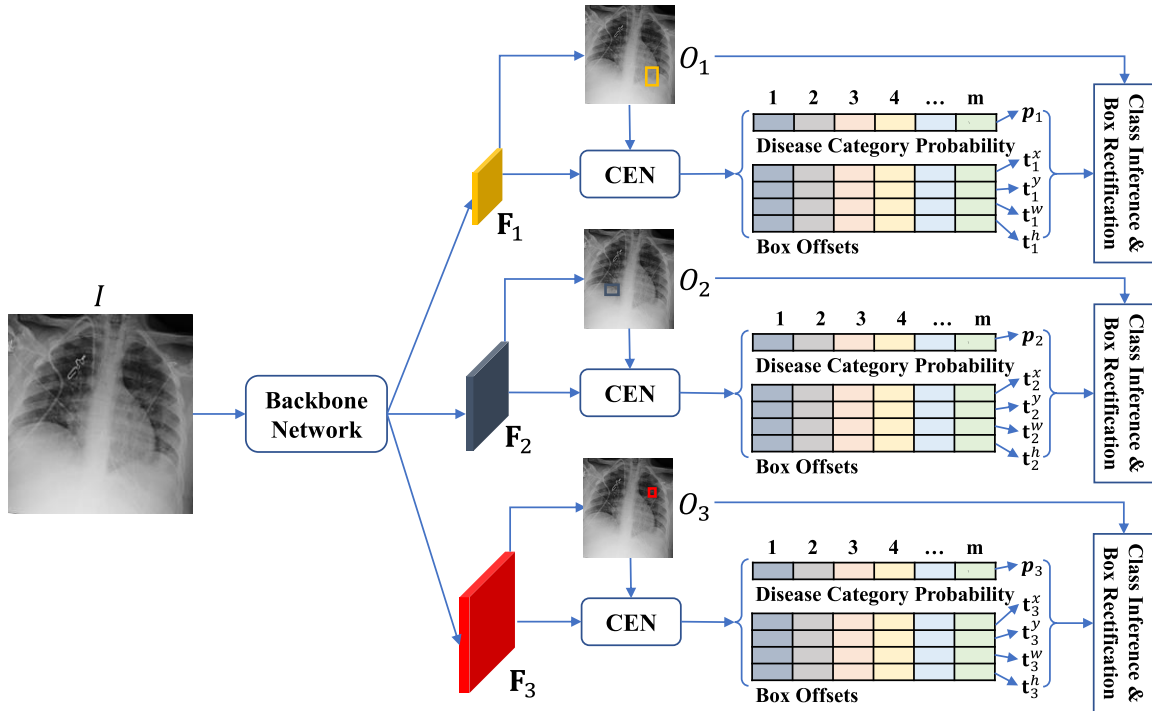
Fig. 4. The disease detection framework after incorporated with our proposed contralaterally enhanced networks (CEN). For detection models built upon feature pyramid networks, multi-scale feature maps are used for extract proposal regions, and each scale of feature map and generated proposals are fed into a specific CEN which produces the disease category probability map and box offsets for adjusting the poses of proposals. In single-scale detection models such as 'Faster R-CNN C4', only one scale of feature map is used for proposal generation. For weakly supervised methods, CEN only produces the category probability vectors without box offsets.

the output of the penultimate convolution layer is used as the input feature map.

We follow the loss function and the training algorithm in [11] to optimize our improved disease detection networks. The loss functions for training the object detection networks During the testing stage, 100 boxes with the highest confidences are selected as disease proposals, and non-maximum suppression with an IoU threshold of 0.7 is used to filter out severely overlapped boxes. The final predictions are post-processed with the non-maximum suppression again. The IoU threshold is set as 0.5. Boxes with confidences larger than 0.05 are considered as positive detections and the maximum number of boxes is set to 20.

B. Weakly Supervised Disease Detection

We can also integrate the contralaterally enhanced networks into the weakly supervised disease framework [5] which is trained with multiple instance learning (abbr. MIL). A probability map $\mathbf{P}$ with size of $\frac{H}{32} \times \frac{W}{32}$ is produced. The vector at position (x, y) in $\mathbf{P}$ indicate the probabilities of the patch $(32x, 32y, 32, 32)$ with respect to disease categories. Here we select top n patches as disease proposals and use the output of the penultimate convolutional layer as the feature map which is a tensor with size of $\frac{H}{32} \times \frac{W}{32}$. n is empirically set as 128. They are fed into the contralaterally enhanced networks, which produces new category probabilities for the selected disease proposals. During the training stage, the other MIL loss function is imposed on the new category probabilities. In the inference phase, the threshold value is set as 0.5 to select the disease proposals from the probability map, and determine the final detection results according to the output of the contralaterally enhanced networks.

V. EXPERIMENTS

A. Datasets and Evaluation Metrics

Fully Supervised Dataset We collect a private dataset to validate fully supervised disease detection methods. The
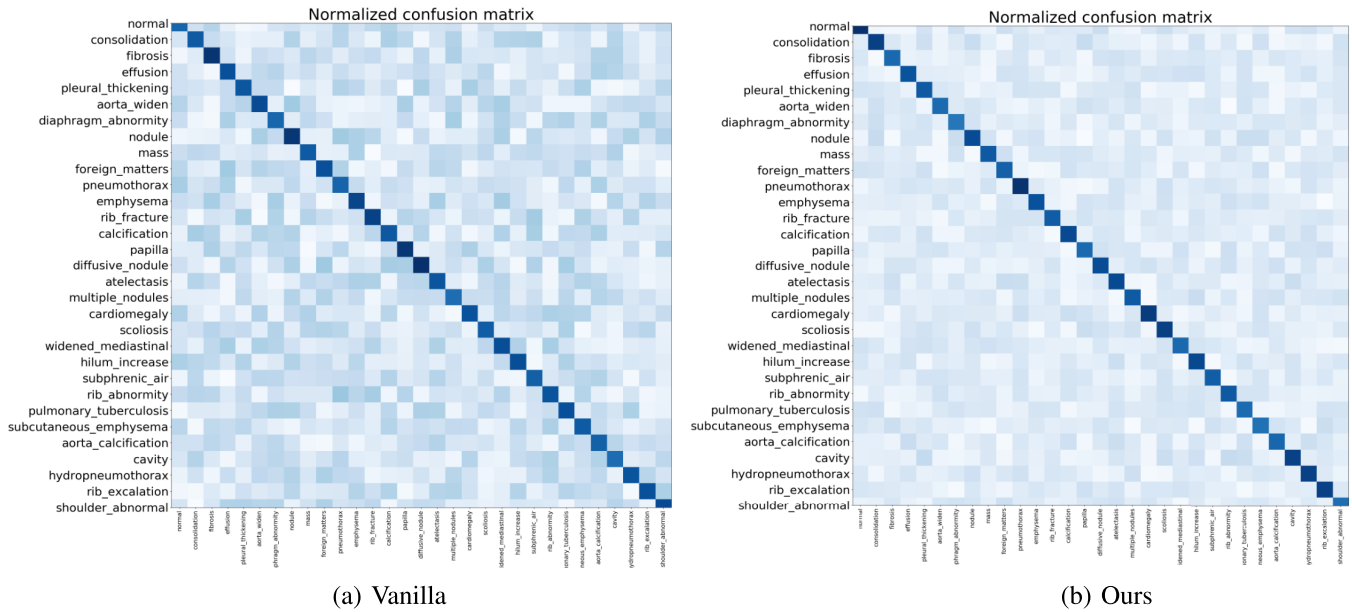
(a) Vanilla

(b) Ours

Fig. 5. Confusion matrices of disease detection on our private chest X-ray dataset: (a) vanilla Faster R-CNN FPN; (b) the variant of Faster R-CNN FPN improved by our method. A prediction successfully hits a ground-truth annotation if their classes are consistent and the IoU is larger than 0.5. It is clearly observed that our method significantly reduces false detections.

TABLE II

COMPARISON WITH BASELINE DETECTION MODELS ON THE FULLY SUPERVISED DATASET. OUR METHOD CAN SIGNIFICANTLY IMPROVE 4 EXISTING BASELINE MODELS WITHOUT LOSING MUCH COMPUTATION EFFICIENCY

Framework	Version	AP-center	AP50	AP75	Para(MB)	FPS
Faster R-CNN C4	Vanilla	35.76	26.11	7.57	130	7.0
	Ours	**39.40**	**29.17**	**9.82**	160	5.1
Faster R-CNN FPN	Vanilla	37.57	28.20	8.60	160	9.5
	Ours	**40.01**	**32.00**	**11.12**	210	7.5
RetinaNet [13]	Vanilla	38.02	28.93	10.01	145	9.2
	Ours	**40.91**	**32.16**	**11.87**	195	7.1
CenterNet [14]	Vanilla	38.19	29.02	10.11	800	3.7
	Ours	**41.02**	**33.17**	**12.34**	830	2.9

TABLE III

ABLATION STUDY ON THE FULLY SUPERVISED DATASET. CORE COMPONENTS IN OUR METHOD, INCLUDING A SPATIAL TRANSFORMER NETWORK (STN) AND OUR PROPOSED ADDITIVE AND SUBTRACTIVE FUSION (ASF), ARE BENEFICIAL FOR IDENTIFICATION AND LOCALIZATION OF DISEASES. CIA INDICATES THE FEATURES OF DISEASE PROPOSALS AND THEIR CONTRALATERAL PATCHES ARE FUSED WITH THE CONTRAST INDUCED ATTENTION [5]. AF MEANS ONLY THE ADDITION OPERATION IS USED FOR FEATURE FUSION

Proposal	STN	Fusion	AP-center	AP50	AP75
C4	w/o	w/o	35.76	26.11	7.57
	w/o	w/ CIA	36.01	27.32	8.83
	w	w/ CIA	37.20	28.27	9.46
	w/o	w/ AF	36.12	26.47	8.69
	w/	w/ AF	37.11	27.14	9.13
	w/o	w/ ASF	36.70	27.92	9.05
	w/	w/ ASF	**39.40**	**29.17**	**9.82**
FPN	w/o	w/o	37.57	28.20	8.60
	w/o	w/ CIA	38.14	30.21	9.54
	w/	w/ CIA	39.01	31.17	10.43
	w/o	w/ AF	38.65	30.11	9.27
	w/	w/ AF	39.21	31.28	10.36
	w/o	w/ ASF	38.73	30.45	9.97
	w/	w/ ASF	**40.01**	**32.00**	**11.12**

dataset includes 31,000 frontal-view X-ray images which belong to 30 disease classes. 155,000 lesions indicated by bounding boxes are carefully annotated by broad-certificated radiologists. In average, there are 5 bounding boxes per image and 5,100 bounding boxes per disease category. The distributions of disease with respect to images and bounding boxes are presented in Table I. The dataset is split into a training set of 27,000 images, a testing set of 2,100 images, and a validation set of 1,900 images. Three metrics based on average precision (AP) are utilized to evaluate the disease detection methods.

1. AP-center: If the center of the predicted box is located inside certain ground-truth box having the same disease category, the box is a true positive; otherwise, it is a false positive.

2. AP50: If the IoU between the predicted box and the ground-truth box is larger than 50%, the box is a true positive; otherwise, it is a false positive.

3. AP75: If the IoU between the predicted box and the ground-truth box is larger than 75%, the box is a true positive; otherwise, it is a false positive.

Weakly Supervised Dataset The NIH chest X-ray dataset [1] is used for weakly supervised disease detection. It contains 112,120 frontal-view X-ray images with 14 disease classes. Bounding box annotations are provided for 880 images. In this paper, we use images with class annotations during the training stage, while the 880 images with bounding box annotations are used for testing. We evaluate the performance of disease detection, following the metrics in [1], [4], [5]. The threshold of IoU for identifying true positive detections varies from 0.1 to 0.7, in step of 0.2.

B. Training Scheme

For weakly supervised disease detection, we use the SGD algorithm with the Nesterov momentum to train all the models
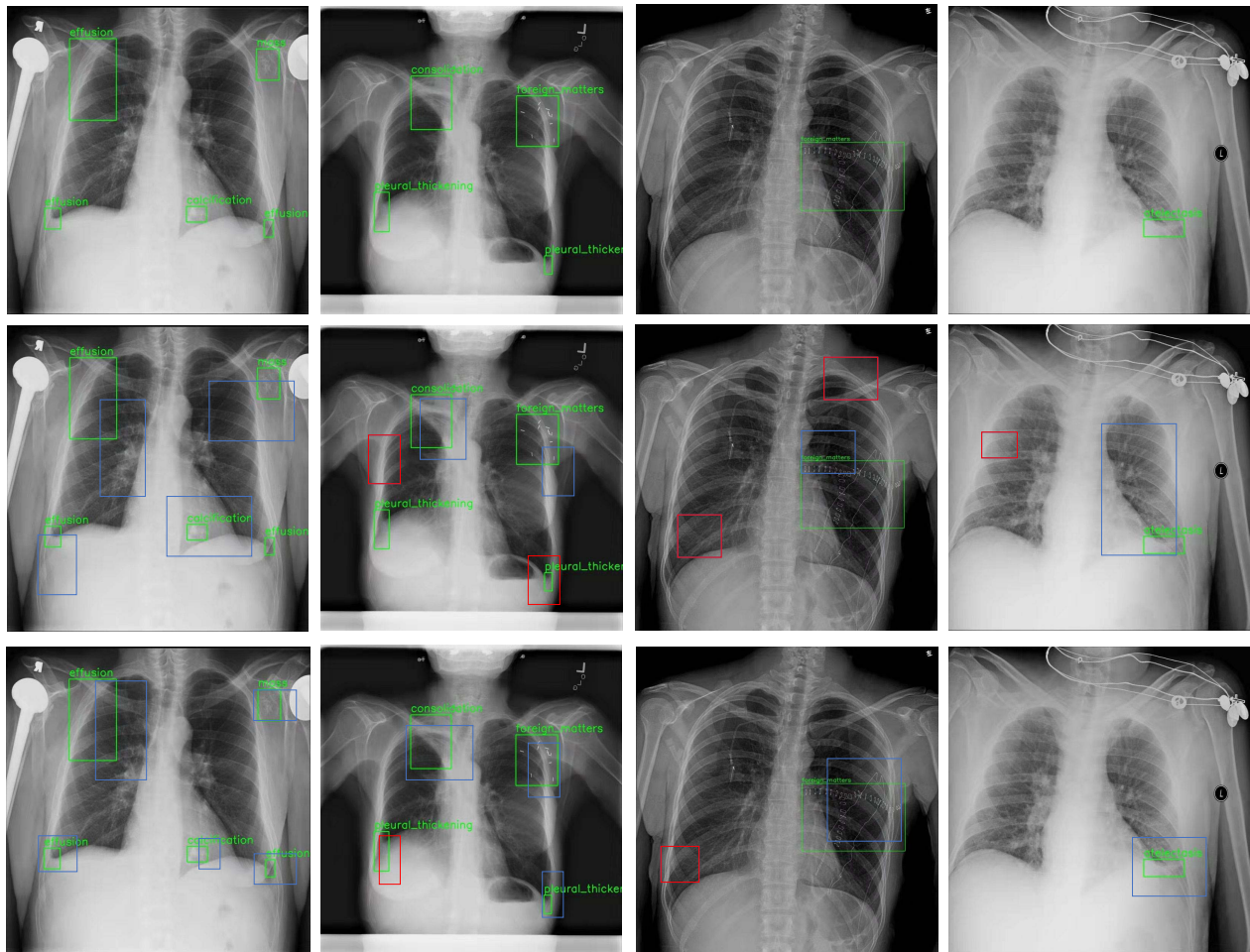
Fig. 6.   Visualization of fully supervised disease detection. The original images, results produced by 'Faster R-CNN C4' [11], and results improved by our method are shown from top to bottom. Green, blue and red boxes represents ground-truths, true positive predictions, and false positive predictions, respectively. We can see that our approach can produce more accurate localizations. The IoU threshold for detecting lesions is set as 0.1.

for 15 epochs on chest X-ray dataset. The learning rate starts with 0.001 and is reduced by a factor of 10 after every 5 epochs. In addition, the weight decay and the momentum are set to 0.0001 and 0.9, respectively. All the weights are initialized by pre-trained ResNet models on ImageNet. Our implementation is based on PyTorch. For testing. We use the threshold of 0.5 to distinguish positive grids from negative grids in the class-wise feature map.

For fully supervised disease detection, we randomly initialize all new layers by drawing weights from a zero-mean Gaussian distribution with standard deviation 0.01. All other layers (i.e., the shared convolutional layers) are initialized with parameters pretrained on ImageNet classification dataset. We use a learning rate of 0.001 for 20 epoches, and 0.0001 for the next 4 epoches. We use a momentum of 0.9 and a weight decay of 0.0005. In addition, for fully supervised tasks, the number of disease proposals is set to 512.

C. Fully Supervised Disease Detection

Spine Segmentation For training the spine segmentation network, the spine regions of 8000 images are carefully

annotated by radiologists. These images are randomly split into 10 folds. Cross validation is conducted to validate the segmentation performance. The Dice similarity coefficient (DSC) and three other metrics including pixel accuracy (P.ACC), mean accuracy (M.ACC), mean IU (M.IU) and frequency weighted IU (F.W.IU) proposed in [37] are used for evaluation. The experimental results are presented in Table V.

Disease Detection The four baseline models, including 'Faster R-CNN C4', 'Faster R-CNN FPN', RetinaNet and CenterNet, are re-implemented using our fully supervised dataset. We attempt to incorporate our proposed module into all of them as introduced in Section IV-A. The quantitative comparisons are presented in Table II. The proposed module can significantly improve the four baseline models. The improvement of AP-center brought by our method is 3.64% (from 35.76% to 39.40%) or 2.44% (from 37.57% to 40.01%), when employing 'Faster RCNN C4' or 'Faster RCNN FPN' to produce disease proposals. Under metrics AP50 and AP75, our method can also outperform all compared methods. For example, the variant of our method equipped with 'Faster R-CNN C4' achieves 29.17% AP50 and 9.82% AP75 which is 3.06 and 2.25 higher than the results of original 'Faster
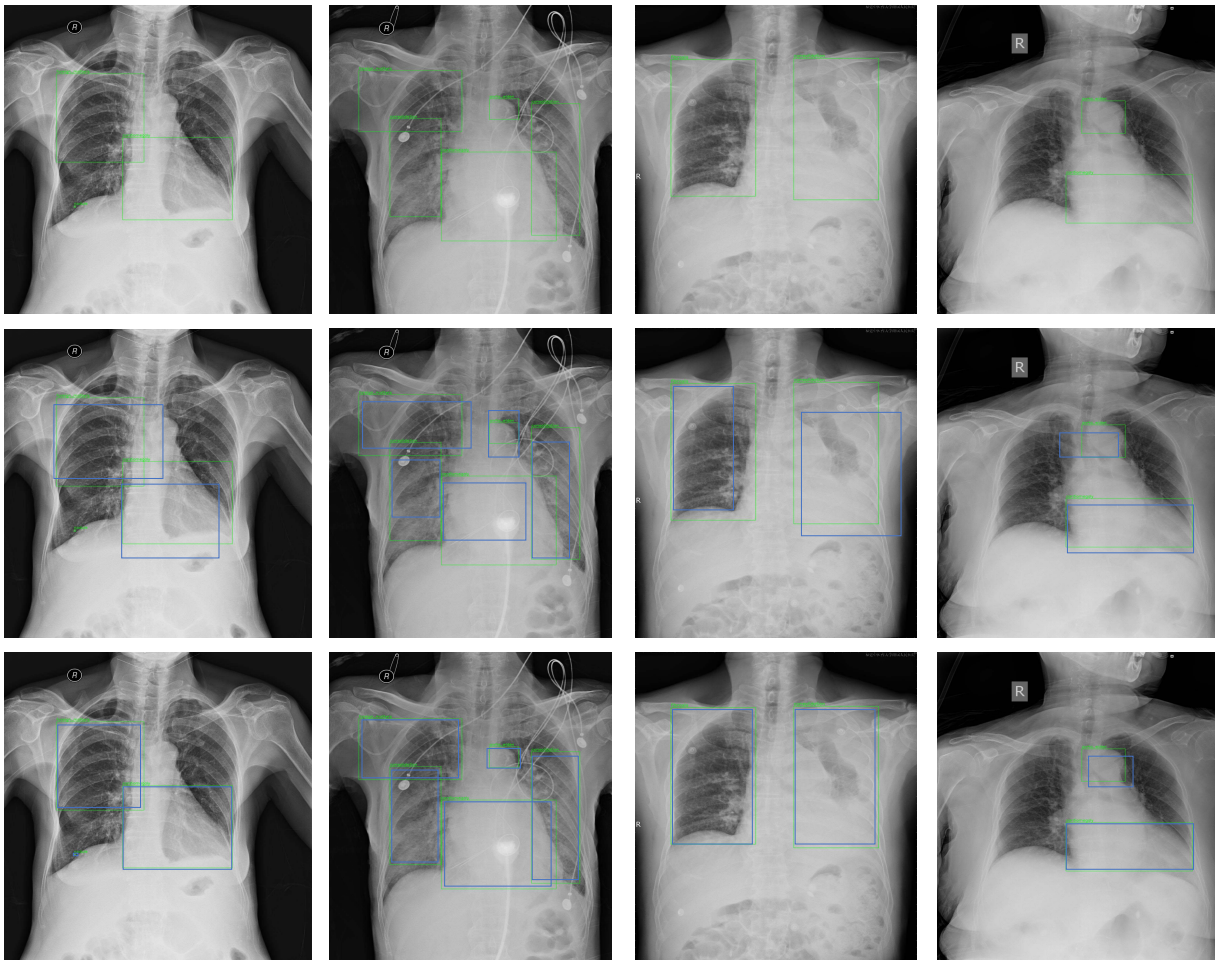
Fig. 7. Visualization of fully supervised disease detection. The original images, results produced by 'Faster R-CNN C4' [11], and results improved by our method are shown from top to bottom. Green, blue and red boxes represents ground-truths, true positive predictions, and false positive predictions, respectively. We can see that our approach can produce more accurate localizations. The IoU threshold for detecting lesions is set as 0.5.
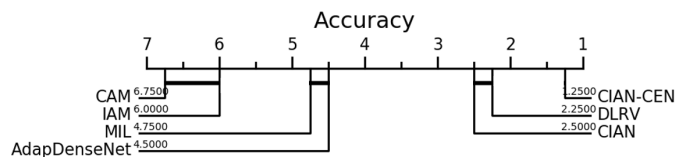

Fig. 8. We generate a critical difference diagram to make comparisons between pairs weakly supervised disease detection methods, based on the Wilcoxon-Holm method. These thick horizontal lines means the linked methods are not significantly different.

R-CNN C4' respectively. The improvements brought by our method are 3.80 (AP50) and 2.52 (AP75) when using the baseline of 'Faster R-CNN FPN'. The confusion matrices [1] of 'Faster R-CNN FPN'and its variant improved by our method are presented in Fig. 5. Results of disease detection in 4 CXRs are visualized in Figs. 6 and 7. From top to down are original images, results of the original 'Faster R-CNN C4', and results of the variant of 'Faster R-CNN C4' which is improved by our module. The green, blue and red boxes stand for ground-truths, true positives and false positives, respectively. The experimental results demonstrate that the contralateral context information extracted by our proposed module is beneficial to the detection of diseases in CXRs.

D. Weakly Supervised Disease Detection

In this task, we adopt images with only image-level labels for training, and images with both bounding box and class annotations for testing. We compare our method against various existing methods proposed in [1], [3]–[5], [29], [30]. As shown in Table IV, our method achieves the best performance in overall. Compared to the baseline model [5], our method achieves consistently higher accuracy under all IoU thresholds. For example, in case of using 0.3 as IoU threshold, it produces results with accuracy of 0.48, surpassing [5] by 0.02. When threshold of IoU is set as 0.5 and 0.7, our approach achieves accuracy of 0.36 and 0.21, with a lead of 0.01 and 0.02 over [5], respectively. In Fig. 9, we illustrate disease detection results of several cases under the condition of the weak supervision. From top to down are original images, results of [5], and results produced by our method. The green boxes stand for ground-truths and blue regions indicate predictions inferred by disease localization models. The results verify the effectiveness of our method in weakly supervised disease detection. Following [38], we generate a critical difference diagram for comparing our method against other weakly supervised methods, based on the Wilcoxon-Holm method.[2]

[1]https://github.com/kaanakan/object_detection_confusion_matrix

[2]https://github.com/hfawaz/cd-diagram/

TABLE IV

COMPARISON WITH OTHER WEAKLY SUPERVISED DISEASE DETECTION METHODS ON NIH CHEST X-RAY DATASET. T(IoU) MEANS THE THRESHOLD VALUE OF IoU USED TO MATCH PREDICTED RESULTS AND GROUND-TRUTHS. DISEASE LOCALIZATION ACCURACY ARE EVALUATED AT T(IoU)-S IN {0.1, 0.3, 0.5, 0.7}. 'MIL*' INDICATES THE RESULTS OF [4] WHICH ARE RE-IMPLEMENTED BY [5]

| T (IoU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
0.1	CAM [1]	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.57
	AdapDenseNet [30]	0.43	0.97	0.69	0.82	0.55	0.22	0.78	0.33	0.60
	IAM [3]	0.68	0.97	0.65	0.52	0.56	0.46	0.65	0.43	0.62
	MIL [4]	0.59	0.81	0.72	0.84	0.68	0.28	0.22	0.37	0.57
	MIL* [4], [5]	0.43	0.82	0.72	0.72	0.52	0.42	0.20	0.70	0.57
	DLRV [29]	0.59	0.99	0.85	0.76	0.61	0.23	0.68	0.49	0.65
	CIAN [5]	0.67	0.86	0.71	0.83	0.77	0.45	0.29	0.40	0.62
	CIAN-CEN	0.71	0.91	0.74	0.85	0.79	0.47	0.40	0.50	**0.67**
0.3	CAM [1]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.22
	AdapDenseNet [30]	0.16	0.92	0.33	0.54	0.18	0.14	0.50	0.14	0.36
	IAM [3]	0.33	0.85	0.34	0.28	0.33	0.11	0.39	0.16	0.35
	MIL* [4], [5]	0.24	0.75	0.47	0.60	0.30	0.05	0.21	0.33	0.37
	DLRV [29]	0.51	0.96	0.56	0.67	0.45	0.16	0.43	0.21	**0.50**
	CIAN [5]	0.45	0.73	0.51	0.74	0.61	0.19	0.17	0.26	0.46
	CIAN-CEN	0.47	0.75	0.55	0.76	0.62	0.21	0.19	0.29	0.48
0.5	CAM [1]	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.06
	AdapDenseNet [30]	0.06	0.78	0.23	0.28	0.08	0.14	0.21	0.05	0.23
	IAM [3]	0.11	0.60	0.10	0.12	0.07	0.03	0.17	0.08	0.16
	MIL* [4], [5]	0.15	0.67	0.30	0.41	0.22	0.02	0.14	0.12	0.25
	DLRV [29]	0.20	0.92	0.19	0.39	0.20	0.06	0.18	0.04	0.27
	CIAN [5]	0.31	0.65	0.37	0.59	0.48	0.07	0.09	0.20	0.35
	CIAN-CEN	0.32	0.68	0.39	0.61	0.49	0.07	0.15	0.21	**0.36**
0.7	CAM [1]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.01
	AdapDenseNet [30]	0.02	0.23	0.22	0.11	0.04	0.14	0.08	0.01	0.11
	IAM [3]	0.01	0.17	0.01	0.02	0.01	0.00	0.02	0.02	0.03
	MIL* [4], [5]	0.03	0.57	0.08	0.04	0.1	0.02	0.06	0.10	0.10
	DLRV [29]	0.04	0.72	0.03	0.15	0.02	0.00	0.03	0.01	0.13
	CIAN [5]	0.11	0.53	0.18	0.27	0.26	0.03	0.04	0.16	0.19
	CIAN-CEN	0.13	0.54	0.19	0.27	0.27	0.04	0.05	0.17	**0.21**
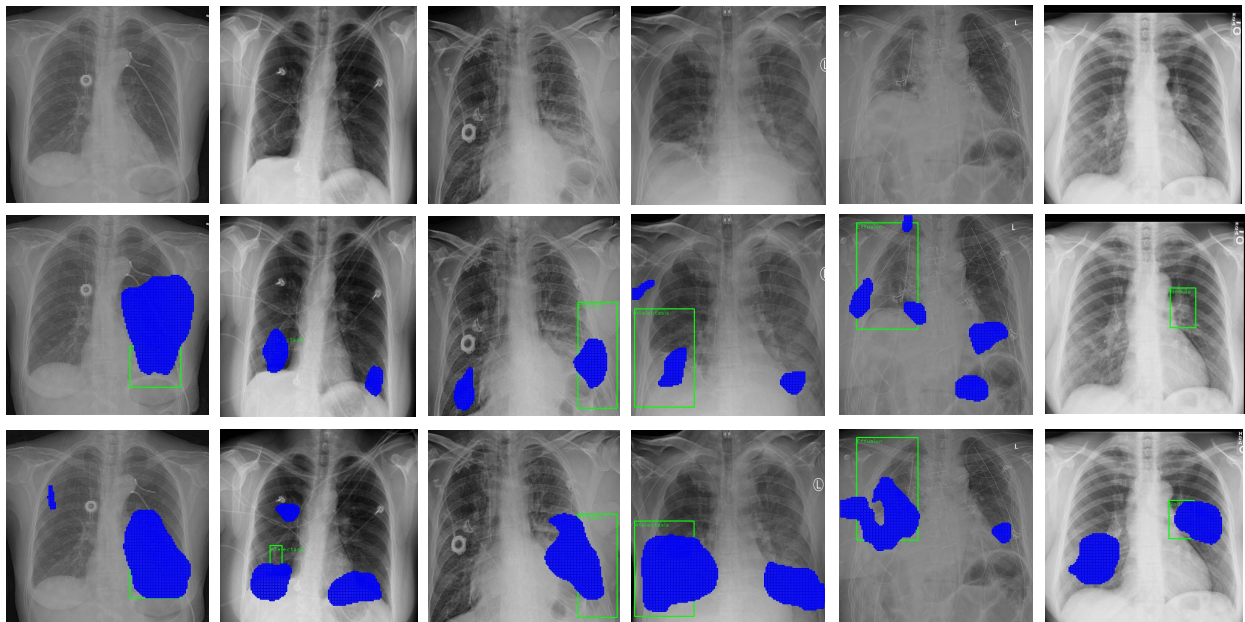
Fig. 9. Visualization of weakly supervised disease detection. The original images are presented in the first row. Results produced by [5] and our method are visualized in the second and third row respectively. Green boxes stand for ground-truths and blue regions indicate predicted results. We can see that our approach can output more accurate localization results.

The mean accuracies under different IoUs are used as the input. As shown in Fig. 8, our proposed method achieves the best rank.

E. Ablation Study

Ablation studies are conducted to discuss the effectiveness of core modules in our method. Both 'Faster R-CNN C4' and 'Faster R-CNN FPN' are used as baseline models. To validate the effectiveness of the STN, a variant of our method without STN, which uses the preliminary contralateral patches to enhance the features of disease proposals, is implemented. In contrast to our proposed additive and subtractive fusion (ASF) strategy, we also apply a counterpart of our feature module with only additive operation (AF) or the contrast
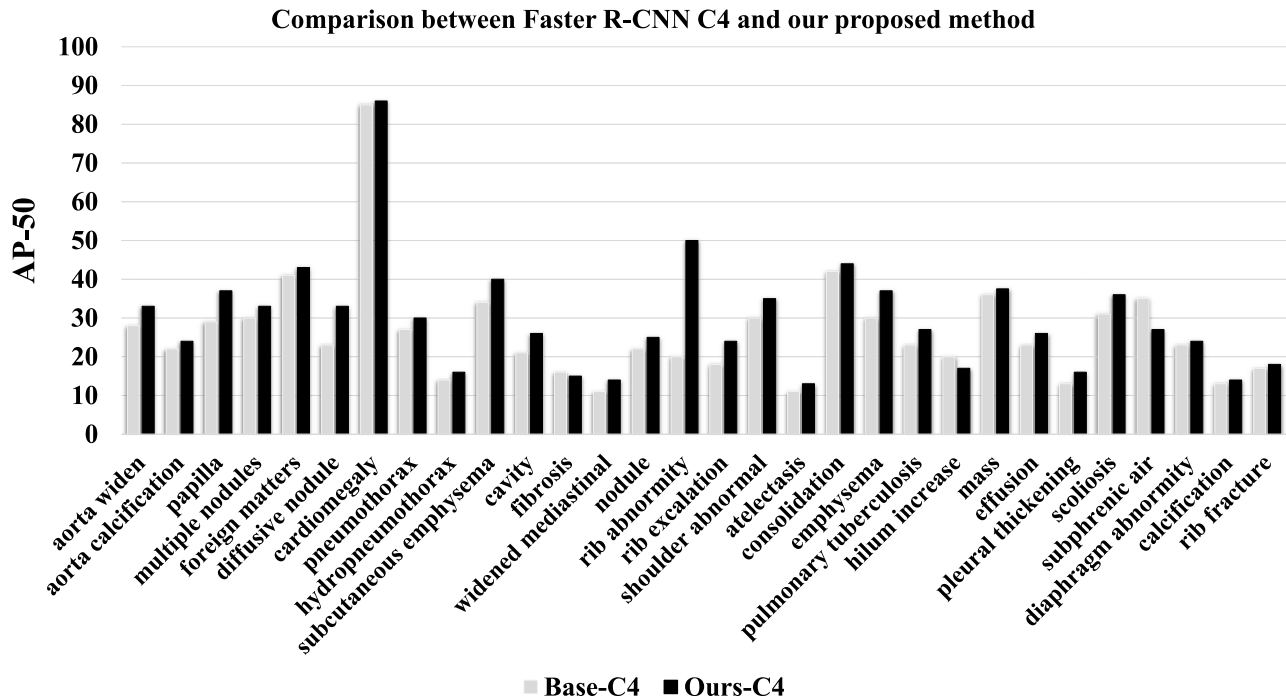
Fig. 10. The AP50 metric values of 30 disease categories. 'Faster R-CNN C4' is used to extract disease proposals in our method. The incorporation of contralateral patches is beneficial for improving the detection performance of most diseases.

TABLE V

THE PERFORMANCE OF PSPNET IN THE SPINE SEGMENTATION TASK.
DICE SIMILARITY COEFFICIENT (DSC), PIXEL ACCURACY (P.ACC),
MEAN ACCURACY (M.ACC), MEAN IU (M.IU), FREQUENCY
WEIGHTED IU (F.W.IU) ARE USED FOR EVALUATION

DICE	P.ACC	M.ACC	M.IU	F.W.IU
96.4%	97.1%	80.4%	64.2%	89.2%

induced attention (CIA) [5] which is based on the subtractive operation to fuse features of each disease proposal and its contralateral patch. As we can see in III, compared to the baseline models, the usage of preliminary contralateral patches can give rise to marginally better performance. The adoption of STN is capable of improving all detection evaluation metrics in all settings. For example, the value of AP-center is promoted from 36.70 to 39.40, once STN is adopted in the variant of 'Faster R-CNN C4' with ASF. For feature fusion strategies, our proposed ASF outperforms CIA under all conditions. For example, ASF induces detection results with 2.20 (39.40 vs 37.20) or 1.00 (40.01 vs 39.01) higher AP-center than the results of CIA when using 'Faster R-CNN C4' or 'Faster R-CNN FPN' as the baseline model, respectively. Besides, the detection performance is degraded without using any of the additive and subtractive operations. It indicates that the two operations can complement each other. In summary, exhaustive ablation studies demonstrate the superiorities of core modules devised in this paper.

F. Performance in Individual Disease Categories

The AP50 metric values of 30 disease categories produced by 'Faster R-CNN C4' and our proposed method are presented in Fig. 10. Compared with the baseline 'Faster R-CNN C4', our proposed method achieves better results in the detection of

most disease categories. Our method is particularly advantageous at detecting abnormalities residing in structures having similar counterparts in the contralateral side of the chest, such as lungs and ribs. Especially, the AP-50 of rib abnormality detection is improved from around 20% to 50% after incorporating our method into 'Faster R-CNN C4'. For diseases of organs which only exist in single half of the chest, such as cardiomegaly, our method can still improve the identification performance, since the contralateral context information can also be used to ignore overlapping or surrounding distortion signals.

For statistical analysis, the Wilcoxon signed-rank test[3] is conducted to compare the distributions of AP50 across diseases, produced by the original 'Faster R-CNN C4' and its improved variant. The sum of the ranks of the differences is 41.5, and the p-value for the test is 0.00039. Hence, we would reject the null hypothesis that the two groups of AP50 are from the same distribution under a confidence level of 5%.

The above experiments demonstrate that, the retrieved contralateral patches are able to strengthen the feature representations of proposals when identifying and locating most disease lesions.

VI. CONCLUSION

We propose a novel module, *Contralaterally Enhanced Networks*, for disease localization in chest X-ray images. Our method aims at taking advantage of the thoracic contralateral context information to enhance the feature representations of disease proposals. The spine line is regarded as the symmetry axis to obtain an initial contralateral patch for each disease proposal. Then a spatial transformer network is devised to

[3]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html

refine the pose of the initial contralateral patch. The disease proposal and its retrieved contralateral patch are fused to predict final disease classification and localization. Experiments on a carefully annotated dataset demonstrate our proposed module improves existing two-stage and one-stage detection methods, Faster R-CNN C4, Faster R-CNN RPN, RetinaNet and CenterNet, by 3.06, 3,80, 3.23 and 4.15 on the AP50 metric, respectively. Our method can also be applied in weakly supervised disease localization and achieves state-of-the-art performance on the NIH chest X-ray dataset.

VII. FUTURE WORK

As discussed above, the contralaterally enhanced networks can effectively improve the disease performance in both fully and weakly supervised disease localization in chest X-ray images. The limitations of our method are as follows: 1) The retrieval of the contralateral patches is dependent to the localization of the spine line; 2) The disease regions neglected by the proposal extraction model cannot be recovered. To settle these issues, it deserves further study to design more efficient module to explore the contralateral contextual information. Furthermore, the replication of small thoracic structures exists in the same side of the chest. Thus, how to exploit such relationships for disease detection is also an interesting research topic.

REFERENCES

[1] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[2] A. I. Aviles-Rivero *et al.*, "GraphX^{NET} chest X-ray classification under extreme minimal supervision," 2019, *arXiv:1907.10085*. [Online]. Available: https://arxiv.org/abs/1907.10085

[3] J. Cai, L. Lu, A. P. Harrison, X. Shi, P. Chen, and L. Yang, "Iterative attention mining for weakly supervised thoracic disease pattern localization in chest X-rays," in *Proc. Int. Conf. Med. Imag. Comput. Comput. Assist. Intervent. (MICCAI)*. Springer, 2018, pp. 589–598.

[4] Z. Li *et al.*, "Thoracic disease identification and localization with limited supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8290–8299.

[5] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10632–10641.

[6] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: http://arxiv.org/abs/1711.05225

[7] C. Qin, D. Yao, Y. Shi, and Z. Song, "Computer-aided detection in chest radiography based on artificial intelligence: A survey," *BioMed. Eng. OnLine*, vol. 17, p. 113, Aug. 2018.

[8] C. Brestel, R. Shadmi, I. Tamir, M. Cohen-Sfaty, and E. Elnekave, "RadBot-CXR: Classification of four clinical finding categories in chest X-ray using deep learning," in *Proc. Med. Imag. Deep Learn.*, 2018, pp. 1–9. [Online]. Available: https://openreview.net/forum?id=B1tMhcIDM

[9] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 590–597.

[10] 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.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 91–99.

[12] K. Yan, X. Wang, L. Lu, and R. M. Summers, "Deeplesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *J. Med. Imag.*, vol. 5, no. 3, 2018, Art. no. 036501.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[14] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[15] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 2015, 2015, pp. 2017–2025.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[18] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[19] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[21] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 21–37.

[22] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[23] J. Hung and A. Carpenter, "Applying faster R-CNN for object detection on malaria images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jul. 2017, pp. 56–61.

[24] N. M. Noor *et al.*, "Texture-based statistical detection and discrimination of some respiratory diseases using chest radiograph," in *Advances in Medical Diagnostic Technology* Singapore: Springer, 2014, pp. 75–97.

[25] M. Cicero *et al.*, "Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs," *Investigative Radiol.*, vol. 52, no. 5, pp. 281–287, 2017.

[26] V. Chouhan *et al.*, "A novel transfer learning based approach for pneumonia detection in chest X-ray images," *Appl. Sci.*, vol. 10, no. 2, p. 559, Jan. 2020.

[27] 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.

[28] A. T. Sahlol, M. A. Elaziz, A. T. Jamal, R. Damaševičius, and O. F. Hassan, "A novel method for detection of tuberculosis in chest radiographs using artificial ecosystem-based optimisation of deep neural network features," *Symmetry*, vol. 12, no. 7, p. 1146, Jul. 2020.

[29] C. Zhang, F. Chen, and Y.-Y. Chen, "Thoracic disease identification and localization using distance learning and region verification," 2020, *arXiv:2006.04203*. [Online]. Available: http://arxiv.org/abs/2006.04203

[30] B. Zhou, Y. Li, and J. Wang, "A weakly supervised adaptive DenseNet for classifying thoracic diseases and identifying abnormalities," 2018, *arXiv:1807.01257*. [Online]. Available: http://arxiv.org/abs/1807.01257

[31] 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.

[32] Q. Ke *et al.*, "A neuro-heuristic approach for recognition of Lung diseases from X-ray images," *Expert Syst. Appl.*, vol. 126, pp. 218–232, Jul. 2019.

[33] K. C. Santosh and S. Antani, "Automated chest X-ray screening: Can Lung region symmetry help detect pulmonary abnormalities?" *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1168–1177, May 2018.

[34] D. Eppstein, M. Overmars, G. Rote, and G. Woeginger, "Finding minimum area k-gons," *Discrete Comput. Geometry*, vol. 7, no. 1, pp. 45–58, Jan. 1992.

[35] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[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] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[38] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: A review," *Data Mining Knowl. Discovery*, vol. 33, no. 4, pp. 917–963, Jul. 2019.