Align, Attend and Locate: Chest X-ray Diagnosis via Contrast Induced Attention Network with Limited Supervision

Jingyu Liu[1*], Gangming Zhao[2*], Yu Fei[1], Ming Zhang[1], Yizhou Wang[1,2,3], Yizhou Yu[2†]

[1]Department of Computer Science, Peking University
[2]Deepwise AI Lab [3]Peng Cheng Laboratory

Abstract

Obstacles facing accurate identification and localization of diseases in chest X-ray images lie in the lack of high-quality images and annotations. In this paper, we propose a Contrast Induced Attention Network (CIA-Net), which exploits the highly structured property of chest X-ray images and localizes diseases via contrastive learning on the aligned positive and negative samples. To force the attention module to focus on abnormalities, we also introduce a learnable alignment module to adjust all the input images, which eliminates variations of scales, angles, and displacements of X-ray images generated under bad scan conditions. We show that the use of contrastive attention and alignment module allows the model to learn rich identification and localization information using only a small amount of location annotations, resulting in state-of-the-art performance in NIH chest X-ray dataset.

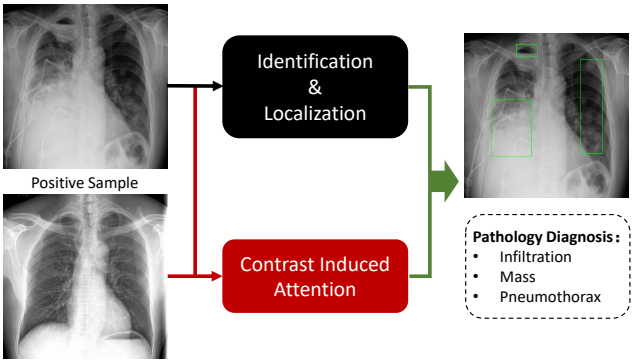

Figure 1. Our proposed framework consists of two branches. The upper branch extracts convolutional features from the input image. The lower branch computes the contrast induced attention on the extracted feature map. The information from each branch is merged to produce disease identification and localization results for the input chest X-ray images.

1. Introduction

Chest X-ray image analysis serves as a crucial role in clinical diagnosis of chest diseases. Traditionally, it requires years of accumulated expertise and consistent concentration to finish the task, adding heavy workloads to radiologists. Fortunately, we can formulate chest X-ray image analysis as a classification task, which assigns a particular type of disease to an image, together with a detection task, which provides location annotation of the abnormality. Therefore, automatic image analysis systems could be implemented with the help of deep Convolutional Neural Network (CNN) methods [28, 5, 11, 12, 27].

To achieve good performance in natural images, classic CNN approaches require tons of samples with image-level labels for image classification, and samples with both category and location labels for object detection. However, these requirements raise two challenges when it comes to chest X-ray image diagnosis. First, accurate location labels are expensive to acquire for chest X-ray images, making it hard to train an accurate detector. Second, the diversity of location, shape and texture make certain categories of abnormality vague and mutually confused.

In this paper, we propose a novel Contrast Induced Attention Network (CIA-Net) (Figure 1) to address these problems. The motivation of CIA-Net originates from the consistency of thoracic structure among humans. Through contrastive study, which exploits visual contrast between a pair of positive image (with diseases) and negative image (without diseases), CIA-Net captures additional identification and localization information in the lack of annotation. Specifically, we extract high-level image feature representations of the image pair from CNN. Then, to utilize the highly structured property of inputs, we compute L1-distance between corresponding pixels in the negative image and the positive one, the result of which serves as an indication of possible lesion sites on the latter. However, some images,

*Equal contribution
†Corresponding author

especially the positive ones, suffer from geometric deformation caused by poor scan conditions. Therefore, to rationalize the process of contrastive learning, we propose a learnable alignment module to adjust input images to be geometrically canonical. Finally, to further utilize the limited location annotation, we apply Multiple Instance Learning (MIL) to perform end-to-end training on CIA-Net. We show that with the help of the alignment module and CIA-Net, even for vague, tiny and randomly appeared lesions, CIA-Net makes more accurate predictions than previous methods.

Above all, our contribution lies in three folds:

- We propose CIA-Net, which is the first to capture information by contrasting positive and negative images. More generally, it provides inspiration to address vision tasks with samples sharing high similarity in their visual structure.

- We propose a learnable alignment module, which is effective in transforming and aligning images in different scan conditions. This technique can also be utilized in other medical images analysis tasks requiring alignment.

- We achieve state-of-the-art results on both classification and localization on ChestX-ray14.

2. Related Work

Automatic Chest X-ray Analysis The release of large scale chest X-ray datasets allows wide applications of deep learning methods on automatic chest X-ray analysis. Wang et al. [30] introduce the ChestX-ray14 dataset, which is by far the largest with 112,120 front-view images in 14 types of thoracic diseases. Before that, a large public dataset Open-i [15] containing 3,955 radiology reports and 7,470 associated chest x-rays enables usage of early deep models. However, chest X-ray datasets usually suffer from limited annotation and data.

Recent surveys [15, 25]have indicated the potential of deep learning methods in chest X-ray image classification [30, 19, 4, 34, 32] and detection [19, 18, 31]. Technically, Rajpurkar et al. [19] and Wang et al. [30] apply CNN models developed for more comprehensive datasets to address the classification task, and use class activation map (CAM) [34] to obtain locations of diseases. Yan et al. [31] add Squeeze-and-Excitation Block [6] to DenseNet [7] and utilize multi-map and max-min pooling techniques. Later, Li et al. propose to use fully convolutional neural network (FCN) [14] to address the problem. They unify the training of image-level and box-level data in one framework, with customized MIL loss. Different from previous approaches, which mainly adapt models or losses developed for other tasks, our proposed contrastive attention exploits the property of chest X-rays to address the problem.

Many works apply attention mechanism to chest X-ray analysis. Ypsilantis et al. [32] propose a recurrent attention model to identify cardiomegaly regions. Later, Pesce et al. [18] introduce a soft attention mechanism, which locates lesions with highlighting part of the saliency map generated by CNN. Guan et al. [4] use attention to generate masks, which help to amplify lesion regions. Most of these attention mechanisms are implicitly built and highly relied on the results of classification. They may suffer from noisy labels considering that the image labels are not directly from manual annotation by experts, but are mined from associated radiological reports using natural language processing [30]. While our CIA-Net focuses on relations between images and explicitly builds attention utilizing the highly structure property of data.

Object detection Object detection has long been the fundamental and studied a lot in computer vision. After the advent of deep learning, two main lines of approaches have become maturely developed in object detection. The first are the two-stage detectors, mainly based on the Region-CNN series. The second are the one-stage detectors, mainly represented by YOLO [20] and SSD [16]. In Faster R-CNN [21], the region proposal network (RPN) in the first stage pre-computes an objectness score for each candidate region and preserve the top K candidates. Then the Fast R-CNN [3] network in the second stage classifies each candidate region and adjusts their locations via regression. In YOLO, objectness score, classification and location regression are computed in the same stage. Our approach is similar in spirit with one-stage detectors. We split the image into cells and decides whether the cell is positive or not based on its overlap with the ground truth, which mimics the roles of anchor boxes in detectors.

Based on the problem setting, weakly object detection [26, 1, 33, 24, 13, 10, 22, 29] is also closely related to our approach. Given only the image-level labels, most approaches formulate object detection as a MIL problem. The loss is often designed based on the paradigm that a positive image contains at least one positive object, while a negative image contains none. Another effective method is to find the peak in the feature map or heat map, among which CAM is the most commonly used. One drawback of this line of approaches is that the localization is always partial, and heavy engineering work is needed to tune the results. Our approach performs end-to-end training and does not rely on any post-processing techniques.

3. Approach

Our proposed framework is illustrated in Figure 2 and comprises of two parts: 1. Alignment modules that automatically adjust the input images towards canonical by

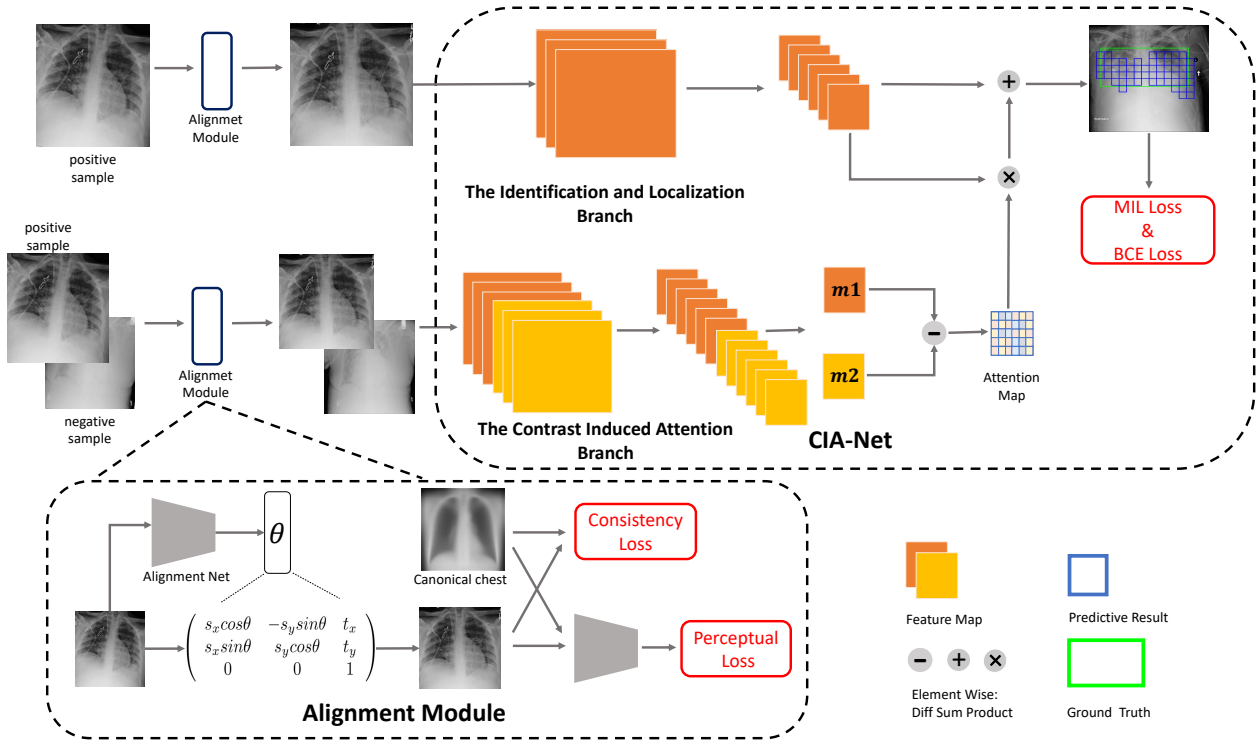

Figure 2. Our proposed framework consists of two parts: (a) The alignment module that automatically affine-transforms the input images towards canonical. (b) The CIA-Net that consists of two branches. The upper branch extracts convolutional features from the input image. The lower branch computes the contrast induced attention on the extracted feature map. The BCE loss and MIL loss take charge of box-level annotated data and class-level data respectively.

affine transformation. 2. CIA-Net that consists of two branches. The contrast induced attention branch generates attention for every class of diseases from a pair of positive and negative images. The attention that contains localization information assists the identification and localization branch to make predictions. Next, we introduce the details of every key component of our framework.

A standard high-quality front-view chest X-ray image should be upright and symmetrical. However, sometimes scanned X-ray images are far from standard due to improper distance, angle or displacement between the camera and patients. The geometric deformation of images might be approximate as affine transform, as shown in Figure 3. To enable chest X-rays to share the same structure, we introduce an alignment network to align all the images. Our alignment network is in spirit similar to the spatial transformer network (STN) [8], but we frame it with more explicit supervision. We align all the images to a single target image, which we term *Canonical Chest*. To obtain the canonical chest image, we simply randomly collect 500 negative images from the dataset. And average them to obtain an averaged image. After that, we crop out the central view tightly bounding the two lungs. The final canonical chest is shown in Figure 3(a). Statistically, we believe that the averaged chest x-ray image should approach to a standard image.

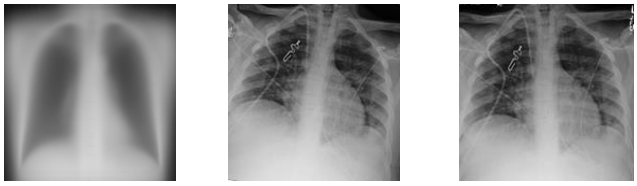

Figure 3. From left to right are the canonical chest, the original image and the aligned image, respectively.

3.1. Alignment Module

After obtaining the canonical chest as the target image, we frame the transformation learning as minimizing the structure divergence between the transformed image and the target image. Formally, let I and T denote the input image to be transformed and the target image respectively. Given I, the alignment network ϕ_A transforms I to $\phi_A(I)$. To let $\phi_A(I)$ have a standard structure, we minimize the structure loss: $L_s = f(\phi_A(I), T)$. Specifically, we use a light-weighted ResNet-18 as the backbone of ϕ_A. The

T (IoU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
0.3	X, Wang [30]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.22
	Z, Li [14]	0.36	**0.94**	0.56	0.66	0.45	**0.17**	0.39	**0.44**	0.49
	Ours	**0.53**	0.88	**0.57**	**0.73**	**0.48**	0.10	**0.49**	0.40	**0.53**
0.5	X, Wang [30]	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.06
	Z, Li [14]	0.14	**0.84**	0.22	0.30	0.22	**0.07**	0.17	0.19	0.27
	Ours	**0.32**	0.78	**0.40**	**0.61**	**0.33**	0.05	**0.37**	**0.23**	**0.39**
0.7	X, Wang [30]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.01
	Z, Li [14]	0.04	0.52	0.07	0.09	0.11	0.01	0.05	0.05	0.12
	Ours	**0.18**	**0.70**	**0.28**	**0.41**	**0.27**	**0.04**	**0.25**	**0.18**	**0.29**

Table 1. Comparison of results trained using 80% annotated and 50% unannotated images. Localization accuracy are evaluated at various T(IoU) in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. The bold values denote the best results and results are rounded to two decimal digits for readability. Our model consistently outperforms previous methods in most cases. The advantage is evident especially at high T(IoU).

T (IOU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
0.1	Z, Li [14]	0.59	0.81	0.72	0.84	0.68	0.28	0.22	0.37	0.57
	Base	**0.61**	0.88	**0.73**	0.78	0.67	0.23	0.09	0.36	0.54
	Ours	0.39	**0.90**	0.65	**0.85**	**0.69**	**0.38**	**0.30**	**0.39**	**0.60**
0.3	Base	0.33	0.71	0.34	**0.68**	0.36	0.06	0.05	0.20	0.34
	Ours	**0.34**	**0.71**	**0.39**	0.65	**0.48**	**0.09**	**0.16**	0.20	**0.38**
0.5	Base	0.19	**0.57**	0.14	**0.49**	0.21	0.01	0.03	0.08	0.21
	Ours	0.19	0.53	**0.19**	0.47	**0.33**	**0.03**	**0.08**	**0.11**	**0.24**
0.7	Base	**0.11**	**0.40**	0.06	**0.29**	0.11	0.00	0.01	0.06	0.13
	Ours	0.08	0.30	**0.09**	0.25	**0.19**	**0.01**	**0.04**	**0.07**	0.13

Table 2. Comparison of results trained using 100% unannotated images and no any annotated images. Disease localization accuracy are evaluated at various T(IoU) in $\{0.1, 0.3, 0.5, 0.7\}$. Our model outperforms [14] and our own implemented baseline model at different IoU thresholds in most cases.

output of the alignment network is a group of parameters $(t_x, t_y, s_x, s_y, \theta)$ of affine transformation. t_x and t_y stand for horizontal and vertical displacements. s_x and s_y stand for horizontal and vertical scaling. θ stands for the rotational angle. To this end, I is transformed to $\phi_A(I)$ following:

$$\phi_A(I) = B\left(\begin{pmatrix} s_x cos\theta & -s_y sin\theta & t_x \\ s_x sin\theta & s_y cos\theta & t_y \end{pmatrix} G(I), I\right) \quad (1)$$

where B stands for a bilinear interpolating function, and G represents a regular grid function.

To encourage $\phi_A(I)$ to have similar structures with T, an ideal solution is to extract the chest structure from X-ray images. However, the structure annotation is not available, so that we need to find an alternative to address the problem. Inspired by *perceptual losses* [9] that is capable of preserving content and structure in style transfer, we adopt it here in our task. Specifically, we adopt the feature reconstruction loss used in [9].

$$L_{feat}(\phi_A(I), T) = \frac{1}{CHW} \left\| N_{feat}(\phi_A(I)) - N_{feat}(T) \right\|_2 \quad (2)$$

where C, H, W are the feature map size, and N_{feat} is the network to extract features. In practice, we also use the consistency loss which computes Euclidean distances of corresponding pixels of the image pair. An exemplar pair of I and $\phi_A(I)$ are shown in Figure 3(b) and (c).

3.2. CIA-Net

Different from natural images that have flexible structures, chest X-ray images have relatively fixed structures. Basically, a positive sample (image with diseases) might have three types of visual abnormalities: 1. Opacity and complex textures caused by accumulated liquid or abnormal tissues, e.g. effusion, consolidation, and mass. 2. Over transparency caused by air, e.g. emphysema and pneumothorax. 3. Visual abnormal shape of organs, e.g. cardiomegaly. Most diseases in our evaluated dataset lie in the above three types. These abnormalities render apparent visual contrast compared with negative samples. To this end, we propose to use the visual contrast as an attention signal indicating the possible location of the disease.

As shown in Figure 2(a), the CIA-Net is composed of two branches. The upper branch extracts the convolutional feature map F_i^+ of size $c \times h \times w$ from a positive image I_i^+. The lower branch takes the positive image I_i^+ and a negative image I_i^- as a pair of inputs. The shared encoder network encodes I_i^+ and I_i^- into attention maps M_i^+ and M_i^- of sizes $h \times w$ respectively. After that, we compute the absolute difference $\Delta M = \left| M_i^+ - M_i^- \right|$ between M_i^+ and M_i^-. Finally, the spatial-wise attention map ΔM is multiplied with F_i^+ element by element to obtain the weighted feature map $\bar{F}_i^+$ as following:

$$\bar{F}_i^+ = \sum_k^{w \times h} \Delta m_k f_k \quad (3)$$

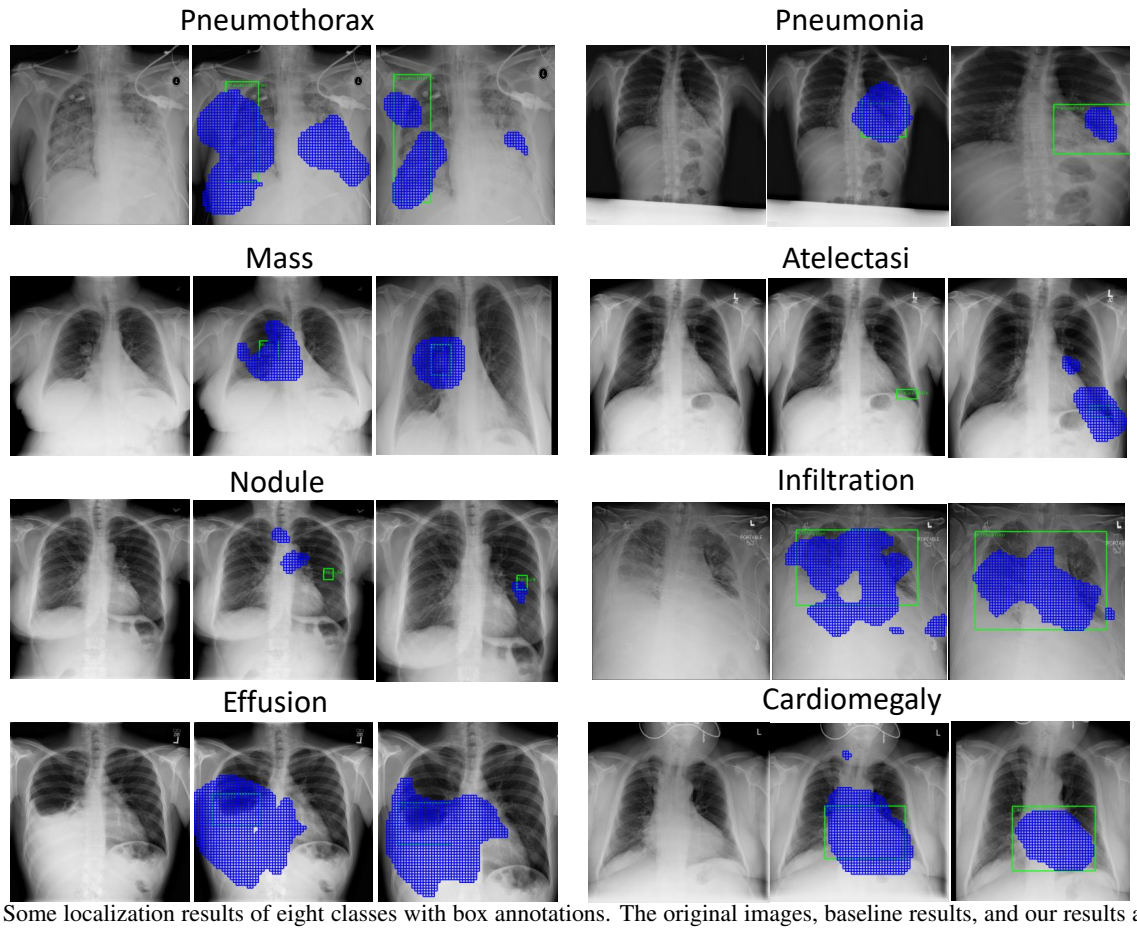

Figure 4. Some localization results of eight classes with box annotations. The original images, baseline results, and our results are shown in the left columns, the middle columns, and the right columns respectively. We can see that our approach can output more accurate localization results.

where Δm_k denotes k_{th} weight in ΔM, and f_k denotes k_{th} grid in F_i^+. We normalize ΔM to make $\sum_k \Delta m_k = w \times h$, to keep activations of F_i^+ properly scaled.

More specifically, the input images of both branches are resized to 512×512. ResNet-50 pre-trained from ImageNet dataset is adopted as the backbone for both branches. For the upper branch, we use the feature map F_i^+ after C5 (last convolutional output of 5th-stage), which is 32 times down-sampled and of size $2048 \times 16 \times 16$. For the attention branch, we use C4 (last convolutional output of 4th-stage) as the encoder module and obtain the attention blob of size $1024 \times 32 \times 32$ after 16 times down-sampled. The attention blob is then passed through a 2×2 max pooling layer and a 1×1 convolutional layer to obtain the attention map of size 16×16. **Loss function.** After obtaining the weighted feature map $\bar{F}_i^+$, we pass it through a 1×1 convolutional layer and a sigmoid layer to obtain the class-aware feature map of size $C \times H' \times W'$, where C is the number of classes. Each grid in the feature map denotes the existent probability of a disease. Then we follow the paradigm used in [14], computing losses and making predictions in each channel for the corresponding class. For images with box-level annotations, if the grid in the feature map has overlap with the projected ground truth box, then we assign label 1 to the grid, otherwise we assign 0 to it. Therefore, we use the binary cross-entropy loss for each grid:

$$L_i^k(B) = \sum_j -y_{ij}^k \log(p_{ij}^k) - \sum_j (1-y_{ij}^k) \log(1-p_{ij}^k) \quad (4)$$

where k, i, and j are the index of classes, samples, and grids respectively. y_{ij}^k denotes the target label of the grid and p_{ij}^k denotes the predicted probability of the grid.

For images with only image-level annotations, we use the MIL loss used in [14].

$$\begin{aligned} L_i^k(I) = &-y_i^k \log(1 - \prod_j (1 - p_{ij}^k)) \\ &-(1 - y_i^k) \log(\prod_j (1 - p_{ij}^k)) \end{aligned} \quad (5)$$

where y_i^k denotes the target label of the image.

T (IOU)	anno ratio	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
0.3	80%	Base	0.46	**0.86**	**0.59**	0.77	0.40	0.07	0.63	**0.51**	0.54
		Ours	**0.54**	0.82	0.55	**0.81**	**0.49**	**0.29**	**0.51**	0.40	**0.55**
	40%	Base	0.41	**0.74**	0.53	**0.79**	0.31	0.08	**0.49**	0.29	0.46
		Ours	**0.55**	0.73	**0.55**	0.76	**0.48**	**0.22**	0.39	**0.30**	**0.50**
	0%	Base	0.33	0.71	0.34	**0.68**	0.36	0.06	0.05	0.20	0.34
		Ours	**0.34**	0.71	**0.39**	0.65	**0.48**	**0.09**	**0.16**	0.20	**0.38**
0.5	80%	Base	0.27	**0.79**	**0.44**	0.55	0.23	0.04	**0.55**	**0.38**	0.41
		Ours	**0.38**	0.77	0.42	**0.63**	**0.34**	**0.26**	0.39	0.27	**0.43**
	40%	Base	0.22	**0.60**	0.34	0.56	0.19	0.03	**0.31**	0.17	0.30
		Ours	**0.36**	0.57	**0.37**	**0.62**	**0.34**	**0.13**	0.23	0.17	**0.35**
	0%	Base	0.19	**0.57**	0.14	**0.49**	0.21	0.01	0.03	0.08	0.21
		Ours	0.19	0.53	**0.19**	0.47	**0.33**	**0.03**	**0.08**	**0.11**	**0.24**
0.7	80%	Base	0.11	**0.74**	**0.33**	0.40	0.18	0.03	**0.45**	**0.25**	0.31
		Ours	**0.18**	0.71	0.31	**0.42**	**0.25**	**0.11**	0.26	0.23	0.31
	40%	Base	0.12	0.42	0.15	0.37	0.15	0.00	**0.19**	0.08	0.19
		Ours	**0.19**	**0.47**	**0.20**	**0.41**	**0.22**	**0.06**	0.12	**0.11**	**0.22**
	0%	Base	**0.11**	**0.40**	0.06	**0.29**	0.11	0.00	0.01	0.06	0.13
		Ours	0.08	0.30	**0.09**	0.25	**0.19**	**0.01**	**0.04**	**0.07**	0.13

Table 3. Localization results of models trained using different number of annotated images with 100% unannotated images.

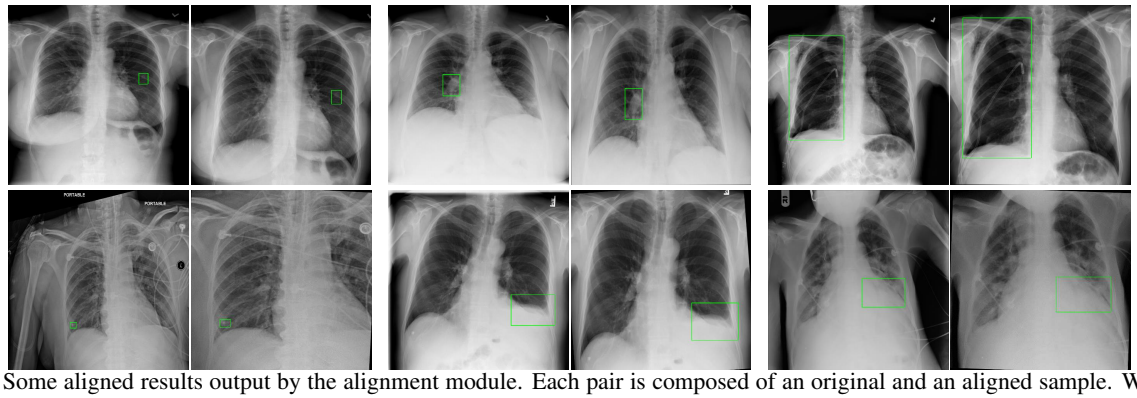

Figure 5. Some aligned results output by the alignment module. Each pair is composed of an original and an aligned sample. We can see that the aligned samples have more canonical views than the original ones.

The total loss across all classes of all samples is:

$$\sum_i \sum_k \lambda_i^k \beta_B L_i^k(B) + (1 - \lambda_i^k) L_i^k(I) \tag{6}$$

where $\lambda_i^k \in 0, 1$ denotes if the k_{th} class in the i_{th} sample has box annotation, and β_B is the balance weight of the two losses.

3.3. Training and Testing

Training. We use the SGD algorithm with the Nesterov momentum to train all the models for 15 epochs on chest X-ray dataset. For CIA-Net, we use a total mini-batch size of 6 on a single GPU. 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 [5] models on ImageNet [23]. Our implementation is based on PyTorch [17].

Testing. We use the threshold of 0.5 to distinguish positive grids from negative grids in the class-wise feature map. In practice, the feature map is up-sampled from the size of 16×16 to 128×128 to achieve more accurate predictions. The up-sampling operation is inserted before the last two convolutions.

4. Experiments

4.1. Dataset and Evaluation Metrics

Dataset. There are 112,120 frontal-view X-ray images of 14 classes of diseases in NIH chest X-ray dataset [30]. Note that each image can have different diseases. Furthermore, the dataset contains 880 images with 984 labeled bounding boxes. We follow the terms in [14] to call 880 images as "annotated" and the remaining 111,240 images as "unannotated". We resize the original 3-channel images from resolution of 1024×1024 to 512×512 without any data augmentation techniques.

Evaluation Metrics. We follow the metrics used in [14]. For localization, the accuracy is calculated by the IoU (Intersection over Union) between predictions and ground truths. Note that predictions can be discrete small rectangles. We only report localization results of the eight diseases with ground truth boxes. The localization result is re-

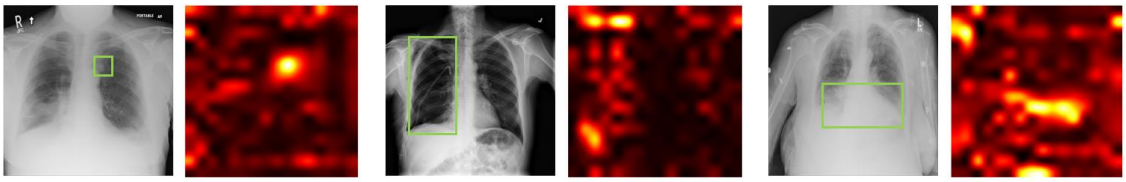

Figure 6. Attention maps generated by CIA-Net. The left shows the predicted images, where green and blue boxes stand for ground truths and predictions respectively. The right shows the generated attention maps, which provide helpful cues for locations of abnormalities.

Model	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Emphysema	Fibrosis	
Z, Li [14]	**0.80**	0.87	**0.80**	0.88	0.87	0.91	0.78	
Ours	0.79	0.87	0.79	**0.91**	**0.88**	**0.93**	**0.80**	
Model	Hernia	Infiltration	Mass	Nodule	Pleural Thickening	Pneumonia	Pneumothorax	**Mean**
Z, Li [14]	0.77	**0.70**	**0.83**	**0.75**	0.79	0.66	0.80	0.80
Ours	**0.92**	0.69	0.81	0.73	**0.80**	**0.75**	**0.89**	**0.83**

Table 4. The AUC scores of our method and the baseline. Here, 70% and 20% images are used for training and testing respectively.

garded as correct when $IoU > T(IoU)$, where T(*) is the threshold. For classification, we also utilize AUC scores (the area under the ROC curve) [2] to measure the performance of our model.

4.2. Comparison with the State-of-the-art

Disease Localization. Following [14], we conduct a 5-fold cross-validation. We design two experiments to verify the effectiveness of our method. In the first experiment, we train our model with 80% annotated images and 50% unannotated images and compare the corresponding localization accuracy with [14] and [30] (Table 1). The model is evaluated on the remaining 20% annotated images. In the second experiment, we train the model with 100% unannotated images without any annotated image and compare the localization accuracy with [14] (Table 2). The model is evaluated on all annotated images.

Table 1 shows the results of the first experiment, we show that our model performs better in most cases. Especially, when T(IoU) increases, our model gradually achieves greater improvement in all 8 classes used for evaluation over the reference models. For example, when evaluated at $T(IoU) = 0.7$, the accuracy of easy classes e.g. "Cardiomegaly" is 0.70, while the reference models achieve 0.52 [14] and 0.03 [30]. For relatively small-object classes e.g. "Nodule" and "Mass", our accuracy achieves 0.27 and 0.04 while the reference models achieve only 0.00 for both classes in [30] and 0.11, 0.01 for [14]. We also calculate the mean accuracy of all classes to compare the general performance of different methods. At $T(IoU) = 0.3$, our approach achieves accuracy of 0.53, with a 0.03 lead over [14]. At $T(IoU) = 0.5$ and $T(IoU) = 0.7$, our approach achieves accuracy of 0.39 and 0.29, with a lead of 0.12 and 0.17 over [14], respectively. Overall, the experimental results shown in Table 1 demonstrate that our approach is more capable of accurate localization, which provides greater support for clinical practices.

Table 2 shows the results of the second experiment. Since [14] only provides the results at $T(IoU) = 0.1$, we utilize the baseline model following [14] implemented by ourselves and evaluate it at $T(IoU) = 0.3, 0.5, 0.7$ for better comparison. The results at $T(IoU) = 0.1$ show that our implemented baseline has similar results with [14], validating the latter comparison. The overall results show that even without annotated data used for training, our approach can achieve decent localization results. Compared with the baseline model, our proposed approach performs better in most classes at $T(IoU) = 0.1, 0.3, 0.5$ demonstrating the advantages of our model over baseline methods. Another interesting observation is that for hard classes like "Nodule" and "Mass", our model achieves comparable results over those in the first experiment without any annotated data. The results show that our model is able to utilize information from unannotated data to make up for the lack of localization annotation and achieve good performance in some hard types of abnormality in chest X-rays.

In Figure 4, we illustrate some qualitative results in eight classes used for evaluation from the second experiment. From left to right are original images, baseline and our results. The green boxes and blue boxes stand for ground truth and prediction, respectively. It shows that our approach can produce more accurate localizations in most cases.

Disease Identification. Table 4 shows the AUC scores for all 14 classes. We compare our results with previous state-of-the-art ones [14]. We follow [14] to use 70% images for training and 20% images for testing. We can see that our model achieves better AUC scores for most diseases. The mean AUC score is improved from 0.80 to 0.83 showing the effectiveness of CIA-Net for identification.

4.3. Ablation Studies

In this section, we conduct ablation studies from three aspects. First, we explore the influence of different numbers of annotated samples on our method. Second, we study the contribution of different modules. Third, we explore different negative sampling strategies used in training and testing.

T (IOU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
0.7	Ours+Canon	0.05	0.62	0.18	0.16	0.12	0.07	0.26	0.20	0.21
	Ours+Rand	0.17	0.62	0.30	**0.46**	0.21	0.08	0.20	0.15	0.27
	Ours+Sim	**0.18**	**0.71**	**0.31**	0.42	**0.25**	**0.11**	**0.26**	**0.23**	**0.31**

Table 5. Influence of different negative sampling strategies. All models are trained using 100% unannotated and 80% annotated images. Rand: randomly sampling negative samples. Canon: always using the canonical chest. Sim: Sampling based on structural similarity.

T (IOU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
0.7	Base	0.11	0.60	0.21	0.42	0.23	0.01	**0.21**	0.11	0.23
	Base+Align	**0.22**	**0.62**	**0.24**	**0.44**	0.23	**0.02**	0.18	0.11	**0.25**
	CIA	0.06	0.64	0.24	0.46	0.24	0.04	0.26	0.14	0.26
	CIA+Align	**0.09**	**0.68**	**0.28**	0.46	**0.26**	**0.06**	**0.29**	**0.15**	**0.28**

Table 6. Influence of alignment module on localization results. All models are trained using 100% unannotated and 80% annotated images.

4.3.1 CIA-Net Gains Localization Information

As shown in Table 3, with the increasing number of annotated images, the localization accuracy will be further improved. Specifically, at $T(IoU) = 0.7$, the mean accuracy is improved from 0.22 to 0.31 when annotated images ratio in training increases from 40% to 80%. Furthermore, by using 40% annotated images, our model gains higher mean accuracy than using 0% annotated images (0.22 vs. 0.13) at $T(IoU) = 0.7$. In addition, as shown in Table 3, CIA-Net has the larger improvement when less annotated images used. Specifically, in most cases our model shows higher mean performance at $anno\ ratio = 0\%\ and\ 40\%$. The experimental results demonstrate that with the help of localization information, provided by CIA-Net, our model can work effectively with limited annotated images.

4.3.2 Negative Sampling

In the training and testing phase, we use perceptual hash algorithm to choose a similarly structured pair image for every training sample. Specifically, we generate a hash code dictionary by resizing all 63,000 negative images to 16×16 and flattening them. During training and testing, we resize every sample to 16×16 and choose the nearest hash code based on cosine distance. The corresponding negative image is then paired with the positive one and sent to the later modules. To justify this approach, we compared it with other two sampling methods: 1. Randomly sampling from negative images. 2. Utilizing the canonical chest as the negative image. From the results in Table 5, we find that structural similarity based sampling is generally better than the other 2 methods in most classes. Randomly sampling introduces to much randomness to the model making it hard to capture meaningful information with contrastive learning. The second method suffers from the domain gap between the real images and averaged one.

4.3.3 Contribution of Different Modules

Figure 5 shows some examples of original images and aligned ones. We can see that the aligned samples are more approaching the canonical chest, which is more symmetrical, vertical and focused on the thoracic cavity. Table 6 shows the quantitative contribution of the alignment module. For the baseline method, our alignment module can improve the mean localization accuracy from 0.23 to 0.25. For CIA-Net, the alignment module can also improve the mean accuracy from 0.26 to 0.28. The results prove the effectiveness of the alignment module.

In addition, by comparing CIA-Net with the baseline model, we demonstrates the effectiveness of CIA-Net. CIA-Net can improve the mean localization accuracy from 0.23 to 0.26 without the alignment module, and from 0.26 to 0.28 with the alignment module. Figure 6 shows visualized attention maps of some examples. we can see that from small lesions like Nodule, to classes of large regions like Pneumothorax and Cardiomegaly, CIA-Net can generate attention maps providing helpful cues of diseases' location.

5. Conclusion

In this paper, we propose CIA-Net to tackle the challenging problem of automatic disease diagnosis in chest X-rays, where the images share similar thoracic structures. Our proposed CIA-Net enables capturing contrastive information from pairs of positive and negative images. The contrastive induced attention can provide localization cues of the possible sites of abnormalities. To rationalize CIA-Net, we also propose a learnable alignment module to adjust all the input images to be canonical. Qualitative and quantitative experimental results on NIH Chest X-ray dataset demonstrate the effectiveness of our approach.

Acknowledgments

This paper is partially supported by Beijing Municipal Commission of Science and Technology under Grant No.Z181100008918005, National Key Research and Development Program of China with Grant No.SQ2018AAA010010, and NSFC-61772039, NSFC-91646202, NSFC-61625201, NSFC-61527804.

References

[1] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, and Deva Ramanananand Thomas S. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural network. *In International Conference on Computer Vision (ICCV)*, 2015.

[2] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[3] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[4] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927*, 2018.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[7] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[10] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. *In European Conference on Computer Vision (ECCV)*, 2016.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *In Advances in Neural Information Processing Systems (NIPS)*, 2012.

[12] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. *In Advances in Neural Information Processing Systems (NIPS)*, 1989.

[13] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, , and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. *In Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8290–8299, 2018.

[15] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[18] Emanuele Pesce, Petros-Pavlos Ypsilantis, Samuel Withey, Robert Bakewell, Vicky Goh, and Giovanni Montana. Learning to detect chest radiographs containing lung nodules using visual attention networks. *arXiv preprint arXiv:1712.00996*, 2017.

[19] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[22] Mrigank Rochan and Yang Wang. Weakly supervised localization of novel objects using appearance transfer. *In Computer Vision and Pattern Recognition (CVPR)*, 2015.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *arXiv:1611.01646*, 2016.

[25] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[26] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. deep inside convolutional networks: Visualising image classification models and saliency maps. *In International Conference on Learning Representations Workshop*, 2014.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions.

[29] Yuxing Tang, Josiah Wang, Xiaofang Wang, Boyang Gao, Emmanuel Dellandrea, Robert Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. *In Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2097–2106, 2017.

[31] Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang. Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 103–110. ACM, 2018.

[32] Ypsilantis, Petros-Pavlos, Montana, and Giovanni. Learning what to look in chest x-rays with a recurrent visual attention model. *arXiv preprint arXiv:1701.06452*, 2017.

[33] Jianming Zhang, Zhe Lin, Shen Xiaohui Brandt, Jonathan, and Stan Sclaroff. Top-down neural attention by excitation backprop. *In European Conference on Computer Vision (ECCV)*, 2016.

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.