CDT-CAD: Context-Aware Deformable Transformers for End-to-End Chest Abnormality Detection on X-Ray Images

Yirui Wu, *Member, IEEE,* Qiran Kong, Lilai Zhang, Aniello Castiglione, *Member, IEEE,* Michele Nappi, *Senior Member, IEEE,* and ShaoHua Wan*, *Member, IEEE,*

Abstract—Deep learning methods have achieved great success in medical image analysis domain. However, most of them suffer from slow convergency and high computing cost, which prevents their further widely usage in practical scenarios. Moreover, it has been proved that exploring and embedding context knowledge in deep network can significantly improve accuracy. To emphasize these tips, we present CDT-CAD, i.e., context-aware deformable transformers for end-to-end chest abnormality detection on X-Ray images. CDT-CAD firstly constructs an iterative context-aware feature extractor, which not only enlarges receptive fields to encode multi-scale context information via dilated context encoding blocks, but also captures unique and scalable feature variation patterns in wavelet frequency domain via frequency pooling blocks. Afterwards, a deformable transformer detector on the extracted context features is built to accurately classify disease categories and locate regions, where a small set of key points are sampled, thus leading the detector to focus on informative feature subspace and accelerate convergence speed. Through comparative experiments on Vinbig Chest and Chest Det 10 Datasets, CDT-CAD demonstrates its effectiveness in recognizing chest abnormities and outperforms 1.4% and 6.0% than the existing methods in AP_50 and AR on VinBig dateset, and 0.9% and 2.1% on Chest Det-10 dataset, respectively.

Index Terms—Chest X-Ray Images, Abnormality Detection, Iterative Context-Aware Feature Extractor, Frequency Pooling Block, Dilated Context Encoding Block; Deformable Transformer Detector

1 Introduction

Medical diagnosis refers to the process for determining which disease or condition explains a patient' symptoms. The required information for medical diagnosis is obtained from a patient's medical history and various medical imaging data, including functional magnetic resonance imaging (fMRI), magnetic resonance imaging (MRI), computed tomography (CT), X-Ray imaging(X-Ray), and other diagnostic tools [1], [2], [3], [4]. Chest X-Ray (CXR) Images are one of the most preferred diagnostic tools in medical practice, which has an important role in the diagnosis of thoracic diseases. There exists an increasing demand in taking CXR images, where it's reported that 129 million CXR images were acquired in the United States [5], equaling that 238 erect-view of CXR images are required for disease diagnosis annually per 1000 persons. The inherent reason of large requirement for CXR images is that CXR has several advantages over another common radiography, i.e., CT.

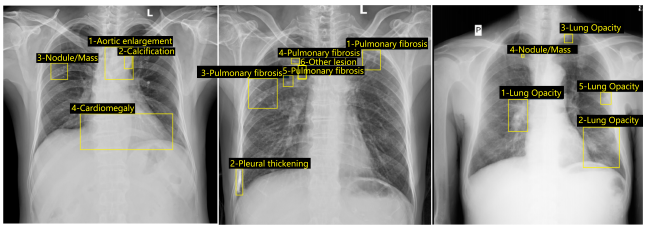


Fig. 1. Difficulties for chest abnormality detection sampled from Vinbig dataset, where (a) refers to occlusions among different abnormalities, (b) corresponds to classifying multiple and complex pathological patterns, and (c) means detecting small or subtle abnormalities.

Firstly, CXR is able to reveal some unsuspected pathologic alterations, thus accurately diagnosing various kinds of chest diseases. Secondly, it has non-invasive characteristics, since the radiation dose is relatively low comparing with CT. Last but not least, it's highly economical and affordable even in the most undeveloped countries.

Considering advantages of CXR, accurately diagnosing chest abnormality via CXR images is a highly professional work, requiring years of expertise and considerable manual efforts. Facing the increasing demand and diagnose complexity brought by CXR images, it's expected to automatically detect chest abnormality with high-potential algorithms, which will not only relieves the burden of doctors for manually recognizing, but also reduces human errors in assisting radiologists to make well-informed decisions.

Since deep learning methods have proved their effec-

- *Yirui Wu, Qiran Kong, Lilai Zhang are with Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University. They are also with the College of Computer and Information, Hohai University, Nanjing City, China, 610023. Yirui Wu is also with Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun City, China, 130117.*
- *A. Castiglione is with Department of Science and Technology, University of Naples - Parthenope, 18993 Napoli, SA, Italy, 80133.*
- *M. Nappi is with Department of Computer Science, University of Salerno, Via Giovanni Paolo II, 132, Fisciano (SA), Italy.*
- *ShaoHua Wan is with Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen City, China, 518000.*
  Email:shaohua.wan@uestc.edu.cn

Manuscript received April 19, 2005; revised August 26, 2015.

tiveness in image analysis tasks, building automatical CXR diagnose tools based on deep learning becomes a hot research topic, where its core idea is to analyze complex CXR data with highly-nonlinear modeling capability of deep learning models. Inspired by the great success of object detection methods, researchers have made tremendous progress on chest abnormality detection. For example, Baltruschat et al. [6] firstly pre-train a neural network on the ImageNet dataset for classification of natural images, and then utilize transfer learning for chest radiography analysis, which proves the efficiency of proper knowledge transfer on medical image analysis domain. Later, Annarumma et al. [7] develop a system for automated and real-time chest radiographs diagnose, which adopts an ensemble of two deep CNNs to predict clinical priorities from radiologic appearances.

Regarding the successful trials in applying deep learning models, superimposition and overlapping of different anatomical structures locate along the projection direction, leading to the diversity of chest abnormalities. Hence, detecting abnormalities from CXR images still requires to deal with difficulties, namely occlusions, multiple and complex pathological patterns, small or subtle abnormalities, as shown in Fig. 1. In fact, various patterns of chest abnormalities leads deep learning models to perform with slow convergence and high computation cost. Moreover, most of the existing methods directly derive their network structure from object detection methods on natural images, thus suffering from domain shift for low accuracy and requiring clinic knowledge to embed to boost performance.

To tackle these problems, we propose CDT-CAD, context-aware deformable transformers for end-to-end chest abnormality detection task. CDT-CAD consists of two modules, i.e., an iterative context-aware feature extractor and a deformable transformer detector. The first module iteratively fuses multi-scale features, which not only enlarges receptive fields to encode multi-scale context information via dilated context encoding blocks **(DCE)**, but also captures unique and scalable feature variation patterns in wavelet frequency domain via frequency pooling blocks **(FP)**. In fact, moving to frequency and pooling in frequency domain offers an alternative view to encode features other than only performing in spatial domain, which helps to better deal with difficulties of realistic occlusions and scale variations existed in CXR images.

Regarded as a powerful architecture for machine translation, transformer structure adaptively aggregates and refines distinguish features, thus achieving superior feature representation to solve complexity of tasks. In CDT-CAD, deformable attention blocks, i.e., the core of deformable transformer detector, attend to a small set of sampling locations as a pre-filter, which focuses on key elements out of the whole feature space, thus greatly decreasing computation and memory cost at training and testing.

The reason to adopt such architecture for medical image analysis is on the basis of two key steps for object detection task, i.e., feature extraction and classifier. Abnormality detection can be regarded as a variance of object detection on X-ray images. Therefore, we propose DCE anf FP for specially designed feature extraction, which corresponds to solve the problem of multi-scale, realistic occlusions and etc with highly distinguished characteristics. Deformable transformer structure is adopted to refine feature map by self-attention scheme and serves as a highly efficient classifier to locate abnormal and compute disease labels.

The contribution of this paper is three-fold:

- The proposed CDT-CAD could efficiently discover inherent patterns of chest abnormality. As far as we know, CDT-CAD is the first work to apply powerful deformable transformer structure for CXR diagnosis.
- The proposed iterative context-aware feature extractor iteratively re-scales and refines feature map by exploiting and fusing multi-scale interdependencies, including DCE and FP.
- A novel frequency pooling block is proposed to encode multi-scale frequency information into feature channels via wavelet transform, thus dealing with realistic occlusions and scale variations.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 presents an overview of the algorithm. Details of network structure, iterative context-aware feature extracotr, and deformable transformer detector are discussed in Section 4. Section 5 presents the experimental results and discussions. Finally, Section 6 concludes the paper.

2 Related Work

We introduce relevant research in this section, including chest X-ray image analysis, object detection methods and transformer structure.

2.1 Chest X-ray Image Analysis

Chest X-ray image analysis includes three important subtasks, i.e., image-level prediction, localization, and segmentation. Approaches adopted for chest X-ray image analysis are summarized in Table 1. Image-level prediction refers to predict a label (classification) or a continuous value (regression) with respect to the entire image. For example, Baltruschat et al. [8] compare the performance of various methods including deep learning methods, where they classify 14 classes of disease based on Chest X-ray 14 dataset. Considering the impact to analyze both front and lateral chest X-ray images, Huang et al. [9] propose Dual-Ray Net as a deep convolutional neural network, which can deal with the front and lateral chest radiography at the same time. Later, Paul et al. [10] train a model with multi-view semantic embedding and self-training technologies, which successfully perform zero-shot diagnosis of chest radiographs. Recently, Janjua et al. [11] propose a framework based on deep machine learning approaches empowered with fuzzy for object detection and classification , which can classify diagnostic objects to determine whether they are malignant or benign. Considering the presence of image artifacts such as lettering often generate a harmful bias in the classifier, Rocha et al. [12] propose Attention-driven Spatial Transformer Network for abnormality detection in chest X-Ray images.

Segmentation refers to assign category label to each pixel, which can be considered as pixel classification. Based on U-Net, i.e., a fully convolutional architecture for natural

TABLE 1
List of recent methods for chest X-ray image analysis.

Authors	Application	Year
Baltruschat et al. [8]	Image-level prediction	2019
Huang et al. [9]	Image-level prediction	2020
Paul et al. [10]	Image-level prediction	2021
Janjua et al. [11]	Image-level prediction	2022
Rocha et al. [12]	Image-level prediction	2022
Kim and Le [13]	Segmentation	2021
Que et al. [14]	Segmentation	2018
Eslami et al. [15]	Segmentation	2020
Zhang et al. [16]	Segmentation	2021
Cho et al. [17]	Localization	2020
Xi et al. [18]	Localization	2021
Han et al. [19]	Localization	2022
Ji et al. [20]	Localization	2022

image segmentation, Kim and Lee [13] use U-Net with self-attention scheme for lung segmentation, meanwhile Que et al. [14] propose DenseNet-based network for segmentation and classification of cardiomegaly. Later, Eslami et al. [15] design an image-to-image translation network for multi-task segmentation in Chest X-ray radiography, successfully generating bone-suppressed images and organ-segmented images at the same time. Recently, ABMDRNet [16] gains complementary information from visible (RGB) and thermal(T) images, and uses bridging-then-fusing strategy to deal with the reduced discriminability caused by modality differences.

Localization refers to the identification of a specific area, typically indicated by drawing a bounding box as well as the corresponding class label. It's noted that the proposed CDT-CAD could be classified as a localization work. For example, Cho et al. [17] adopt a YOLOv2 based structure with multi-scale scheme to detect 5 different classes of chest abnormalities. Later, Xi et al. [18] propose an attention-driven weakly supervised algorithm, which design explicit ordinal attention constraints to enable principled models' training in a weakly-supervised fashion. To extend contrastive learning to medical image domain, Han et al [19] propose an end-to-end semi-supervised knowledge-augmented contrastive learning framework tailored for the medical images, which seamlessly integrates radiomic features as knowledge augmentation means. Recently, Ji et al. [20] propose PBC as an abnormality localization framework, which utilize a small number of fully annotated CXRs with lesion-level bounding boxes and extensive weakly annotated samples by points.

2.2 Object Detection Methods

Many chest abnormality detection methods are inspired by object detection methods. One of the most famous cases is Faster R-CNN [21], which is a typical two-stage detector by generating region proposals via the RPN module at first, and then determing the final detection results based on the region proposals. Another famous one-stage object detector is called as YOLO [22], which divides the input image into many grids and each grid should offer hints on whether objects are inside or not. Afterwards, YOLO v4 [23] appears as a significantly improvement, which adopts technologies of the modified spatial attention module, path aggregated network, cross iteration normalization and many other components to achieve high accuracy and fast speed.

Intending to solve the problem of multi-scale feature encoding, Feature Pyramid Network [24] generates multi-scale feature maps in parallel, and adopts a divide and conquer strategy for the predictions of different sizes. Later, Retina Net [25] presents a novel Focal Loss, which makes training process focus on a sparse set of hard examples and prevents the vast number of negatives to overwhelm the detector during training. Using a single convolution neural network, CornerNet [26] detects an object bounding box, which is regarded as a pair of keypoints and corners on the main diagonal lines of the bounding box. Recently, Sangaiah et al. [27]propose a method for conserving position confidentiality of roaming PBSs users using machine learning techniques. Zhang et al. [28]design a multiple features fusion method and propose a correlation filter object function model called Spatial-Channel Selection and Temporal Regularized Correlation Filters to improve tracker performance. Most recently, FCOS [29] directly makes dense predictions on each pixel of the feature map and introduces centerness, which could be used to filter low-quality results and for refinement.

2.3 Transformer Structure

Transformers are firstly introduced as a new structure block with attention mechanism for machine translation [30]. Since a sequence can be computed by transformer in parallel, transformers are more suitable than RNN when dealing with a long sequence, thus being more popular in language processing problems than traditional RNN. As transformers are limited by a fixed-length context in the setting of language modeling, Transformer-XL [31] is proposed to enable learning dependency beyond a fixed length without disrupting temporal coherence, which manages to generate reasonably coherent, novel text articles with thousands of tokens. To increase computation efficiency, routing transformer [32] proposes a clustering-based attention mechanism that learns the attention sparsity in a data driven fashion. Inspired by the idea that Neural Architecture Search (NAS) could be used to search for more efficient transformers, Wang et al. [33] propose HAT (Hardware-aware Transformers), where hardware efficiency feedback is used as a reward signal.

Transformers have been applied in many other domains, such as recommendation systems [34], computer vision and so on. For example, VIT [35] represents an image as patches of words and uses transformers to process these words as in the NLP task, achieving better results than CNN on extremely large datasets. Since vision transformers usually suffer from high computation costs, Data-efficient image Transformers (DeiT) [36] is proposed, requiring less data and less computing resources to generate a high-performance image classification model.

Most recently, researchers propose DETR [37] as a transformer-based detector, which directly uses transformer to map feature maps into detection results. Since DETR has a slow convergence speed and limited feature resolution, Deformable DETR has been proposed to deal with those problems [38], where its attention module only focuses on several sampling points near the reference point as the key element in the attention module.

3 The Proposed Method

In this section, we first describe the overall network structure and loss function. Then, we introduce the proposed iterative context-aware feature extractor, including dilated context encoding (DCE) block and frequency pooling (FP) block. Finally, we offer descriptions on deformable transformer detector.

3.1 Network Structure Overview

As shown in Fig. 2, we design two main modules, i.e., iterative context-aware feature extractor and deformable transformer detector. The former module contains ResNet-50 backbone, dilated context encoding (DCE) block, frequency pooling (FP) block and positional encoding structure, while the latter one contains transformer encoder, transformer decoder and a feed-forward network for classification and regression tasks.

The proposed context-aware feature extractor first adopts ResNet-50 to process input CXR image as feature map. Then, we design an iterative feature fusion scheme, which iteratively refines feature map with both DCE blocks and FP blocks. The former one is capable to enlarge receptive fields, thus locally encoding multi-scale information by convolutional filters with different sizes, while the latter one directly resize feature map based on output of DCE blocks, which acts as a multi-scale feature encoding scheme in a global sense. Afterwards, we flatten the resulting feature map with abundant local and global information, resulting in a sequence feature map for further processing. Finally, positional encoding scheme is adopted to add spatial context information, thus better dealing with the problem of permutation invariance of transformers.

Regarding the sequence feature map with abundant multi-scale context information in both image and frequency domain as input, the proposed deformable transformer detector is used to directly map the input into a set of abnormality predictions. Specifically, the proposed transformer encoder adopts deformable attention block to select a small set of sampling locations as a pre-filter for prominent key elements, thus acting as feature subspace for computation and memory decreasing. Afterwards, the proposed transformer decoder takes the input as queries, which adaptively aggregates the key contents according to the attention weights that measure the compatibility of query-key pairs. Finally, a feed-forward network is adopted to make final predictions on categories and locations of abnormalities.

Note that CDT-CAD requires a fixed number N_{obj} for possible predictions, each with a coordinate regression results and an abnormality classification result. Let y the ground truth and $\hat{y} = \{\hat{y}_i\}_{i=1}^{N_{obj}}$ denotes the set of N_{obj} predictions. The total loss for both regression and classification tasks is achieved by searching for a permutation $\omega \in \Omega_{N_{obj}}$ of the N_{obj} predictions with Hungarian algorithm, which could be described as:

$$\hat{\omega} = \underset{\omega \in \Omega_N}{\arg\min} \sum_{i=1}^{N} L_{\text{match}} \left(y_i, \hat{y}_{\omega(i)}\right) \tag{1}$$

where y is padded to the size of N_{obj}, $y_{\hat{\omega(i)}}$ is the ith element of the predictions. Each element of the prediction refers to $\hat{y}_{\omega(i)} = \left(\hat{p}_{\omega(i)}\left(c_i\right), \hat{b}_{\omega}(i)\right)$, where $\hat{b}_{\omega(i)}$ represents the bounding box and $\hat{p}_{\omega(i)}\left(c_i\right)$ represents the probability of the class with the maximum probability.

The loss function for training is a combination of the box loss and classification loss, which is defined as:

$$L(\hat{y}, y) = \sum_{i=1}^{N} \left[\alpha_1 L_{cls}\left(c_i, \hat{p}_{\omega(i)}\left(c_i\right)\right) + \alpha_2 L_{loc}\left(b_i, \hat{b}_{\omega_{(i)}}\right)\right] \tag{2}$$

where the classification loss is the cross entropy, represented as

$$L_{cls}\left(c_i, \hat{p}_{\omega(i)}\left(c_i\right)\right) = \sum_{i=1}^{N} -\log \hat{p}_{\omega(i)}\left(c_i\right) \tag{3}$$

and the bounding box loss is

$$L_{loc}\left(b_i, \hat{b}_{\omega_{(i)}}\right) = \sum_{i=1}^{N} \left[\beta_1 L_{iou}\left(b_i, \hat{b}_{\omega_{(i)}}\right) + \beta_2 L_{reg}\left(b_i, \hat{b}_{\omega_{(i)}}\right)\right] \tag{4}$$

which is essentially the summation of IoU loss and L_1 loss. It's noted that α_1, α_2, β_1 and β_2 are all hyper-parameters. Specifically, we adopt GIoU [39] to balance the loss between large and small objects. Parameters of the proposed CDT-CAD method are updated based on the loss obtained by the best search of permutation, which enables the proposed network to be trained in an end-to-end manner without many hand designed components.

3.2 Design of Iterative Context-Aware Feature Extractor

The proposed context-aware feature extractor consists of three parts, i.e., iterative feature fusion scheme, DCE block, and FP block.

Iterative Feature Fusion Scheme. We design the proposed iterative feature fusion scheme for multi-scale feature fusion as shown in Fig. 2. Essentially, the proposed feature fusion scheme builds on top of the Feature Pyramid Networks (FPN) [24] by iteratively and progressively refining scaled feature map from the top layers to the bottom-up ones. Unrolling the iterative structure to a sequential implementation, we obtain feature map for abnormality detector that looks at the images twice or more with structures of multiple stages, and much more carefully with DCE and FP blocks to enhance feature representation in both spatial and frequency domains. Similar to the cascaded detector in Cascade structure, the proposed feature fusion scheme iteratively enhances original feature map of FPN to generate increasingly powerful representations. In other words, the proposed feature fusion scheme acts as a multi-scale feature encoding scheme in a global sense by directly resizing feature map, meanwhile DCE and FP blocks encodes multi-scale information in a local sense by not only enlarging receptive via fields convolutional filters with different sizes, but also encoding the multi-scale frequency information into feature channels via wavelet transform. Such iterative feature fusion operations could be represented as

$$\begin{cases} F_l = F_{l-1} + F_{DCE}(F_{l-1}) \\ F_{l+1} = f_{down}(F_{FP}(F_l)) \end{cases} \tag{5}$$

where F_l refers to the lth feature map after $l-1$ times down sampling operations, functions $f_{DCE}()$, $f_{FP}()$ and $f_{down}()$

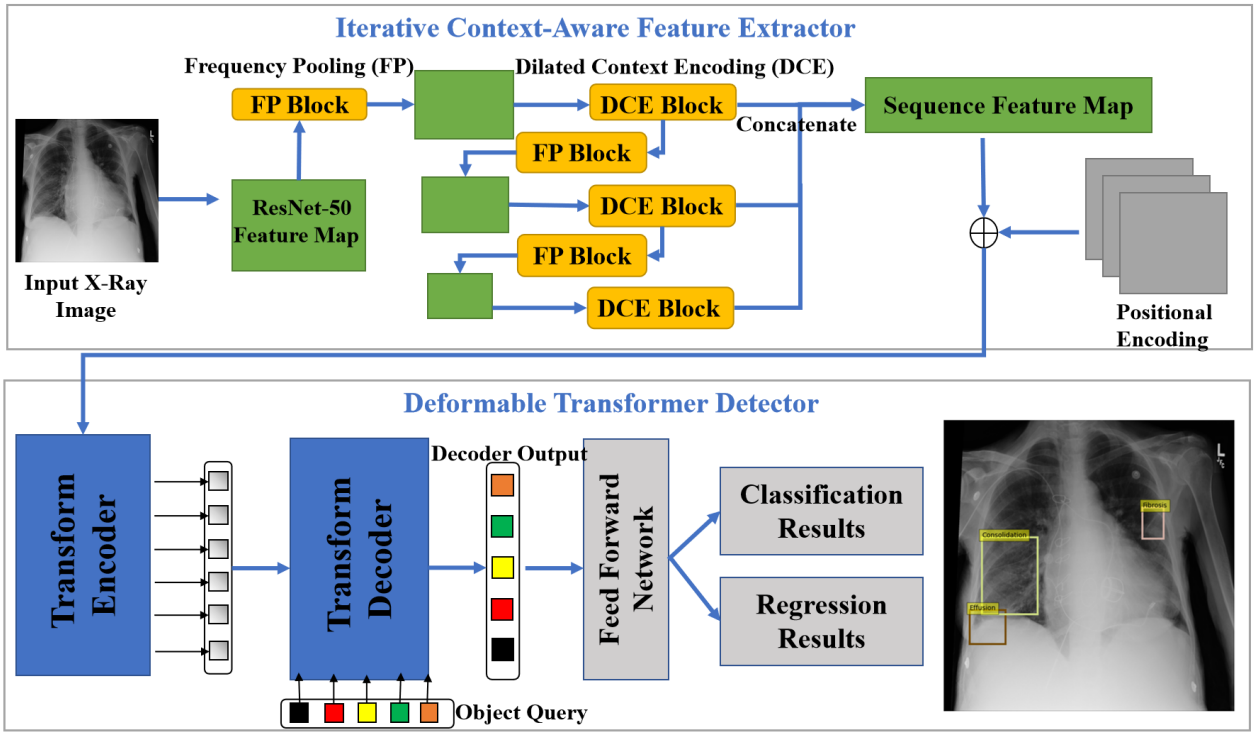


Fig. 2. Network architecture of the proposed CDT-CAD method. It's noted that CDT-CAD could output a set of predictions without pre- and post-processing steps in an end-to-end manner.

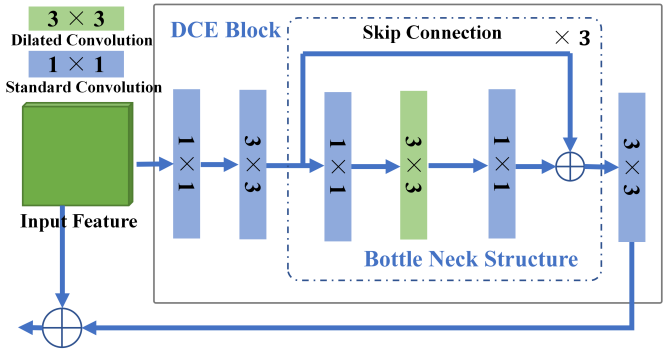


Fig. 3. Architecture design of the proposed DCE block, where the dilated convolution filter is adopted to enlarge the receptive field, thus acquiring quantity of multi-scale context information for further processing.

represent single-in-single-out operator of DCE block, operators of FP block and down-sampling operator, l varies from 2 to 4 in the proposed method.

Dilated Context Encoder Block. Inspired by YOLOF [40], we design structure of DCE block as shown in Fig. 3, where dilated convolution and skip connections are used to enlarge the receptive field and capture more local context information. Essentially, this powerful one-level feature successfully finds a way to generate an output feature with various receptive fields, compensating for the lack of multiple-level features. Therefore, it exceeds the range of scales matching to the scaled feature's receptive field, which benefits the detection performance for abnormalities across various scales.

Specifically, we first design a 1×1 and a 3×3 standard convolution layer as a projector, which is used for feature refinement. The main component in DCE block is the residual block, which consists of two 1×1 convolution layer with a 3×3 dilated convolution layer. Then, we stack several residual blocks with residual connection to build a short-way for gradient flow. Each residual block has a different dilated rates with different receptive field, covering all scales and extracting extensive contextual information. Finally, we sum the resulting feature map with the original feature map for output.

Frequency Pooling Block. Frequency transform is a powerful tool for content analysis, since images or signals are general to be sparse on an appropriate DWT basis. This property makes it easy to filter noise or informative part out of feature channels, which is considered to be an ideal methodology to be adopted in pooling structures to separate useful part from original feature channels. Moreover, we compare the frequency distribution of different CXR images, which generally show obvious differences in the frequency domain. The advantages of analyzing and the corresponding output orders of frequency transform motivate us to construct a frequency based pooling module. Among different possible frequency transform methods, we further choose wavelet transform due to its multi-scale property, which coincides with one of the problems in CXR task, i.e., scale variance. In fact, encoding frequency information with different size of windows in DWT offers a hierarchical and complementary view to analyze input CXR images, which is extremely helpful to filter the unnecessary or informative part for decrease of computation cost.

The structure design of Frequency Pooling Module is shown in Fig. 4. Specifically, we start from revisiting the

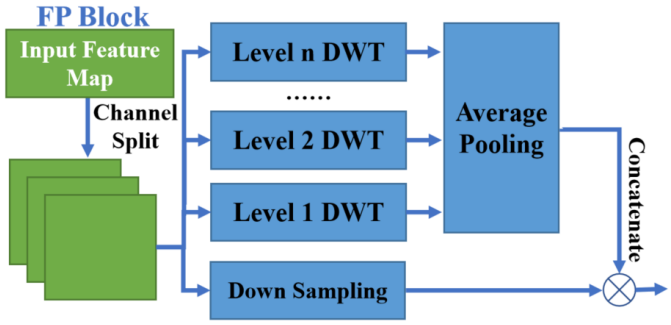


Fig. 4. Architecture design of the proposed FP block, where wavelet transform brings benefits of robustness to degradation of occlusion variations by focusing on dominant frequency information representing saliency parts of CXR content.

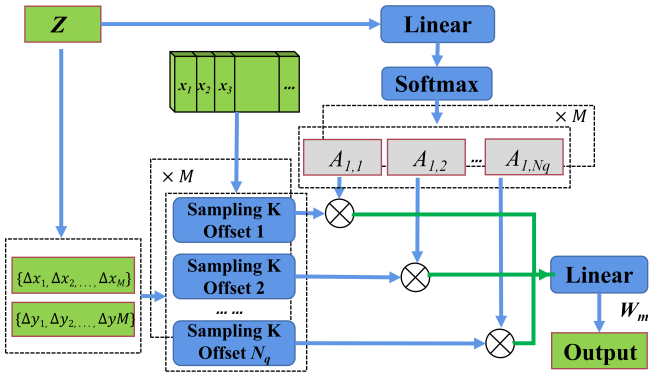


Fig. 5. Architecture design of the proposed deformable attention block, which is the core component of deformable transformer detector. It's noted that K points are sampled from the input multi-scale feature map.

equation of 1D-DWT, which can be represented as:

$$x_{\alpha,L}[n] = \sum_{k=0}^{K-1} x_{\alpha-1,L}[2n-k]g[k] \tag{6}$$

$$x_{\alpha,H}[n] = \sum_{k=0}^{K-1} x_{\alpha-1,L}[2n-k]h[k] \tag{7}$$

where $g[k]$ and $h[k]$ represent low-pass and high-pass filter with window size k, respectively. In this way, the low-frequency component $x_{\alpha,L}$ and the high-frequency component $x_{\alpha,H}$ corresponding to the αth layer are extracted.

Following the definition of 1D-DWT, 2D-DWT separates a signal into low-frequency parts and high-frequency parts along the horizontal direction as L and H at the first step. Afterwards, 2D-DWT separates L and H in the vertical direction, which computes four individual frequency parts, i.e., LL, LH, HL and HH. It's noted the upper left part (LL) contains the dominate content-related information, which can be decomposed into signal parts with different resolutions. In other words, multi-scale frequency domain can be extracted by adopting different number of levels in 2D-DWT. We show this particular design of 2D-DWT in the right part of Fig. 4.

Input feature channels $X \in R^{C\times H\times W}$ are firstly split into n parts, namely $\{X_1, X_2 \dots X_n\}$. After processing via the i-layer wavelet transform, we could get $Y_i \in R^{C/i\times H/(2^i)\times W/(2^i)}$. Then, we perform global average pooling on the decomposed parts to achieve multi-scale frequency domain information. Finally, we concatenate all these compressed parts in channel direction to compute the output feature $Z \in R^{C\times 1\times 1}$:

$$Y_i = DWT^i(X_i), where\ i = 1, 2, 3, \dots n \tag{8}$$

$$Z = GAP(Y_1) + GAP(Y_2) + \cdots + GAP(Y_n) \tag{9}$$

where operator $+$ represents concatenate operation, DWT^k means the kth level DWT processing step, and function $GAP()$ means global average pooling.

It should be noted that the proposed frequency pooling block is designed as a plug and play module, so it is not an integral part of context-aware feature extractor structure. In other words, context-aware feature extractor can maintain functional integrity without it.

3.3 Design of Deformable Transformer Detector

Self-attention transformer is a powerful network that can automatically aggregate key and distinguish features, thus discovering complex and inherent patterns from input scratches. However, directly using standard self-attention transformer is suboptimal, since it will look over all possible locations of the entire feature map to compute reasonable attention weights. Such searching strategy not only brings computation burden and memory cost, but also makes training difficult and slow to converge.

Deformable attention block is capable to focus on informative parts of the input feature, which firstly samples k points from all possible locations and then computes the corresponding attention weights on the sub feature map. As the computation cost is largely reduced by computing on feature subspace, deformable attention can attend to multi-scale feature maps for simultaneous and light-weight computation, thus successfully encoding context information on feature subspace on multi-scale feature map.

The proposed deformable attention block is illustrated in Fig. 5 with single-scale and multi-head attention property. Given a sequence input feature, we first obtain query feature $\boldsymbol{z}$ and feature map $\boldsymbol{x}$ via several linear layers. By applying linear layers on $\boldsymbol{z}$, we can compute multi-head offsets $\{\Delta \boldsymbol{x}_m, \Delta \boldsymbol{y}_m\}_{m=1}^M$ and the corresponding attention weights A. It's noted that each pair of offsets is used to sample k points from the feature map $\boldsymbol{x}$. Afterwards, single-scale and multi-head deformable attention block can be defined as:

$$At(A, x) = \sum_{m=1}^{M} W_m[\sum_{k=1}^{K} A_{m,k} \cdot f_{off}([x, \Delta x_m, \Delta y_m]_k)] \tag{10}$$

where m indexes the attention head, k indexes the sampled keys, M and K are total number of attention heads and sampling points, respectively.

Furthermore, the efficiency property of single-scale and multi-head deformable attention block leads multi-scale deformable attention block to be easily built as:

$$MsAt(\{A^l\}_{l=1}^L, \{x^l\}_{l=1}^L) = \sum_{m=1}^{M} \boldsymbol{W}_m[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{m,k}^L \cdot f_{off}([x^l, \Delta x_m^l, \Delta y_m^l]_k)] \tag{11}$$

where l refers to the index of layers and L is the total number of layers. We stack 6 deformable encoder and decoder layers with deformable attention blocks to achieve decoder output, whose size is (N_{obj}, c_{out}). It's noted that N_{obj} is the number of the abnormalities detected and c_{out} is the output dimension of decoder layers.

4 Experiments and Analysis

In this section, we first give introduction to dataset and measurements. Then, we conduct comparative studies with several existing methods. Afterwards, we perform ablation experiments to show the effectiveness of the proposed DCE blocks and frequency pooling blocks. Additionally, we analyze training performance of CDT-CAD. Finally, we offer implementation details for readers' convenience.

4.1 Datasets and Measurements

We adopt two datasets to conduct chest X-Ray abnormality detection, i.e., Vinbig Chest X-Ray Dataset and ChestX Det-10 Dataset. For former dataset, we select a subset for experiments, which contains 5000 training images and 1063 testing images in total. With annotations of bounding boxes and the corresponding class labels, all images are labeled by a panel of experienced radiologists for the presence of 14 critical radiographic findings. The latter one is a subset Dataset with box annotations of a public dataset NIH Chest-14, which contains 3001 and 541 images in the training set and testing set. Specifically, we follow each of their guidance use total 3000 images and 1000 images per class for training, total 1800 images and 600 images per class for validation, and 1604 images for testing. Respectively, It's noted each image is annotated with 10 common categories of diseases.

To evaluate the performance of detection results, we follow the evaluation rules of both datasets, where AP is defined as the mean precision value over multiple IoU thresholds (Intersection over Union) and all the object classes:

$$IoU = \frac{\text{area}\left(B_p \cap B_{gt}\right)}{\text{area}\left(B_p \cup B_{gt}\right)} \tag{12}$$

IoU is defined as the area of the intersection divided by the area of the union of the predicted bounding box.

$$AP_{U_j} = \frac{1}{10 \times C} \sum_{i=1}^{C} \sum_{j=1}^{10} P\left(i, U_j\right) \tag{13}$$

where i and j refer to the index of class and threshold respectively, C is the total number of classes, the IoU values U_j corresponds to a range from 0.5 to 0.95 with a step size of 0.05, and the function $P(i, U_j)()$ calculates precision values for the ith object class under a fixed IoU threshold U_j. Moreover, AP_{50} and AP_{75} refer to mAP(Mean Average Precision) values over the IoU thresholds of 0.5 and 0.75 respectively, while AP_S,AP_M and AP_L are the AP for small, medium and large objects, respectively.

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{14}$$

mAP is defined as the average of K-class APs.

Furthermore, we apply AR for evaluation, where AR represents the average recall at IoU threshold from 0.5 to 1, and mAP is adopted as evaluation metric with IoU threshold setting as 0.5 and 0.75 respectively.

$$AR = 2 \int_{0.5}^{1} \text{recall}(o) do \tag{15}$$

where o is the number of detected objects. Specifically, a prediction is considered as positive only if it has a larger IoU than threshold (0.5 and 0.75) with any ground truth.

4.2 Ablation Study

In this subsection, we conduct three groups of ablation experiments, where the first one is to prove the effectiveness of the proposed iterative context-aware feature extractor, the second group is designed to show the effectiveness of CDT-CAD with or without bottleneck structure in DCE block, and the last one is performed to compare performance of CDT-CAD with different number of DWT levels in FP block.

The Effectiveness of Proposed Extractor. We show statics of the first ablation experiment in Table. 2, where we can observe that a larger number of layers (as 4 for both experimental datasets) settled in iterative context-aware feature extractor contributes to a higher AP_{50} ,AP_{75},AP_S,AP_M and AP_Lvalue. The inherent reason lies in the enhancement ability for feature representation of the serial-connected DCE and FP blocks in each iteration, where DCE blocks enlarge receptive fields to encode multi-scale context information via dilated context encoding blocks, and FP blocks capture unique and scalable feature variation patterns in wavelet frequency domain. However, it's noted that AP_{50},AP_{75},AP_S,$AP_M$and AP_Lfails to increase after exceeding 4 layers in layer number, where we believe too many layers bring noisy information for feature map, thus preventing to achieve accurate CXR diagnose results.

The Performance of Bottle Neck Structure. Details of the second ablation experiment are presented in Table. 3. Higher AP_{50} ,AP_{75},AP_S,$AP_M$and AP_L achieved by CDT-CAD with 3 bottle neck structures for VingBig and ChestX Det-10 dataset demonstrate the capability in improving CXR detection results with a few number of bottle neck structures. In fact, more bottle neck structures bring advantages of serial dilated convolutions to generate feature map, thus offering diversity on different size of receptive fields.

The Performance of Different DWT Levels. Table. 4 shows quantitative comparative results with various number of DWT levels in FP block. We observe that more DWT layers lead to a higher AP_{50}, AP_{75},AP_S,$AP_M$and AP_L,which proves that encoding multi-scale frequency information could help in accurately localizing abnormalities in chest images. It's shown that all the filters perform well when the DWT level is fixed to 3. When the level is increased to 4, the accuracy remained unchanged. The reason that more DWT levels fail in contributing to higher accuracy lies in the fact, that high-level of down-sampling frequency parts contain nearly rare information for disease diagnosis task. As explained in the theory of DWT, low-frequency part corresponds to the dominate information of image content, which is proved by the fact that we can recover an image based on only low-frequency part. With higher levels of sampling in DWT, high-frequency part proves to be useless for disease diagnosis task, which should be filtered for benefits of low computation cost.

TABLE 2
Performance comparison with different number of layers in iterative context-aware feature extractor.

Dataset	No.of Layers	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Vin Big	0	33.9	15.1	3.2	12.2	21.1
	1	34.5	15.3	3.4	12.3	21.3
	2	35.1	15.5	3.4	12.3	21.4
	3	35.7	15.7	3.5	12.5	21.6
	4	**36.3**	**15.8**	**3.6**	**12.6**	**21.7**
	5	36.1	15.6	3.4	12.4	21.5
ChestX Det-10	0	41.9	15.3	4.9	15.0	25.1
	1	42.5	15.7	5.1	15.1	25.3
	2	43.0	15.9	5.2	15.2	25.4
	3	43.4	16.2	5.4	15.2	25.7
	4	**43.6**	**16.5**	**5.6**	**15.4**	**25.8**
	5	43.5	16.3	5.5	15.3	25.6

TABLE 3
Performance comparison with different number of serially-connected bottle neck structures in DCE block.

Dataset	No.of Bottle Neck	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Vin Big	0	34.7	15.2	3.1	12.0	21.0
	1	35.3	15.4	3.2	12.2	21.2
	2	36.1	15.5	3.4	12.5	21.4
	3	**36.3**	**15.8**	**3.6**	**12.6**	**21.7**
	4	36.1	15.7	3.5	12.4	21.5
ChestX Det-10	0	41.9	15.9	4.9	15.0	25.1
	1	42.5	16.1	5.1	15.1	25.3
	2	43.3	16.2	5.5	15.2	25.5
	3	**43.6**	**16.5**	**5.6**	**15.4**	**25.8**
	4	43.4	16.4	5.4	15.3	25.4

4.3 Comparison with Existing Methods

Experimental results of performance comparison on VinBig Dataset and ChestX Det-10 Dataset are shown in Table. 5. Among comparative studies, Yolov3 uses a single neural network to predict bounding boxes and class probabilities, directly estimating from images in one-round evaluation. Next, DETR reasons about the relations of the objects and the global image context to directly output the final set of predictions. Meanwhile, Cascade R-CNN, a multi-stage object detection framework, is designed to avoid the problems of overfitting at training and mismatch at inference. Afterwards, we use DenseNet, a convolutional network architecture, to introduce direct connections between any two layers with the same feature-map size, which alleviate the vanishing-gradient problem and strengthen feature propagation. It is claimed that we achieve public codes of Cho et al. [17] and Ji et al. [20] for testing. It's noted that YoLo Modified adopts DenseNet as its backbone, Faster R-CNN Modified adopts special design of data augmentation for better performance, and Faster R-CNN with FPN adopts FPN to bring property of multi-scale feature map. We follow the code instructions online to implement three ensemble network structures of 5 detectors, 3 detectors and 3 detectors with lighter structure design.

From Table. 5, we could observe that accuracy in VinBig Dataset is generally lower than ChestX Det-10 Dataset, since CXR images in VinBig Dataset not only correspond to more categories of abnormalities, but also vary in appearance with more complex patterns. It's observed that the proposed CDT-CAD has achieved the highest AP_{50},AP_{75},AR, AP_S,$AP_M$and AP_L on both datasets, which outperforms Cho et al. [17] and Ji et al. [20], Faster R-CNN, YoLo and their modified versions by a large margin. All these facts prove structures of deformable transformer detector, iterative dilated context encoder and frequency pooling are helpful to improve detection accuracy.

On the challenging VinBig dataset, CDT-CAD achieves competitive performance comparing with the ensemble baseline 1, which is a complicated structure that ensembles the results of five different detectors. So are the other two ensemble baseline methods. All these facts point out that complexity in structure design not always brings advantages on performance boosting. When comparing with DETR, the better performance obtained by CDT-CAD shows that the proposed deformable attention block can help focus on informative feature subspace without having to look over the entire space, which might bring noise information to decrease accuracy of detection results. When comparing with Cho et al. [17], the proposed deformable transformer structure brings high distinguish capability with large receptive field due to its self-attention scheme, being larger than that of convolutional filters adopted in Cho et al. [17]. Moreover, We outperforms Ji et al. [20] which serves as a benchmark for weakly semi-supervised abnormality localization in chest x-rays, focusing on inferences based on partly labeled situation rather than labeled samples.

In Fig. 7, we compare the abnormality detection accuracy between ground truth and detection results achieved by CDT-CAD and Faster R-CNN, where we can view that CDT-CAD is capable to detect hard cases ignored by Faster R-CNN, such as nodules proved by the first column of examples, and complicated pattern of disease proved by the

TABLE 4
Performance comparison with different number of DWT levels in FP Block.

Dataset	No.of DWT Levels	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Vin Big	0	35.6	15.2	3.2	12.1	21.2
	1	36.0	15.4	3.3	12.3	21.3
	2	36.0	15.6	3.5	12.4	21.5
	3	**36.3**	**15.8**	**3.6**	**12.6**	**21.7**
	4	36.1	15.7	3.4	12.5	21.6
ChestX Det-10	0	43.0	15.7	5.0	15.1	25.2
	1	43.3	15.9	5.2	15.2	25.4
	2	43.5	16.1	5.4	15.3	25.6
	3	**43.6**	**16.5**	**5.6**	**15.4**	**25.8**
	4	43.2	16.3	5.5	15.3	25.5

TABLE 5
Performance comparison among CDT-CAD and the existing methods, bold texts refer to the best performance.

Dataset	Method	AP_{50}	AP_{75}	AR	AP_S	AP_M	AP_L
VinBig	Faster R-CNN with FPN	29.1	13.6	29.1	1.7	7.3	14.4
	Yolov3 [22]	26.2	11.5	24	1.4	6.8	12.2
	DETR [37]	33.7	15.1	32.9	2.5	10.5	18.4
	Cascade R-CNN [41]	33.5	14.4	29.9	2.4	10.8	18.6
	Yolo Modified	29.5	13.4	28.2	2.1	8.5	16.8
	Faster R-CNN Modified	30.3	14.2	31.5	2.5	8.9	17.5
	Ensemble Model 1	35.7	15.2	32.4	3.3	12.5	20.8
	Ensemble Model 2	34.3	14.8	32.6	3.2	11.9	19.7
	Ensemble Model 3	33.9	13.6	33.1	3.1	11.5	19.8
	Cho et al. [17]	35.3	15.1	33.1	3.0	12.3	20.9
	Ji et al. [20]	35.8	15.3	33.4	3.2	12.5	21.4
	CDT-CAD	**36.3**	**15.8**	**35.4**	**3.6**	**12.6**	**21.7**
Chest Det-10	Faster R-CNN with FPN	39.3	15.6	45.3	4.2	13.2	22.8
	Yolov3 [22]	37.7	15.9	39.3	4.1	13.8	23.2
	DETR [37]	41.5	16.3	47.7	4.4	13.4	24.6
	Cascade R-CNN [41]	41.1	14.4	46.5	4.2	13.5	23.9
	DenseNet [42]	42.7	15.1	47.9	4.8	14.8	24.5
	Cho et al. [17]	42.8	15.9	47.2	5.2	15.3	25.3
	Ji et al. [20]	43.2	16.2	47.3	5.3	15.2	25.4
	CDT-CAD	**43.6**	**16.5**	**48.2**	**5.6**	**15.4**	**25.8**

TABLE 6
Comparison results on mean precision and convergence speed between CDT-CAD and DETR.

Dataset	Method	AP_{50}	AP_{75}	Epoch
VinBig	DETR	33.5	12.8	1000
	CDT-CAD	36.3	15.6	400
ChestX Det-10	DETR	41.5	14.4	1000
	CDT-CAD	43.5	16.5	400

last column of examples.

4.4 Training Time Analysis

As shown in Table. 6, CDT-CAD achieves much lower training epochs when comparing with DETR on both VinBig and ChestX Det-10 datasets. Meanwhile, CDT-CAD grantees performance to be much higher than that of DETR on both datasets. In other words, CDT-CAD performs not only better in accuracy , but also faster in convergence than DETR. The reason of such performance lies in the fact that the proposed iterative context-aware attention extractor successfully extracts multi-scale attention information to boost accuracy performance without increasing computation burden.

Fig. 6 shows the training loss of CDT-CAD with either standard transformer or deformable transformer. We can clearly view that deformable transformer offers higher loss speed than the standard one, thus leading to fast convergence and less computation cost. Such phenomenon can be explained that deformable transformer detector utilize a small set of key points for further calculation, thus leading the detector to focus on informative feature subspace and accelerate convergence speed.

Using the hardware configuration in implementation details, on the Chest Det-10 Dataset, it take an average of 1.251 seconds for a single image to complete the target detection process. Considering high-performance video cards are not popular in the medical environment of most developing countries, we use CPUs(in implementation details) for testing. In this case, the time consumption become 5.32 seconds.

4.5 Implementation Details

All our experiments were conducted on a server with two Intel Xeon E5-2620 v4 (@2.1GHz) CPUs and 4 NVIDIA GTX TITAN XP graphic cards. Our experimental codes are mainly based on the PyTorch framework. For data augmentation, we use random clip, resize and crop. We train our network will multiple size. Our initial learning rate is set as 0.0001, weight decay is 0.0001 and the momentum is 0.9. Due to the linear warm up mechanism, the learning rate increases from $1/3 \times 0.01$ to 0.01 in the first 500 iterations.

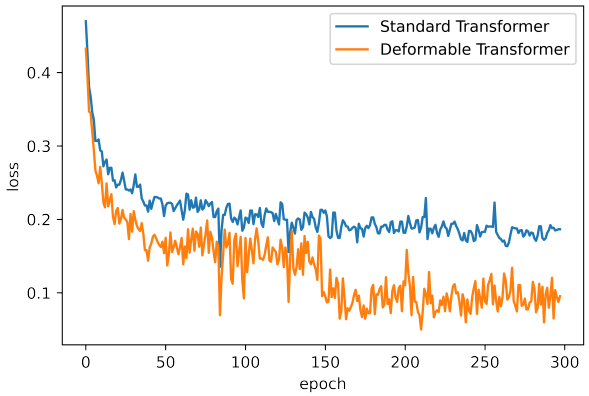


Fig. 6. Comparisons of training loss of CDT-CAD with standard transformer and deformable transformer on chestX Det 10 dataset.

We firstly use Adam optimizer to pre-train our network on MS COCO dataset and then finetuned on the chest X-Ray image dataset. We choose ResNet-50 as the backbone network and finetune backbone network with ImageNet pretrained backbones. On VinBig and Chest Det-10 Dataset, finetune 800 epochs for the detection frame subnet, reduce the learning rate to 1/10 of the original for every 300 epochs, and finetune 100 epochs for the segmentation head.

5 Conclusion

In this paper, we present a context-aware deformable transformers for end-to-end chest abnormality detection on X-Ray images. The proposed method firstly constructs an iterative context-aware feature extractor to not only encode multi-scale context information by dilated context encoding blocks, but also encode multi-scale frequency information into feature channels via frequency pooling blocks. Afterwards, we build a deformable transformer detector for abnormality detection with properties of accelerating convergence speed. Comparative experiments on Vinbig Chest and ChestX Det-10 Dataset prove that the proposed CDT-CAD method is effective and efficient for chest abnormality detection on X-Ray images. Our feature work includes ideas of interpretable deep learning methods for knowledge embedding explanation on abnormality detection results.

Medical diagnosis and analysis has become the most important and core application scenario of artificial intelligence in the medical field. CNNs for medical diagnosis has been successful, but their conventional formulation is limited to data structured in an ordered, grid-like fashion. Now, graph-based deep learning for medical diagnosis has attracted attention because graph neural networks can exploit implicit information present in biological systems.

Acknowledgments

This work was supported in part by a grant from National Key R&D Program of China under Grant No. 2021YFB3900601, the Fundamental Research Funds for the Central Universities under Grant B220202074, the Fundamental Research Funds for the Central Universities, JLU, and the Natural Science Foundation of China under Grant 61702160.

References

[1] S. Ding, H. Wang, H. Lu, M. Nappi, and S. Wan, "Two path gland segmentation algorithm of colon pathological image based on local semantic guidance," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[2] B. Ni, Z. Liu, X. Cai, M. Nappi, and S. Wan, "Segmentation of ultrasound image sequences by combing a novel deep siamese network with a deformable contour model," *Neural Computing and Applications*, pp. 1–15, 2022.

[3] H. Wang, D. Zhang, S. Ding, Z. Gao, J. Feng, and S. Wan, "Rib segmentation algorithm for x-ray image based on unpaired sample augmentation and multi-scale network," *Neural Computing and Applications*, pp. 1–15, 2021.

[4] Y. Zhao, H. Li, S. Wan, A. Sekuboyina, X. Hu, G. Tetteh, M. Piraud, and B. Menze, "Knowledge-aided convolutional neural network for small organ segmentation," *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1363–1373, 2019.

[5] F. A. Mettler, M. Bhargavan, K. Faulkner, D. B. Gilley, J. E. Gray, G. S. Ibbott, J. A. Lipoti, M. Mahesh, J. L. McCrohan, M. G. Stabin, B. R. Thomadsen, and T. T. Yoshizumi, "Radiologic and nuclear medicine studies in the united states and worldwide: frequency, radiation dose, and comparison with other radiation sources–1950-2007." *Radiology*, vol. 253, no. 2, pp. 520–531, 2009.

[6] I. M. Baltruschat, L. Steinmeister, H. Ittrich, G. Adam, H. Nickisch, A. Saalbach, J. von Berg, M. Grass, and T. Knopp, "When does bone suppression and lung field segmentation improve chest x-ray disease classification?" in *Proceedings of 16th IEEE International Symposium on Biomedical Imaging*, 2019, pp. 1362–1366.

[7] M. Annarumma, S. J. Withey, R. J. Bakewell, E. Pesce, V. Goh, and G. Montana, "Automated triaging of adult chest radiographs with deep artificial neural networks," *Radiology*, vol. 291, no. 1, pp. 196–202, 2019.

[8] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest x-ray classification," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[9] X. Huang, Y. Fang, M. Lu, F. Yan, J. Yang, and Y. Xu, "Dual-ray net: automatic diagnosis of thoracic diseases using frontal and lateral chest x-rays," *Journal of Medical Imaging and Health Informatics*, vol. 10, no. 2, pp. 348–355, 2020.

[10] A. Paul, T. C. Shen, S. Lee, N. Balachandar, Y. Peng, Z. Lu, and R. M. Summers, "Generalized zero-shot chest x-ray diagnosis through trait-guided multi-view semantic embedding with self-training," *IEEE Trans. Medical Imaging*, vol. 40, no. 10, pp. 2642–2655, 2021.

[11] J. I. Janjua, T. A. Khan, and M. Nadeem, "Chest x-ray anomalous object detection and classification framework for medical diagnosis," in *2022 International Conference on Information Networking (ICOIN)*. IEEE, 2022, pp. 158–163.

[12] J. Rocha, S. C. Pereira, J. Pedrosa, A. Campilho, and A. M. Mendonça, "Attention-driven spatial transformer network for abnormality detection in chest x-ray images," in *Proceedings of IEEE International Symposium on Computer-Based Medical Systems,*, 2022, pp. 252–257.

[13] M. Kim and B.-D. Lee, "Automatic lung segmentation on chest x-rays using self-attention deep neural network," *Sensors*, vol. 21, no. 2, p. 369, 2021.

[14] Q. Que, Z. Tang, R. Wang, Z. Zeng, J. Wang, M. Chua, T. S. Gee, X. Yang, and B. Veeravalli, "Cardioxnet: automated detection for cardiomegaly based on deep learning," in *Proceedings of IEEE Engineering in Medicine and Biology Society*, 2018, pp. 612–615.

[15] M. Eslami, S. Tabarestani, S. Albarqouni, E. Adeli, N. Navab, and M. Adjouadi, "Image-to-images translation for multi-task organ segmentation and bone suppression in chest x-ray radiography," *IEEE Trans. Medical Imaging*, vol. 39, no. 7, pp. 2553–2565, 2020.

[16] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2633–2642.

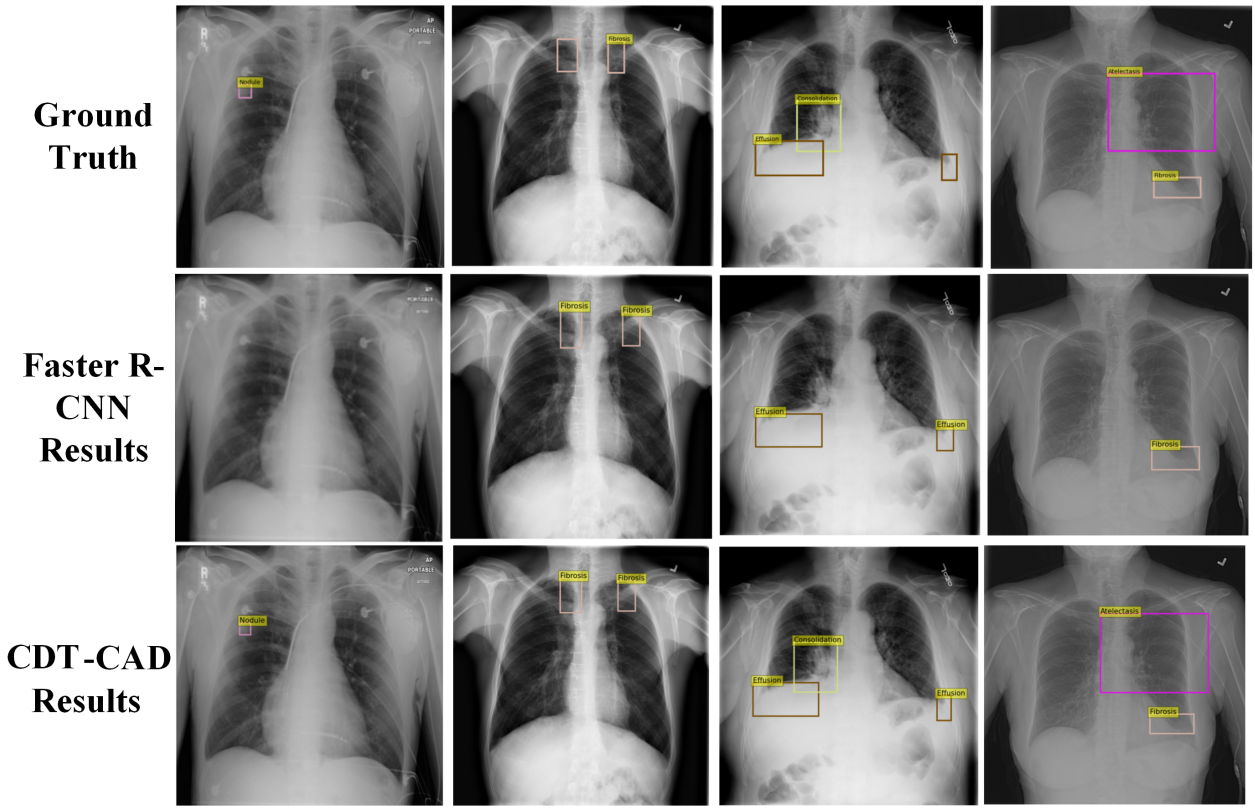


Fig. 7. Comparisons between Ground truth and detection results achieved by the proposed CDT-CAD and Faster R-CNN.

[17] Y. Cho, Y.-G. Kim, S. M. Lee, J. B. Seo, and N. Kim, "Reproducibility of abnormality detection on chest radiographs using convolutional neural network in paired radiographs obtained within a short-term interval," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[18] X. Ouyang, S. Karanam, Z. Wu, T. Chen, J. Huo, X. S. Zhou, Q. Wang, and J. Cheng, "Learning hierarchical attention for weakly-supervised chest x-ray abnormality localization and diagnosis," *IEEE Trans. Medical Imaging*, vol. 40, no. 10, pp. 2698–2710, 2021.

[19] Y. Han, C. Chen, A. Tewfik, B. Glicksberg, Y. Ding, Y. Peng, and Z. Wang, "Knowledge-augmented contrastive learning for abnormality classification and localization in chest x-rays with radiomics using a feedback loop," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2022.

[20] H. Ji, H. Liu, Y. Li, J. Xie, N. He, Y. Huang, D. Wei, X. Chen, L. Shen, and Y. Zheng, "Point beyond class: A benchmark for weakly semi-supervised abnormality localization in chest x-rays," in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 249–260.

[21] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[23] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv:2004.10934*, 2020.

[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[26] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[27] A. K. Sangaiah, D. V. Medhane, T. Han, M. S. Hossain, and G. Muhammad, "Enforcing position-based confidentiality with machine learning paradigm through mobile edge computing in real-time industrial informatics," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4189–4196, 2019.

[28] J. Zhang, W. Feng, T. Yuan, J. Wang, and A. K. Sangaiah, "Scstcf: spatial-channel selection and temporal regularized correlation filters for visual tracking," *Applied Soft Computing*, vol. 118, p. 108485, 2022.

[29] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of IEEE International Conference on Computer Vision*, 2019, pp. 9627–9636.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[31] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp. 2978–2988.

[32] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, "Efficient content-based sparse attention with routing transformers," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 53–68, 2021.

[33] H. Wang, Z. Wu, Z. Liu, H. Cai, L. Zhu, C. Gan, and S. Han, "HAT: hardware-aware transformers for efficient natural language processing," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7675–7688.

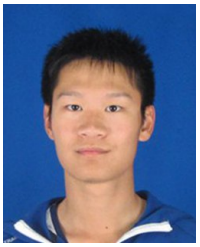
[34] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *CoRR*, vol. abs/1710.10903, 2017.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of 9th International Conference on Learning Representations*, 2021.

[36] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of International Conference on Machine Learning*, 2021, pp. 10 347–10 357.

[37] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of European Conference on Computer Vision*, 2020, pp. 213–229.
[38] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: deformable transformers for end-to-end object detection," in *Proceedings of International Conference on Learning Representations*, 2021.
[39] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 993–13 000.
[40] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 039–13 048.
[41] Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high quality object detection," in *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.
[42] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

Yirui Wu is currently an Associate Professor at Hohai University. Before coming to Hohai, he obtained his Ph.D. degree from Nanjing University in 2016. He received his B.S. Degree from Nanjing University in 2011 as well. His current research interests include computer vision and multimedia understanding.



Qiran Kong received a B.S. degree in computer science and technology from Hohai University, Nanjing, China, in 2019. He is currently working toward an M.S. degree in the college of computer and information, Hohai University. His current research interests include scene text detection, biometrical image analysis, and instance segmentation.

Lilai Zhang received a B.E. degree in computer science and technology from Taiyuan University of Technology, Taiyuan, China, in 2021. He is currently working toward a M.E. degree in the College of Computer and Information, Hohai University. His current research interests include Computer Vision and Artificial Intelligence.

Aniello Castiglione received the Ph.D. degree in Computer Science from the University of Salerno, Italy. He is currently with the Department of Science and Technology, University of Naples Parthenope, Italy. He authored over 200 papers in international journals (14 journal papers are published on IEEE / ACM Transactions and) and conferences. His current research interests include Information Forensics, Digital Forensics, Security and Privacy on distributed systems, Communication Networks, Applied Cryptography, and Sustainable Computing. Dr. Castiglione has served in the organization (mainly as the Program Chair and a TPC member) of more than 200 international conferences. He served as a Reviewer for approximately 100 international journals and the Managing Editor of two ISI-ranked international journals. He was a Guest Editor of around 20 special issues and served as an Editor on more than 10 Editorial Boards of international journals. One of his papers (published in the IEEE Transactions on Dependable and Secure Computing) was selected as the "Featured Article" in the "IEEE Cybersecurity Initiative" in 2014, while in 2018 another paper (published in the IEEE Cloud Computing Magazine) was selected as the "Featured Article" in the "IEEE Cloud Computing Initiative."

Michele Nappi received the Laurea degree (cum Laude) in computer science from the University of Salerno, Italy, in 1991, the M.Sc. degree in information and communication technology from I.I.A.S.S."E.R. Caianiello", Vietri sul Mare, Salerno, and the Ph.D. degree in applied mathematics and computer science from the University of Padova, Italy. He is currently a Full Professor of computer science with the University of Salerno. He is also a Team Leader of the Biometric and Image Processing Lab (BIPLAB). His research interests include multibiometric systems, pattern recognition, image processing, compression and indexing, multimedia databases, human–computer interaction, and VR/AR. He has coauthored more than 190 papers in international conference, peer review journals and book chapters in these fields. He was a member of IAPR. He received several international awards for scientific and research activities. He was the President of the Italian Chapter of the IEEE Biometrics Council, from 2015 to 2017.

Shaohua Wan received the Ph.D. degree from School of Computer, Wuhan University in 2010. Since then, he has been an associate professor with the School of Information and Safety Engineering, Zhongnan University of Economics and Law.From 2016 to 2017, he was a visiting professor at with the Department of Electrical and Computer Engineering, Technical University of Munich, Germany. His main research interests include deep learning for Internet of Things and edge computing. He is an author of over 150 peer-reviewed research papers and books, including over 40 IEEE/ACM Transactions papers such as TII, TITS, TOIT, TNSE, TMM, TCSS, TOMM, TETCI, PR, etc., and many top conference papers in the fields of Edge Intelligence.