Computers in Biology and Medicine 168 (2024) 107742

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

DualAttNet: Synergistic fusion of image-level and fine-grained disease attention for multi-label lesion detection in chest X-rays

Qing Xu [*], Wenting Duan

The School of Computer Science, University of Lincoln, Lincolnshire, LN6 7TS, United Kingdom

ARTICLE INFO

Keywords:
Multi-label lesion localisation
Pyramid feature refinement
Dual attention supervision
Computer-aided diagnosis

ABSTRACT

Chest radiographs are the most commonly performed radiological examinations for lesion detection. Recent advances in deep learning have led to encouraging results in various thoracic disease detection tasks. Particularly, the architecture with feature pyramid network performs the ability to recognise targets with different sizes. However, such networks are difficult to focus on lesion regions in chest X-rays due to their high resemblance in vision. In this paper, we propose a dual attention supervised module for multi-label lesion detection in chest radiographs, named DualAttNet. It efficiently fuses global and local lesion classification information based on an image-level attention block and a fine-grained disease attention algorithm. A binary cross entropy loss function is used to calculate the difference between the attention map and ground truth at image level. The generated gradient flow is leveraged to refine pyramid representations and highlight lesion-related features. We evaluate the proposed model on VinDr-CXR, ChestX-ray8 and COVID-19 datasets. The experimental results show that DualAttNet surpasses baselines by 0.6% to 2.7% mAP and 1.4% to 4.7% AP_{50} with different detection architectures. The code for our work and more technical details can be found at https://github.com/xq141839/DualAttNet.

1. Introduction

Chest X-ray (CXR), as a cornerstone of X-ray exam, has become the most frequent radiological screening since the last century. Due to cost efficiency and low radiational doses, CXR is usually an essential approach for diagnosing different types of thoracic diseases [1], where images of posteroanterior and anteroposterior views are mainly used to identify the position of the disease [2]. With the development of computing devices, constructing models used for automatic CXR analysis is no longer impossible. Various algorithms have been introduced, which can optimise monotonous tasks, improve the sensitivity for rare cases and assist with long-range diagnosis [3–5]. Multi-label abnormality detection plays a crucial role in thoracic disease diagnosis. Due to the high visual similarity of CXRs and the complexity of interpretation, it can be also known as the fine-grained image recognition task. Existing models, combined with Deep Neural Networks (DNNs) and various learning strategies, have made a significant breakthrough in automatic localisation of thoracic diseases [6–11]. The architecture of these methods mainly combines deep-layer encoders [12,13] and Feature Pyramid Network (FPN) [14] to extract features from CXRs. The main idea of FPN is to adopt multi-scale fusion to improve the performance of identifying targets with different sizes. The final layers of these networks involve two sub-networks consisting of a header for regressing the bounding box and predicting classification. These proposed algorithms are also called one-stage detection networks, which usually perform faster inference but lower precision compared to the two-stage models. To improve the performance of the one-stage models, embedding an attention mechanism module is a common solution. Attention algorithms work in a way to guide networks to focus on lesion zones by applying different non-linear transformations and combinations on the feature map. However, the additional computation costs reduce the inference speed of the network and practical applicability. Also, CXR images may include different disease types. Existing methods lack the supervision of classification and have the risk of false positive detections, which can be a serious issue for practical diagnostic applications.

To address this issue, we propose a dual attention module for multi-label lesion detection in chest X-rays, called DualAttNet, involving two components. The multi-scale feature maps extracted from FPN will be first fed into an Image-level Attention (ILA) block that adaptively recalibrates the weight of each channel. In this case, ILA is able to capture global classification information from the whole feature map but lacks supervision for the location of different disease types. For this reason, we design another fine-grained disease attention (FGDA) algorithm to guide the network to pay attention to the area of interest. It is connected to the header of the detector, utilises predicted

[*] Corresponding author.
E-mail address: xq14183925@gmail.com (Q. Xu).

https://doi.org/10.1016/j.compbiomed.2023.107742
Received 2 July 2023; Received in revised form 14 November 2023; Accepted 20 November 2023
Available online 22 November 2023

anchors to enhance spatial representations, and combines with global attention features from ILA. Finally, the fusion feature map, including classification and location information, will be compared to the lesion-level one-hot labels using binary cross entropy loss function so that the backward gradients can help suppress feature activation in irrelevant regions. The main contributions of this work can be summarised as follows:

(1) An ILA model is presented to receive global classification information from the features in FPN layers. Additionally, we introduce an FGDA algorithm to represent local attention information of features from the header of the detection architecture.
(2) We use FGDA to improve the representations from ILA and integrate a new DualAttNet. The outputs will compare with the ground truth from the image level. As the proposed model is not in charge of generating refined feature maps, embedding the module will not decrease the inference speed of the detector.
(3) Our experiments are conducted with three different CXR datasets: VinDr-CXR [15], ChestX-ray8 [16], and COVID-19 [17]. Evaluation results demonstrate that our proposed DualAttNet performs better than other attention models in terms of standard detection metrics — average precision (AP). It can be a new state-of-the-art (SOTA) method for multi-label lesion detection in CXRs.

2. Related work

2.1. Feature pyramid network for object detection

Feature Pyramid Network (FPN) [14] has been widely used in one-stage and two-stage object detection algorithms. To fully utilise the semantic information from each down-sampling layer, it constructs a top-down architecture with horizontal connections, which combines the feature maps at different scales. Many studies have demonstrated that FPN can better recognise objects with various sizes [18]. Existing detectors usually leverage FPN to fuse multi-scale features. RetinaNet [19], proposed by Lin et al. simply uses multi-layer Convolutional Neural Networks (CNNs) to transform every pyramid feature map to potential bounding boxes and corresponding classes of targets. To address the issue of imbalanced datasets, RetinaNet is trained with the focal loss function that adjusts the weight between easy and hard samples. Redmon et al. [20] presented a one-stage object detection model, called You Only Look Once (YOLO). The model first divides the input image into an S × S (e.g., 7 × 7) grid. The grid cell containing the centre of the labelled bounding box is responsible for detecting the object. The experimental results show that YOLO surpasses the baseline and achieves faster inference. However, YOLO has a limited ability to detect adjacent objects because only one object can be recognised by each grid. To improve the performance of YOLO, various updated versions, replacing the backbone or header and embedding FPN, have been published [21–23]. Furthermore, Zhang et al. [24] dedicated to the improvement of object localisation and introduced a VFNet for dense object detection. The model corrects the bounding box feature representation from FPN and is trained with a new varifocal loss function, which refines the predicted bounding boxes using an IoU-aware classification score. Extensive experiments demonstrate the VFNet model outperforms previous SOTA methods. However, these architectures mainly focus on the optimisation of localising targets. In contrast, our method improves the sensitivity for thoracic disease detection. In other words, it suppresses the detection of irrelevant disease regions and generates a low false positive rate, which is beneficial for clinical applications.

2.2. Attention mechanism

In the last few years, the attention mechanism has achieved significant success in computer vision. It aims to emphasise target regions while suppressing irrelevant features in an image. Hu et al. [25] proposed a Squeeze-and-Excitation Network (SENet) that captures channel-wise information and re-weights feature representations. This module can be flexibly embedded into any feature-extract layer to guide what networks should pay attention to. Following SENet, different variants [26–28] have been designed to further handle channel features efficiently. Spatial attention is another mechanism used for adaptive spatial area selection. Especially, Vision Transformer (ViT), proposed by Dosovitskiy et al. [29], has received the most attention recently. Multi-head attention module is the main idea of ViT, which calculates dynamic spatial–temporal correlations from an input sequence. Inspired by ViT, various transformer-based frameworks [30–32] have been constructed and show better performance than CNNs. However, such models usually include a large number of parameters and require a high cost to configure a trainable environment. In addition, Woo et al. [33] presented the convolutional block attention module (CBAM) that combines two different dimensional information: channel and spatial. It activates valuable channels as well as highlights informative local regions. For lesion detection in medical imaging, attention mechanism has been widely used to highlight disease regions. Zihao et al. [34] proposed a multi-view FPN with position-aware attention (MVP-Net) for universal disease detection in CT scans. A CBAM-like architecture is used to re-calibrate the weight of fused pyramid feature maps and optimise the proposals of potential lesion regions. Hao et al. [35] designed a slice attention transformer (SATr) block that can be embedded into any convolutional encoder to generate a hypernetwork structure. Benefiting from a novel cascaded self-attention module, SATr performs better in the universal lesion detection task. Wong et al. [36] introduced a multi-scale attention network for automatic pneumonia detection. The key idea is to refine feature multi-level features via spatial attention networks. As a result, the final concatenated feature map has the ability to represent correct disease regions. However, existing attention modules receive features from the encoder and directly produce optimised feature maps, which are integrated into the whole model. On the contrary, our approach leverages both global and local gradient flows to refine the feature representation from FPN.

3. Methodology

In this section, we describe the components of the proposed DualAttNet. They connect with each feature pyramid layer and classification head of the detection architecture. The outputs integrate global and lesion-level local attention gradient flows that are used for guiding the detector to focus the disease regions in chest X-rays.

3.1. Image-level attention block

For the detection architecture in CXRs, the feature map of each layer in FPN contains thoracic disease objects at different scales. To distinguish various lesion types efficiently, we propose an image-level attention (ILA) block. The architecture of ILA is provided in Fig. 1. Firstly, a multi-scale feature map X_k obtained by FPN is considered as input. It will go through two different branches. On one branch (i.e. lower branch shown in Fig. 1), we use 1 × 1 convolution, followed by a BatchNorm and ReLU activation, to compress the channel-wise information, which can be represented as:

$$F_k = ReLU(BN(Conv(X_k))) \tag{1}$$

Where the channel of the output feature F_k is equal to the number of disease classes. The transformed multi-level pyramid feature involves stacked heatmaps for all lesion types. Each heatmap has the ability to represent the intensity of one disease in the image level. On the other branch, a global average pooling (GAP) is used to capture global spatial

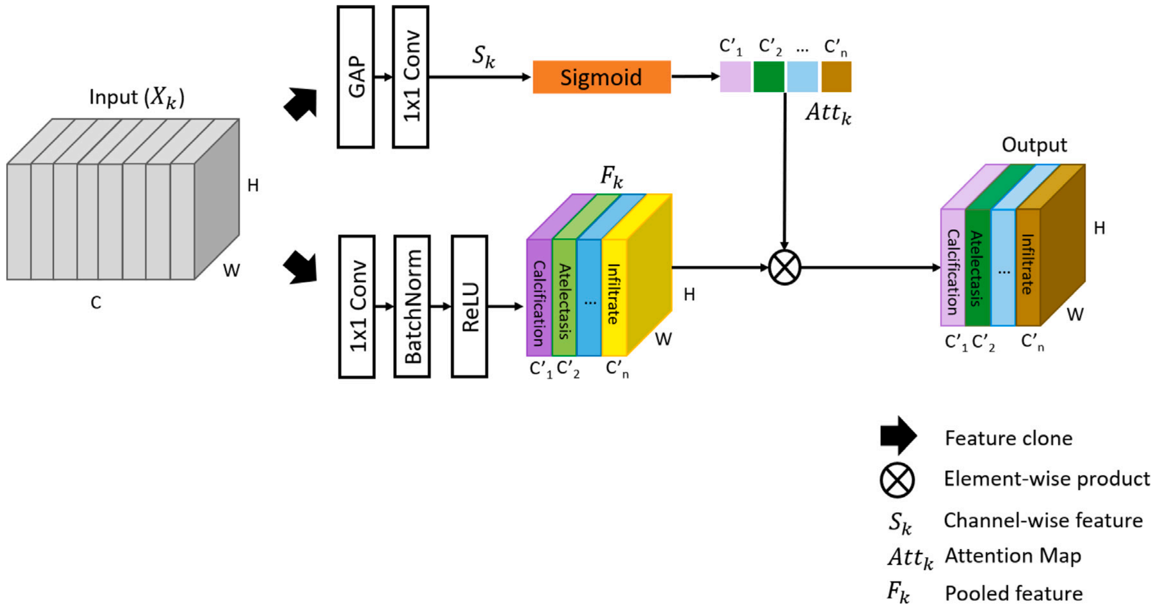


Fig. 1. The structure of image-level attention block.

information from the feature map X_k. The cth channel of the pooled feature S_c^k is calculated as:

$$S_c^k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c^k(i,j) \tag{2}$$

Where $X_c^k(i,j)$ stands for each feature value associated with the cth channel of input feature maps. H and W indicate their scales. The GAP compresses spatial dimensions and generates channel-wise statistics. We also connect 1×1 convolution with GAP for channel alignment with F_k. A soft attention is used across channels to adaptively select different spatial scales, which is guided by S_k. A soft assignment weight is designed by:

$$Att_k = \sigma(S_c^k) = \frac{1}{1 + e^{(-S_c^k)}} \tag{3}$$

Here we apply the sigmoid to re-calibrate the weight of the feature channel (i.e. different lesions). For the softmax function, it aims to improve the probability of the correct category while suppressing the remaining categories. Instead, there may be more than one lesion in CXRs. Therefore, we do not consider the softmax function in our module. Finally, an operation of element multiplication $\odot$ is performed on the refined weight Att_k and the corresponding feature map F_k:

$$Y_k' = F^k \odot Att_k, \quad k \in 1,2,3 \cdots N \tag{4}$$

Where k is the number of feature maps extracted from FPN. In sum, the ILA module aggregates global contextual information for different thoracic diseases.

3.2. Fine-grained disease attention

Anchor-based networks [37,38] have achieved great success in object detection. Anchors can be known as a set of predefined bounding boxes with different sizes and aspect ratios. To select a suitable anchor for detecting objects, we compute the intersection over union (IoU) between each anchor and targets. The head of detection architectures usually adopts fully convolutional networks to classify and regress the number of bounding boxes based on the anchor configuration. As the output feature map involves position and class information of disease, it can provide a local attention region for the specific disease in CXRs. Therefore, we construct a fine-grained disease attention (FGDA) algorithm, which is shown in Fig. 2. For each disease type n, it first extracts the anchor information (Height × Width × Anchor) from the classification feature map X_k'. Where the anchor with the maximum prediction is used as the attention map, which can be represented as:

$$A = \arg\max(X_k'), \quad A \in R^{H \times W \times Anchor} \tag{5}$$

The intensity of the attention map may have a large range of fluctuations due to the backpropagation during the training phase. To make the model more stable and achieve a smooth convergence, a simple normalisation method is performed:

$$A_{i,j}^n = \frac{A_{i,j}^n}{\sum_{\alpha=1}^{H} \sum_{\beta=1}^{W} A_{\alpha,\beta}^n} \tag{6}$$

Finally, we concatenate the attention map of all classes:

$$Y_k'' = Cat([A^1, A^2, \ldots, A^n]) \tag{7}$$

As illustrated by the above analysis, our proposed FGDA module can extract the location information of different lesions from CXRs.

3.3. DualAttNet architecture

Modern object detection architectures often focus on the enhancement of location precision. For lesion recognition in CXRs, false positive detection is a major issue in clinical applications. In this case, we embed the newly designed ILA and FGDA attention modules into a so-called DualAttNet to guide the network focusing on the symptomatic areas while suppressing unrelated detection. An overview of the proposed architecture is presented in Fig. 3. Particularly, we select RetinaNet [19] as our main detector. In the training phase, the input image first goes through an encoder and FPN. The feature map of the FPN layers will first be fed to the ILA module to extract global attention information. Then local attention is obtained from the FGDA module. These attention maps provide two different regions of interests (ROIs). Then, we combine the global attention map with local information and leverage GAP to transform class-aligned heatmaps into vectors. The final prediction fusing every pyramid layer can be computed as:

$$\hat{Y} = \sigma\left(\frac{1}{K \times H \times W} \sum_{k=1}^{K} \sum_{i=1}^{H} \sum_{j=1}^{W} Y_k'(i,j) + Y_k''(i,j)\right), \quad \hat{Y} \in R^{Batch \times Class} \tag{8}$$

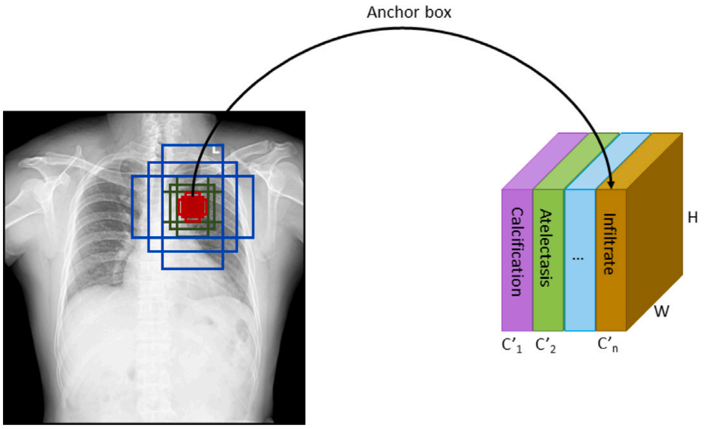


Fig. 2. The illustration of fine-grained disease attention algorithm.

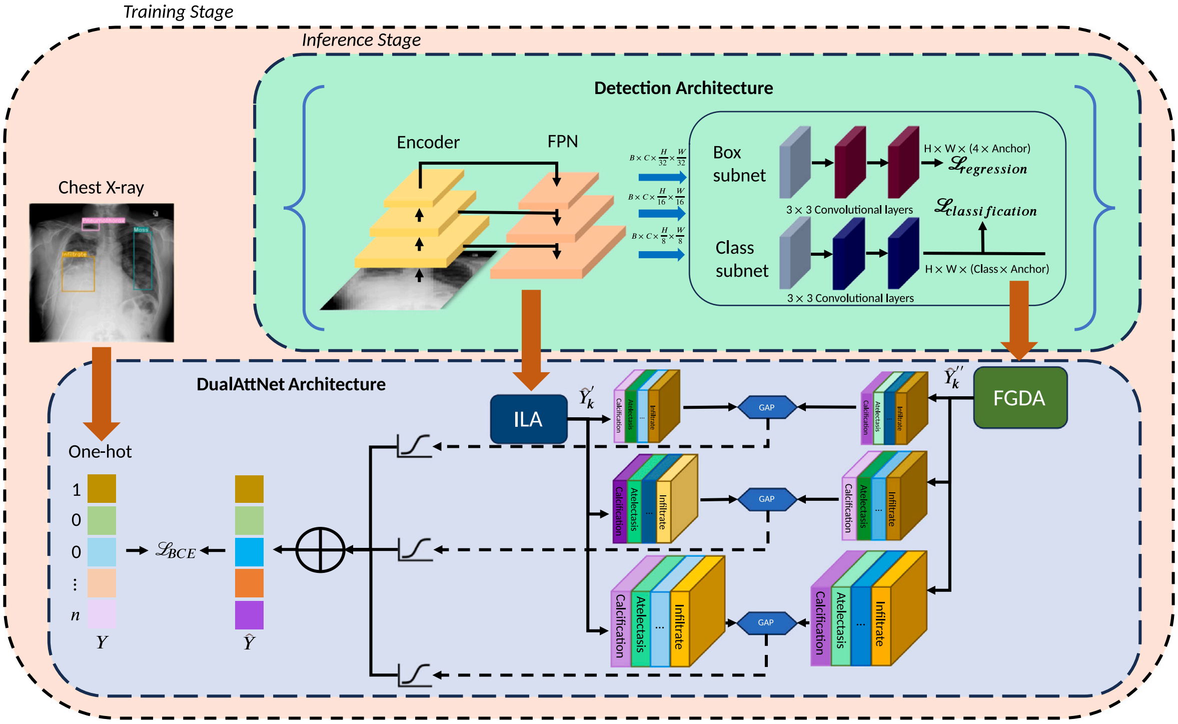


Fig. 3. The overview of our DualAttNet architecture connected with RetinaNet [19].

Where σ is the sigmoid function and k is the index of FPN layers. This operation also combines multi-scale feature maps in each pyramid layer, which can perceive different sizes of lesions in medical imaging. The cross entropy loss connected with a softmax function is a common algorithm for the classification task. However, the result of softmax usually highlights the probability of one category while suppressing the remaining types, which cannot handle multi-label scenarios. To refine the inclusion of multiple symptomatic lesions being represented in the pyramid feature, we adopt the binary cross entropy loss as the objective function. It can be defined as:

$$\mathcal{L}_{\mathrm{BCE}}(\hat{Y}, Y) = -\sum_{n=1}^{N} \left[Y_n \cdot \log\left(\hat{Y}_n\right) + \left(1 - Y_n\right) \cdot \log\left(1 - \hat{Y}_n\right) \right] \tag{9}$$

Where Y_n is the lesion-level annotations and $\hat{Y}_n$ is the prediction. As a result, the gradient from both global and local branches can optimise the feature representation of FPN for lesion classification in CXRs. In addition, the baseline model (i.e. RetinaNet) still keeps its original two sub-networks to perform bounding box and object classification predictions. The final loss function can be organised as:

$$\mathcal{L}_{\mathrm{loss}} = \mathcal{L}_{\mathrm{regression}} + \mathcal{L}_{\mathrm{classification}} + \mathcal{L}_{\mathrm{BCE}} \tag{10}$$

In the inference stage, the detection architecture can be directly decoupled from DualAttNet as no gradient calculation is required. Overall, the proposed model can adapt to any FPN-based detection framework without affecting the inference speed of the original detection architecture.

4. Experiments and results

4.1. Dataset

To demonstrate the effectiveness of DualAttNet, we evaluate it on three public datasets of multi-label thoracic disease detection:

- VinDr-CXR [15] comprises 4394 Digital Imaging and Communications in Medicine (DICOM) files with 14 types of thoracic abnormalities from chest radiographs. It is also the training database

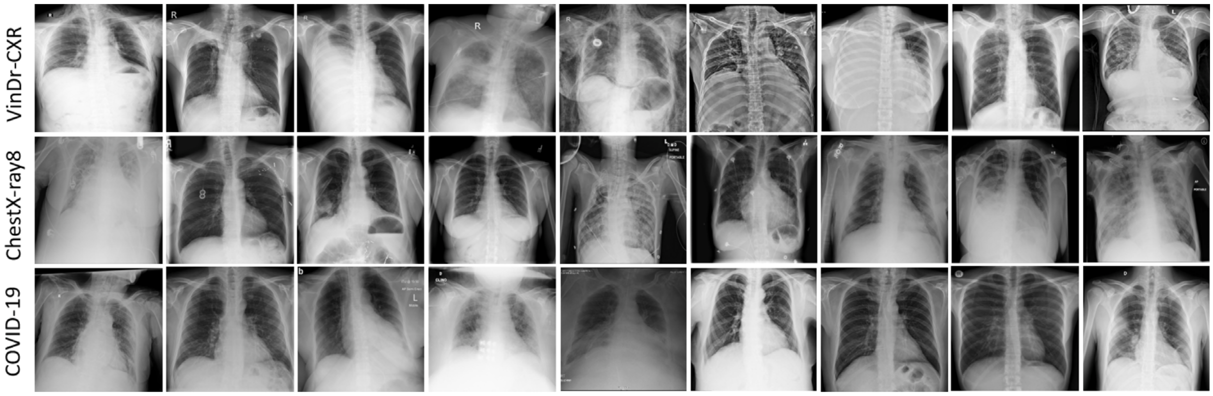


Fig. 4. Examples of chest radiographs under various circumstances in three datasets.

Table 1
Details of the chest X-ray datasets used in our experiments.

Dataset	Images	Input size	Train	Valid	Test
VinDr-CXR	4394	Variable	3074	880	440
ChestX-ray8	880	1024 × 1024	616	176	88
COVID-19	223	Variable	154	46	23

for the Kaggle 2021 VinBigData Chest X-ray Abnormalities Detection Competition. All of the samples are collected from the Hospital 108 as well as the Hanoi Medical University Hospital, and manually annotated by a total of 17 experienced radiologists.

- The second dataset used in this study is ChestX-ray8 [16]. It includes 108,948 posteroanterior X-ray images of 32,717 unique patients with eight disease labels. Particularly, 880 images have annotated objects with bounding boxes, which can be used as the ground truth to evaluate the disease localisation performance. To guarantee high confidence for each labelled disease, images and their disease keywords were provided to board-certified radiologists. They can only identify the disease instance in the image that corresponds to the given keyword.
- In order to evaluate the robustness of the proposed architecture on a small dataset, we add a COVID-19 dataset [17] to our experiment. There are only 223 frontal X-ray images for the detection of five pneumonia types: COVID-19, SARS, Streptococcus, Pneumocystis, and ARDS. These images are extracted from different hospitals around the world (e.g. China, Germany, Vietnam) in order to illustrate the diversity of radiographs.

Furthermore, Fig. 4 presents some chest X-rays converted from DICOM files. A fixed random seed is used to divide all datasets into three sets: train, validation, and test, in the ratio of 7:2:1. More details about the data split are provided in Table 1.

4.2. Evaluation metrics

In the field of object detection, average precision (AP) is the standard evaluation metric. Specifically, we calculate the mean AP (mAP) from AP_{50} to AP_{95} with an interval of 5, following the COCO evaluator. Recall plays an important role in clinical applications [39]. Therefore, we also compute average recall (AR) for small (AR_S), medium (AR_M), and large (AR_L) lesions. They are respectively defined as being between 0^2 to 32^2, 32^2 to 96^2, and 96^2 to $1e5^2$ pixels in area. Moreover, floating point operations (FLOPs) is calculated to investigate the complexity of each module. All reported statistics were averaged over the number of abnormalities in the dataset.

4.3. Data augmentation

As illustrated in Section 4.1, medical datasets contain a limited number of samples. The model is easily prone to overfitting in the training stage. To mitigate this issue, we adopt data augmentation approaches to enrich the diversity of images and improve the robustness of the model. In the experiment, horizontal flip, rotation, brightness, contrast, and cutout transformations are randomly applied to the training set of each dataset with a probability of 0.1.

4.4. Implementation details

All experiments are implemented using PyTorch 1.10.0 framework on a single NVIDIA RTX 2080Ti GPU, 4-core CPU, and 28 GB RAM. In order to compare DualAttNet with other attention modules, we use a commonly FPN-based detector, RetinaNet [19] as the fundamental architecture, a pretrained ResNet-50 [12] as the encoder, focal loss as the main objective function and Adam as the optimiser with a learning rate of 1e-5. The number of batch sizes and epochs are set to 4 and 40 respectively. In the training phase, the images from all three datasets are resized to 512 × 512. We also apply ReduceLROnPlateau to adjust the learning rate, where the hyperparameters: patience and factor are set to 3 and 0.1 respectively. All experiments on three datasets are conducted on the same training, validation, and testing datasets.

4.5. Results

In this section, we show quantitative results on three different CXR datasets and compare our proposed model with other SOTA methods.

4.5.1. Comparison on VinDr-CXR dataset

The quantitative result on VinDr-CXR dataset is presented in Table 2. For lesion detection in chest X-rays, a series of AP metrics are able to reveal the precision of the location. From the Table, DualAttNet shows better performance than other SOTA models in most of the metrics. Specifically, our module achieves an mAP of 0.116 and an AP_{50} of 0.241, which outperforms PSA by 1.3% in terms of mAP and 2.0% in AP_{50}. Also, recall is an essential metric in clinical applications. In the experimental result, DualAttNet also displays higher AR_M and AR_L scores than other methods.

4.5.2. Comparison on ChestX-ray8 dataset

The annotation in medical imaging is usually expensive and time-consuming. Therefore, many CXR datasets contain limited box annotations, which is a challenge for generalisation of the model. For lesion detection in chest X-rays, the performance of the architecture on mAP and AP_{50} are the most witnessed metrics. A comparison between each model is provided in Table 3. We are not able to calculate AP_S and AR_S metrics because there are no small targets based on COCO criteria. The

Table 2
Results on the VinDr-CXR dataset.

Methods	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L	FLOPs
SENet [25]	0.103	0.219	0.104	0.004	0.085	0.114	0.023	0.168	0.247	53.81
CBAM [33]	0.098	0.214	0.096	**0.010**	0.087	0.092	**0.032**	0.170	0.225	53.82
PSA [40]	0.103	0.221	0.099	0.005	0.086	0.116	0.025	0.167	0.230	54.04
CCNet [41]	0.101	0.217	0.102	0.002	0.084	0.107	0.025	0.169	0.242	54.23
ACmix [42]	0.099	0.215	0.099	0.005	0.082	0.107	0.029	0.171	0.236	55.00
CoT [43]	0.104	0.218	0.102	0.003	0.090	0.112	0.025	0.167	0.240	56.89
ParNet [44]	0.102	0.221	0.105	0.006	0.085	0.117	0.024	0.163	0.233	57.35
DualAttNet (Ours)	**0.116**	**0.241**	**0.114**	0.005	**0.098**	**0.121**	0.027	**0.177**	**0.257**	**53.79**

Table 3
Results on the ChestX-ray8 dataset.

Methods	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L	FLOPs
SENet [25]	0.054	0.124	0.036	–	0.002	0.065	–	0.034	0.140	53.81
CBAM [33]	0.057	0.126	0.050	–	0.008	0.074	–	0.057	0.142	53.82
PSA [40]	0.061	0.123	0.050	–	0.019	0.076	–	0.043	0.152	54.04
CCNet [41]	0.063	0.132	0.077	–	0.006	0.074	–	0.039	0.154	54.23
ACmix [42]	0.053	0.104	0.063	–	0.008	0.061	–	**0.069**	0.142	55.00
CoT [43]	0.062	0.133	0.042	–	0.016	0.075	–	0.028	0.160	56.89
ParNet [44]	0.061	0.108	**0.078**	–	0.001	0.073	–	0.020	0.146	57.35
DualAttNet (Ours)	**0.071**	**0.145**	0.076	–	**0.026**	**0.076**	–	0.056	**0.161**	**53.79**

Table 4
Results on the COVID-19 dataset.

Methods	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L	FLOPs
SENet [25]	0.015	0.049	0.010	–	–	0.015	–	–	0.040	53.81
CBAM [33]	0.017	0.041	0.005	–	–	0.017	–	–	0.042	53.82
PSA [40]	0.014	0.041	0.003	–	–	0.014	–	–	0.036	54.04
CCNet [41]	0.013	0.050	0.000	–	–	0.013	–	–	0.037	54.23
ACmix [42]	0.013	0.032	0.002	–	–	0.013	–	–	0.045	55.00
CoT [43]	0.016	0.045	0.003	–	–	0.016	–	–	0.037	56.89
ParNet [44]	0.016	0.051	**0.014**	–	–	0.018	–	–	0.047	57.35
DualAttNet (Ours)	**0.021**	**0.064**	0.010	–	–	**0.019**	–	–	**0.053**	**53.79**

result demonstrates that DualAttNet has an increase of 0.8% over CCNet in mAP and 1.2% in AP_{50}. Particularly, our proposed model presents a significant enhancement over the recent self-attention architecture, where the mAP of DualAttNet is 2.8% higher than ACmix, and the AP_{50} of DualAttNet is 4.1% higher than this model.

4.5.3. Comparison on COVID-19 dataset

COVID-19 has tremendously impacted patients and the medical system globally. Current deep learning-based studies are expected to distinguish COVID-19 from other types of pneumonia automatically. To meet the requirement, we evaluate all models on the COVID-19 dataset. As existing annotations are based on the whole area of the lung, the dataset does not include any small or medium objects, and related metrics are not computed in the experiment (i.e. hence denoted as — in Table 2, 3, 4, 5). The quantitative results are provided in Table 4. We can observe that DualAttNet performs an mAP of 0.021 with a rise of 0.4% over CBAM and 1.3% in AP_{50} compared to the ParNet model.

4.6. Ablation study

4.6.1. Efficiency of ILA block

The DualAttNet model adopts the ILA block to capture global spatial information of different thoracic diseases at image level, which can refine multi-scale feature maps from FPN layers. The effectiveness of ILA block can be evaluated by comparing the configurations: RetinaNet and RetinaNet + ILA in Table 5. In terms of the mAP and AP_{50}, the ILA block respectively produces an improvement of 0.7% and 0.9% on the VinDr-CXR dataset, 1.7% and 1.8% increase on the ChestX-ray8 dataset, as well as 0.7% and 1.6% enhancement on the COVID-19 dataset. Thus, it can be concluded that the ILA block enhances the performance of the original RetinaNet.

4.6.2. Significance of FGDA algorithm

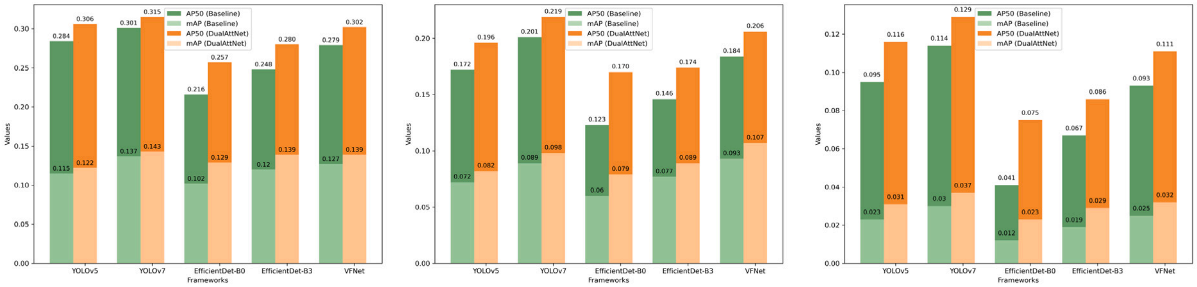
The FGDA algorithm is an essential part of the proposed DualAttNet model. It extracts anchor information of each lesion from the classification sub-network and then transformed into local attention feature maps uses with a normalisation method. We compare the network configurations: RetinaNet and RetinaNet + FGDA to evaluate the effectiveness of the FGDA algorithm. From the mAP and AP_{50} metrics in Table 5, FGDA respectively shows an improvement of 1.1% and 1.3% on the VinDr-CXR dataset, 1.3% and 1.3% improvement on the ChestX-ray8 dataset, as well as 0.4% and 0.8% improvement on the COVID-19 dataset. We can argue that the FGDA-embedded architecture performs better than the RetinaNet model. Overall, ILA has a more significant impact than the FGDA algorithm. By taking advantage of both modules, the AligNet model (RetinaNet + ILA + FGDA) can further improve the mAP by 0.1% to 0.9% and the AP_{50} by 0.9% to 2.6% compared to the RetinaNet with a single ILA or FGDA module.

4.7. Applicability of DualAttNet

We also transfer DualAttNet to other popular detection frameworks. In detail, the ILA and FGDA modules are connected with their feature pyramid layers and classification header respectively. Fig. 5 displays the comparison results between original and DualAttNet-based detectors, including YOLOv5 [45], YOLOv7 [46], EfficientDet-B0 [47], EfficientDet-B3 [47] and VFNet [24]. It can be demonstrated that adding DualAttNet can further boost the mean mAP by around 1.2% and mean AP_{50} by around 2.5% on three CXR datasets. Moreover, vision transformer (ViT) [29] has achieved great success in the field of computer vision. In the field of disease detection, we construct a new ViT-DualAttDet by combining our DualAttNet with a transformer-based detector, ViTDet [48]. The comparison results between other SOTA one-stage architecture and ViT-DualAttDet are provided in Fig. 6.

Table 5
Detailed ablation study of the DualAttNet architecture.

Dataset	Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L
VinDr-CXR	RetinaNet [25]	0.104	0.219	0.097	0.003	0.091	0.119	0.022	0.155	0.219
	RetinaNet + ILA	0.111	0.228	0.103	**0.007**	0.093	0.122	0.030	0.175	0.242
	RetinaNet + FGDA	0.115	0.232	0.108	0.006	0.092	**0.124**	**0.032**	0.162	0.230
	RetinaNet + DualAttNet	**0.116**	**0.241**	**0.114**	0.005	**0.098**	0.121	0.027	**0.177**	**0.257**
ChestX-ray8	RetinaNet [25]	0.049	0.106	0.033	–	0.004	0.061	–	0.059	0.145
	RetinaNet + ILA	0.066	0.124	0.056	–	0.013	0.072	–	**0.066**	0.160
	RetinaNet + FGDA	0.062	0.119	0.061	–	0.006	0.074	–	0.024	0.154
	RetinaNet + DualAttNet	**0.071**	**0.145**	**0.076**	–	**0.026**	**0.076**	–	0.056	**0.161**
COVID-19	RetinaNet [25]	0.011	0.037	0.000	–	–	0.011	–	–	0.037
	RetinaNet + ILA	0.018	0.053	0.007	–	–	0.016	–	–	0.038
	RetinaNet + FGDA	0.015	0.045	0.009	–	–	0.013	–	–	0.042
	RetinaNet + DualAttNet	**0.019**	**0.064**	**0.010**	–	–	**0.019**	–	–	**0.053**

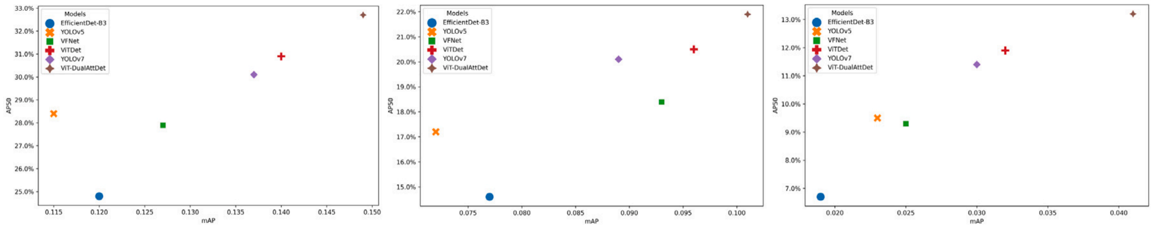


(a) VinDr-CXR

(b) ChestX-ray8

(c) COVID-19

Fig. 5. The performance comparison of baselines vs DualAttNet-embedded detection architectures on three CXR datasets.



(a) VinDr-CXR

(b) ChestX-ray8

(c) COVID-19

Fig. 6. The performance comparison of SOTA one-stage detectors vs ViT-DualAttDet on three CXR datasets.

Specifically, ViT-DualAttDet improves the mAP by 0.5% to 0.9% and the AP_{50} by 1.3% to 1.8% compared to ViTDet. As a result, ViT-DualAttDet can be considered as a new SOTA model for lesion detection.

5. Discussion

Multi-label lesion detection in chest X-rays has received increasing attention in the field of medical image analysis. Due to the high visual similarity of various thoracic diseases in medical imaging, attention mechanism has been widely inserted into general detection architectures to highlight different lesion regions in the feature map. For existing popular methods, such operations usually bring additional time costs in the inference phase. In contrast, DualAttNet focuses on current pyramid feature-based detectors and constructs an attention sub-network and leverage loss function and the gradient from global and local attention branches to support the refinement of features from pyramid layers. In this case, the attention module is no longer responsible for generating the optimised feature map and will not degrade the inference speed. To illustrate the effectiveness of our approach, on the one hand, we have conducted a series of quantitative comparison experiments following COCO evaluation standards. On the other hand, FROC [49] is a crucial metric in clinical applications [50]. For each

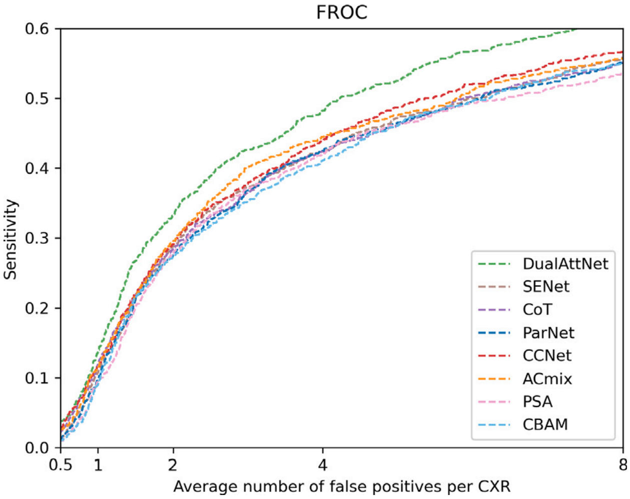


Fig. 7. The comparison of our DualAttNet and other SOTA attention models with FROC curve.

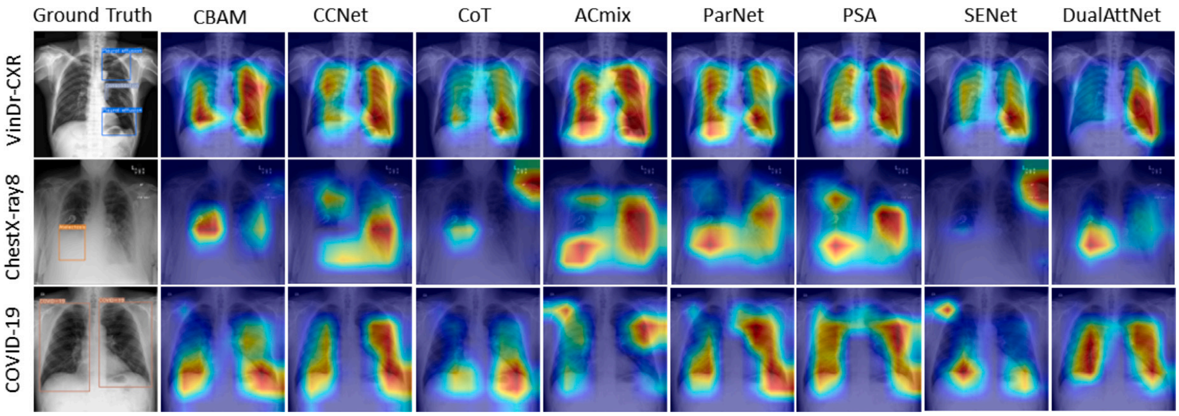


Fig. 8. The Visualisation of the attention regions in the method of DualAttNet vs other eight attention models.

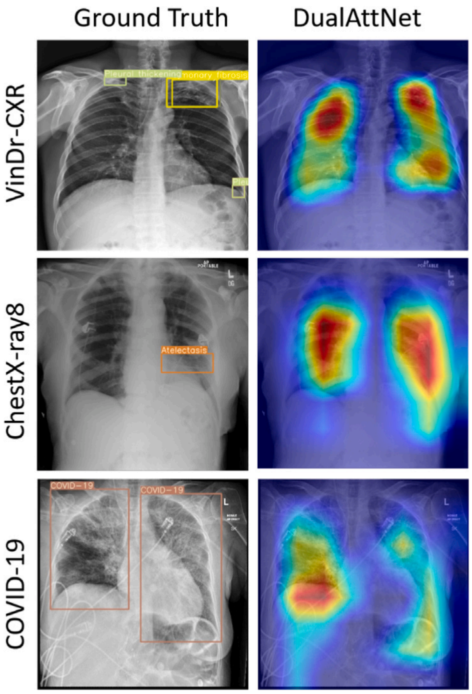


Fig. 9. The scenarios that DualAttNet fails to focus on disease regions in three CXR datasets.

CXR image, it illustrates the average sensitivity of models based on different number of false positives per scan. As ChestX-ray8 and COVID-19 datasets contain too few samples in the test set which lack statistical significance, we combine all three datasets in this evaluation. Following the algorithm in the Camelyon16 challenge [51], the result is presented in Fig. 7. It can be observed that from an average of one false positive per scan, the sensitivity of DualAttNet is significantly higher than other methods. Consequently, our proposed model performs a much lower false positive rate with the same sensitivity.

Furthermore, to further explain why the DualAttNet-based detection framework is more effective than other attention approaches in the multi-label lesion detection task, we use Eigen-CAM [52] to visualise the attention region of different models on three datasets, which is shown in Fig. 8. It calculates the principal components of the trained feature representations from the convolutional layers. Eigen-CAM has demonstrated robustness to misclassifications caused by fully connected layers within convolutional neural networks. Moreover, it operates independently of gradient backpropagation, class correlation scores, maximum activation positions, or any other kinds of weighted features. From the visualisation, we can see that the existing attention algorithms may not focus on correct disease areas. On the contrary, the proposed module is able to suppress other non-lesion regions and generate a more interpretable attention map. Thus, DualAttNet can reduce the detection of false positive cases and improve targeting accuracy. In addition, our proposed method still exists several limitations. Firstly, as illustrated on AP_S and AR_S metrics (Table 2), DualAttNet does not perform well in small object detection compared to other attention models. In Fig. 9, we visualise some Eigen-CAM results in which DualAttNet fails to localise disease regions due to the small-size lesions or the interference from medical devices (e.g. instrument detection heads as shown in the bottom image of Fig. 8). Secondly, although the model can be connected with most of the FPN-based detectors, some of the modern studies remove pyramid features in their architectures. Therefore, extending our method to such frameworks is also our future work. In summary, DualAttNet displays its robustness and superior performance on the multi-label lesion detection task and we believe it can be considered as a new SOTA method for computer-aid diagnosis in chest X-rays.

6. Conclusion

In this paper, we propose a gradient-based attention module for multi-label lesion detection in CXRs, called DualAttNet. The introduced module is comprised of an ILA block and an FGDA algorithm. The former collects global classification information from each pyramid feature layer. The latter focuses on the local attention extraction from anchor feature maps, which combines with the global attention from the ILA block, refines representations of pyramid features and guide the detector to pay attention to lesion regions using the gradient flow of binary cross entropy loss function. We evaluate our proposed method on three CXR datasets. The results demonstrate that compared to other attention models, DualAttNet not only achieves higher scores in standard mAP and AP_{50} metrics but also performs lower false positive rates with the same sensitivity. In the future, we will explore the application of DualAttNet in non-FPN detection architectures and other medical image analysis tasks, such as instance segmentation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Raoof, D. Feigin, A. Sung, S. Raoof, L. Irugulpati, E.C. Rosenow III, Interpretation of plain chest roentgenogram, Chest 141 (2) (2012) 545–558.

[2] V. Lai, W.K. Tsang, W.C. Chan, T.W. Yeung, Diagnostic accuracy of mediastinal width measurement on posteroanterior and anteroposterior chest radiographs in the depiction of acute nontraumatic thoracic aortic dissection, Emergency Radiol. 19 (2012) 309–315.

[3] Z. Yan, J. Zhang, S. Zhang, D.N. Metaxas, Automatic rapid segmentation of human lung from 2D chest X-ray images, in: Proc. of MICCAI Workshop on Sparsity Techniques in Medical Imaging, Citeseer, 2012.

[4] I. Solti, C.R. Cooke, F. Xia, M.M. Wurfel, Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches, in: 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop, IEEE, 2009, pp. 314–319.

[5] P. Yu, H. Xu, Y. Zhu, C. Yang, X. Sun, J. Zhao, An automatic computer-aided detection scheme for pneumoconiosis on digital chest radiographs, J. Digit. Imaging 24 (2011) 382–393.

[6] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017, arXiv preprint arXiv:1711.05225.

[7] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

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

[9] S. Akter, F.J.M. Shamrat, S. Chakraborty, A. Karim, S. Azam, COVID-19 detection using deep learning algorithm on chest X-ray images, Biology 10 (11) (2021) 1174.

[10] F.J.M. Shamrat, S. Azam, A. Karim, R. Islam, Z. Tasnim, P. Ghosh, F. De Boer, LungNet22: a fine-tuned model for multiclass classification and prediction of lung disease using X-ray images, J. Pers. Med. 12 (5) (2022) 680.

[11] F.J.M. Shamrat, S. Azam, A. Karim, K. Ahmed, F.M. Bui, F. De Boer, High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images, Comput. Biol. Med. 155 (2023) 106646.

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[13] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

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

[15] H.Q. Nguyen, K. Lam, L.T. Le, H.H. Pham, D.Q. Tran, D.B. Nguyen, D.D. Le, C.M. Pham, H.T. Tong, D.H. Dinh, et al., VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations, Sci. Data 9 (1) (2022) 429.

[16] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.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, 2017, pp. 2097–2106.

[17] J.P. Cohen, P. Morrison, L. Dao, COVID-19 image data collection, 2020, arXiv preprint arXiv:2003.11597.

[18] Z. Zou, K. Chen, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years: A survey, Proc. IEEE (2023).

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

[20] J. Redmon, S. Divvala, R. Girshick, 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.

[21] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, J. Sun, You only look one-level feature, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13039–13048.

[22] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: Exceeding yolo series in 2021, 2021, arXiv preprint arXiv:2107.08430.

[23] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al., YOLOv6: A single-stage object detection framework for industrial applications, 2022, arXiv preprint arXiv:2209.02976.

[24] H. Zhang, Y. Wang, F. Dayoub, N. Sunderhauf, Varifocalnet: An iou-aware dense object detector, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8514–8523.

[25] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[26] Z. Gao, J. Xie, Q. Wang, P. Li, Global second-order pooling convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3024–3033.

[27] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.

[28] Z. Qin, P. Zhang, F. Wu, X. Li, Fcanet: Frequency channel attention networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 783–792.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[30] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F.E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[33] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.

[34] Z. Li, S. Zhang, J. Zhang, K. Huang, Y. Wang, Y. Yu, MVP-Net: multi-view FPN with position-aware attention for deep universal lesion detection, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, Springer, 2019, pp. 13–21.

[35] H. Li, L. Chen, H. Han, S. Kevin Zhou, Satr: Slice attention with transformer for universal lesion detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 163–174.

[36] P.K. Wong, T. Yan, H. Wang, I.N. Chan, J. Wang, Y. Li, H. Ren, C.H. Wong, Automatic detection of multiple types of pneumonia: Open dataset and a multi-scale attention network, Biomed. Signal Process. Control 73 (2022) 103415.

[37] S. Zhang, L. Wen, X. Bian, Z. Lei, S.Z. Li, Single-shot refinement neural network for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4203–4212.

[38] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.

[39] S.M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, M.K. Khan, Medical image analysis using convolutional neural networks: a review, J. Med. Syst. 42 (2018) 1–13.

[40] H. Zhang, K. Zu, J. Lu, Y. Zou, D. Meng, EPSANet: An efficient pyramid squeeze attention block on convolutional neural network, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 1161–1177.

[41] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 603–612.

[42] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, G. Huang, On the integration of self-attention and convolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 815–825.

[43] Y. Li, T. Yao, Y. Pan, T. Mei, Contextual transformer networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2022).

[44] A. Goyal, A. Bochkovskiy, J. Deng, V. Koltun, Non-deep networks, Adv. Neural Inf. Process. Syst. 35 (2022) 6789–6801.

[45] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, Z. Yifu, C. Wong, D. Montes, et al., ultralytics/yolov5: V7. 0-YOLOv5 SOTA realtime instance segmentation, Zenodo (2022).

[46] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7464–7475.

[47] M. Tan, R. Pang, Q.V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.

[48] Y. Li, H. Mao, R. Girshick, K. He, Exploring plain vision transformer backbones for object detection, in: European Conference on Computer Vision, Springer, 2022, pp. 280–296.

[49] J.P. Egan, G.Z. Greenberg, A.I. Schulman, Operating characteristics, signal detectability, and the method of free response, J. Acoust. Soc. Am. 33 (8) (1961) 993–1007.

[50] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, S. Yu, A survey on incorporating domain knowledge into deep learning for medical image analysis, Med. Image Anal. 69 (2021) 101985.

[51] B.E. Bejnordi, M. Veta, P.J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J.A. Van Der Laak, M. Hermsen, Q.F. Manson, M. Balkenhol, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, JAMA 318 (22) (2017) 2199–2210.

[52] M.B. Muhammad, M. Yeasin, Eigen-cam: Class activation map using principal components, in: 2020 International Joint Conference on Neural Networks, IJCNN, IEEE, 2020, pp. 1–7.