IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 September 2024, accepted 8 October 2024, date of publication 16 October 2024, date of current version 1 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3482102

RESEARCH ARTICLE

YOLO-CXR: A Novel Detection Network for Locating Multiple Small Lesions in Chest X-Ray Images

SHENGNAN HAO[1], XINLEI LI[1], WEI PENG[1], ZHU FAN[1,2], ZHANLIN JI[3,4], (Member, IEEE), AND IVAN GANCHEV[4,5,6], (Senior Member, IEEE)

[1]College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063210, China
[2]Department of Respiratory Medicine, Affiliated Hospital, North China University of Science and Technology, Tangshan 063009, China
[3]College of Mathematics and Computer Science, Zhejiang A&F University, Hangzhou, Zhejiang 311300, China
[4]Telecommunications Research Centre (TRC), University of Limerick, Limerick, V94 T9PX Ireland
[5]Faculty of Mathematics and Informatics, University of Plovdiv "Paisii Hilendarski," 4000 Plovdiv, Bulgaria
[6]Institute of Mathematics and Informatics-Bulgarian Academy of Sciences, 1040 Sofia, Bulgaria

Corresponding authors: Zhu Fan (Fanzhu1333@163.com), Zhanlin Ji (zhanlin.ji@gmail.com), and Ivan Ganchev (ivan.ganchev@ul.ie)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0135700; in part by Bulgarian National Science Fund (BNSF) under Grant KP-06-IP-CHINA/1; and in part by the Telecommunications Research Centre (TRC), University of Limerick, Ireland.

ABSTRACT Chest X-ray is one of the most widely used methods for clinical diagnosis of chest diseases. In recent years, the development of deep learning technologies has driven progress in chest disease detection, but existing methods still face numerous challenges. Current research primarily focuses on detecting specific chest diseases. However, when chest X-ray images indicate multiple diseases, the diverse and complex characteristics of different disease types make it challenging to extract effective information. Additionally, the detection accuracy of small lesions remains low, which lessens the overall lesion detection rate. To address these issues, a novel network, named YOLO-CXR, is proposed in this paper for multiple disease detection, which is able to effectively locate multiple small lesions in chest X-ray images. Firstly, the proposed network enhances the YOLOv8s backbone by replacing the ordinary convolutional layers with RefConv layers to improve its feature extraction capabilities w.r.t. various diseases. Secondly, it utilizes a novel Efficient Channel and Local Attention (ECLA) mechanism to increase its sensitivity to the spatial location information of different lesions. Thirdly, to enhance its detection of small lesions, YOLO-CXR incorporates a dedicated small-lesion detection head and the Selective Feature Fusion (SFF) technique. Due to these improvements, the proposed network significantly enhances its detection of lesions at different scales and multiple small lesions in particular. Experiments conducted on the publicly available VinDr-CXR dataset demonstrate that YOLO-CXR achieves an *mAP@0.5* of 0.338, a *mAP@[0.5:0.95:0.05]*of 0.167, and *recall* of 0.365, outperforming all state-of-the-art networks considered.

INDEX TERMS X-ray detection, image processing, computer aided diagnosis, object detection.

I. INTRODUCTION

Since the beginning of the 21st century, chest diseases have become a major global health concern due to factors such as population aging and air pollution [1]. Chest diseases can stem from simple lesions, making the time and accuracy of diagnosis critical factors for radiologists. Earlier, and more accurate, diagnoses can significantly improve patient outcomes and save lives.

Currently, the early diagnosis of chest diseases primarily relies on medical imaging. Chest X-rays (CXR) offer a non-invasive means to detect lesions with minimal harm to the body, allowing for the localization of abnormalities and providing initial morphological characteristics to assist doctors in further diagnosis. Compared to computed tomography (CT) and magnetic resonance imaging (MRI), CXR

The associate editor coordinating the review of this manuscript and approving it for publication was R. K. Tripathy.

is more cost-effective, faster to process, and sufficiently sensitive to various pathologies [2]. Thus, it remains the primary method for early screening of chest diseases [3]. Given that the chest contains vital organs and is prone to a wide range of diseases, it is common for CXR images to indicate the presence of multiple health-related conditions, especially during a patient's initial visit when the specific type of disease may not yet be determined. Therefore, detecting multiple diseases in clinical settings is crucial. We believe that having a unified model to detect multiple diseases is beneficial, as it reduces complexity and preserves computational resources compared to using multiple separate models. This is particularly meaningful in busy emergency departments during initial patient evaluations. Previous single-disease detection models have shown limitations in clinical practice [4].

Annually, approximately 360 million medical images are acquired. Even slight improvements in the accuracy or speed of radiologists may have a significant impact on patient care [5]. However, due to the complex principles and structures involved in CXR imaging, professional radiologists require considerable time for careful examination. The development of computer-aided diagnosis (CAD) can effectively improve the accuracy of lesion screening and reduce the workload of radiologists [6]. This is particularly crucial in underdeveloped regions lacking sufficient medical professionals. For auxiliary diagnosis, the automated medical image analysis, based on deep learning, has garnered significant attention. The rapid advancement of deep learning technologies provides a foundation for improving diagnostic efficiency, especially with convolutional neural networks (CNNs) widely applied in medical imaging. CNNs utilize large datasets and computational resources to enhance their prediction performance, breaking through potential limitations [7]. In the past, CNNs were constrained by limited datasets, preventing them from learning complex features and limiting their applicability in diverse, real-world environments. Additionally, due to restricted computational resources, CNNs were slow to train, which in turn limited their scale and complexity, thereby constraining the potential improvement in their performance.

In recent years, with advancements in the field of computer vision, research on automatic diagnosis of diseases, based on CXR images, has intensified, and the accuracy of diagnosis has gradually improved. However, during the initial diagnostic assistance to physicians, the inability to localize disease regions necessitates radiologists to spend additional time confirming diagnostic results. This is unacceptable in environments where rapid diagnosis is required. Therefore, chest disease detection models have proven indispensable in assisting radiologists in identifying and diagnosing diseases based on CXR images [8]. Current studies have introduced object detection for localization of diseases in CXR images, but automatic chest disease localization is still a challenging task. Firstly, current research primarily focuses on detecting a single abnormality, such as lung nodules or pneumothorax. However, chest diseases are often comorbid, meaning that a single CXR image may indicate the presence of multiple diseases, which is not comprehensive enough for aiding early patient screening. Single-disease detection methods also struggle to capture the characteristics of multiple diseases. Secondly, lesion regions corresponding to different diseases vary significantly in size, shape, and texture, and, in addition, may overlap. Due to the limited receptive field of convolutional kernels, it is challenging to obtain sufficiently rich feature information, which may result in a network being less sensitive to the features of different diseases, thereby reducing its localization accuracy. Lastly, many subtle lesions that are difficult to detect could be present in the localized areas and, in addition, may be with low contrast. Existing methods lack multi-scale information and tend to introduce redundant information when using shallow-layer features, leading to the omission of such diseases. As shown in Figure 1, CXR images typically may indicate multiple disease types, with corresponding varying lesion sizes, overlapping regions, and small lesions.

To address these issues, a chest disease detection network, based on YOLOv8s [9], named YOLO-CXR, is proposed in this paper. Firstly, it uses an improved YOLOv8s backbone by replacing ordinary convolutional layers with RefConv layers, as to enhance its feature extraction capability. Secondly, a novel Efficient Channel and Local Attention (ECLA) mechanism is utilized by YOLO-CXR to effectively extract information at different scales and capture spatial location information of lesions from feature maps at various scales. Lastly, a special detection head is used to enhance the YOLO-CXR's detection capability w.r.t. small lesions by utilizing the Selective Feature Fusion (SFF) [10], thus improving its multi-scale feature representation and enhancing its detection capability w.r.t. various chest lesions. Conducted experiments demonstrate that the proposed YOLO-CXR network overall outperforms the state-of-the-art (SOTA) networks taking part in the experiments.

The main contributions of this paper could be summarized as follows:

- An improved YOLOv8s backbone is proposed for use by YOLO-CXR, by replacing the ordinary downsampling convolutional layers with RefConv layers, thus improving the lesion detection in CXR images without incurring additional inference costs to the proposed network.
- A novel ECLA mechanism is elaborated, which utilizes both channel and spatial information to enhance the extraction of critical features from feature maps, thereby improving the localization accuracy of the proposed network.
- An enhanced small-lesion detection is elaborated for improving the detection accuracy of the proposed network w.r.t. small lesions, by adding a small object detection head to YOLOv8s, which is more sensitive to small objects.

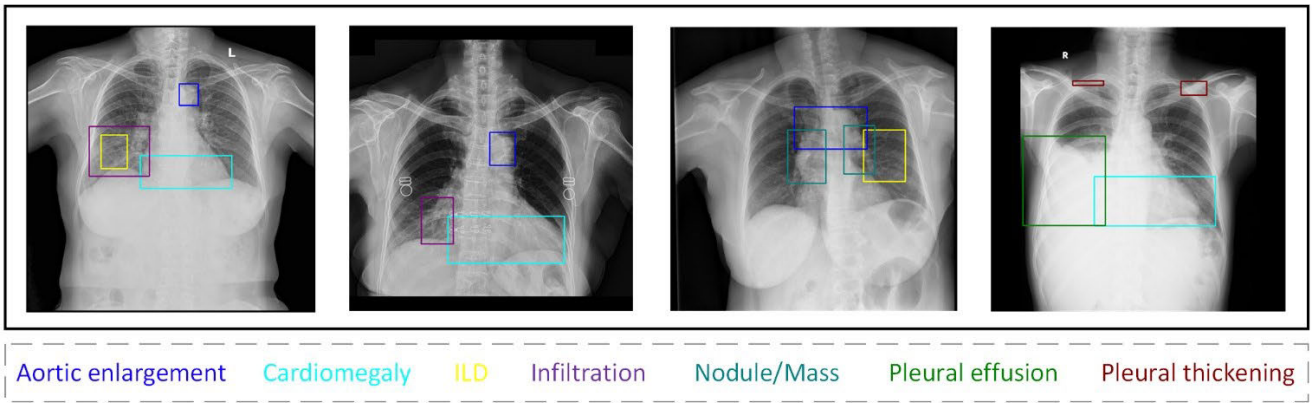


FIGURE 1. Sample chest X-ray images indicating multiple diseases, with corresponding varying lesion sizes, overlapping regions, and small lesions (the bounding boxes correspond to the disease types listed at the bottom, matched by color).

The rest of the paper is structured as follows. Section II reviews the related work in the field of chest disease detection. Section III presents the proposed YOLO-CXR network by focusing on the implementation process. Section IV introduces the dataset, evaluation metrics, and settings used in the experiments, and analyzes the obtained results. Section V discusses the key findings and limitations of the proposed network. Finally, Section VI concludes the paper.

II. RELATED WORK

A. DETECTION OF SINGLE DISEASE

Previous methods used for chest disease detection have primarily focused on detecting specific diseases. For instance, Harsono et al. [11] proposed a novel lung nodule detection and classification model, called I3DR-Net, based on dilated 3D CNNs [12], achieving significant improvements on the Moscow private dataset and LIDC public dataset. Li et al. [13] proposed a deep learning-based method for lung nodule detection, utilizing a patch-based multi-resolution convolutional network for feature extraction. This method achieved promising detection performance on the publicly available Japanese Society of Radiological Technology (JSRT) dataset. Xu et al. [14] employed a High-Resolution Network (HRNet) as a backbone of Cascade R-CNN [15] for lung nodule detection. Experiments, conducted on the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset, demonstrated that Cascade R-CNN improves the accuracy of detecting lung nodules, particularly w.r.t. small nodules. Mendoza et al. [16] proposed and analyzed a pipeline for detecting lung nodules in CXR images, including lung region segmentation, potential nodule localization, and candidate nodule classification. Their model was evaluated on all nodule-containing images in the JSRT dataset, and the experiments demonstrated its competitive results. Dahshan et al. [17] developed the RESCOVIDTC-Nnet model, integrating Empirical Wavelet Transform [18], ResNet-50 [19], and Temporal Convolutional Networks [20]. They employed artificial neural networks and support vector machines (SVM) for data classification, improving the efficiency of the COVID-19 diagnostic. Tolkachev et al. [21] combined U-Net [22] with various backbones for pneumothorax detection and introduced data augmentation and different preprocessing steps, achieving promising results compared to three experienced radiologists. Agrawal et al. [23] proposed an Attention-based Lightweight CNN (ALCNN), which utilized stacked convolutional layers with an attention mechanism to recalibrate the feature maps of channels. Additionally, these authors evaluated three different transfer learning methods for comparison with ALCNN. The experimental results indicated that their network performed well, and transfer learning did not significantly impact pneumothorax detection in CXR images. Bhatt et al. [24] developed a highly lightweight, deployable, and accurate model for pneumonia detection by utilizing three different models with varying kernel sizes. Their model achieved high accuracy, making it a deployable diagnostic aid solution. Hashmi et al. [25] proposed a pneumonia recognition model trained on CXR images, using data augmentation techniques to expand the training dataset. They also employed transfer learning during model training. Experiments on the Guangzhou Women and Children's Medical Center pneumonia dataset demonstrated that the proposed model could assist radiologists in clinical decision-making.

The presented methods demonstrate good detection performance for specific diseases, including COVID, pneumothorax, and pneumonia. However, in real clinical scenarios, CXR images are often used for early screening and comprehensive diagnosis, meaning that a single image may contain lesions of multiple, potentially unknown, diseases. The complexity of these images, with the coexistence of different types of diseases, makes single-disease detection inadequate for handling multiple occurring diseases, limiting its ability to maintain high accuracy.

B. DETECTION OF MULTIPLE DISEASES

Table 1 summarizes the main studies in this subarea w.r.t. covered disease types, and datasets and methods used. Guo et al. [27] proposed a method for diagnosing and localizing diseases in CXR images using CNNs, incorporating

TABLE 1. Summary of reviewed multiple-disease detection studies: datasets, count of disease types, and methods used.

Study	Dataset(s)	Count of Disease Types	Method(s) used
Guo et al. [27]	Shenzhen Hospital CXR & NIH CXR	7	Artificial bee colony algorithm and linear average-based ensemble method.
Fan et al. [28]	VinDr-CXR	14	Based on YOLOv5.
Bharati et al. [29]	NIH CXR	14	Combined neural networks, data augmentation, and spatial transformer networks with CNNs.
Lian et al. [30]	ChestX-Det & DR-private	13	Anatomical relation encoding and contextual clue aggregation.
Lin et al. [31]	VinDr-CXR	14	Utilizing a multi-convolution feature fusion block, tree-structured aggregation module, and scalable channel and spatial attention.
Yuan et al. [32]	VinDr-CXR	14	Utilizing a multi-scale lesion feature extraction network.
Nguyen et al. [33]	VinDr-CXR	14	Using ResNet-34 for extracting lung regions from chest X-ray images and YOLOv5 for detecting abnormalities within them.
Sheng et al. [34]	NIH CXR & VinDr-CXR	14	Using adjusted Barlow Twins and Faster R-CNN with FPN.
Xu and Duan [35]	VinDr-CXR, ChestX-ray8 & COVID-19	14, 8 & 5	Based on the proposed dual attention supervised module for multi-label lesion detection.
Lin et al. [36]	VinDr-CXR	14	Utilizing a combined RepVGG block and Resblock for information fusion extraction.
Ngo et al. [37]	VinDr-CXR	14	Data enhancement and optimization of the RetinaNet model using ResNet101 in the FPN backbone for optimal performance.

a deep learning network, fine-tuned with an artificial bee colony algorithm, and implementing a linear average-based ensemble method to detect pulmonary abnormalities. Experiments conducted on the Shenzhen Hospital CXR and NIH CXR datasets indicated that the method performed exceptionally well in detecting pulmonary abnormalities and diagnosing specific tuberculosis-related manifestations. Fan et al. [28] introduced an anomaly detection model using YOLOv5 [38] to detect 14 types of chest anomalies. The experimental results, obtained on the VinDr-CXR dataset, demonstrated that its performance surpasses other anomaly detection models. Bharati et al. [29] developed a new hybrid deep learning framework, named VDSNet, which combines data augmentation and spatial transformer networks with CNNs to detect lung diseases, based on images of the NIH CXR dataset. Although these methods have advanced chest disease detection, they primarily focus on extracting local information from images and insufficiently utilize global spatial information, thereby limiting their feature extraction capabilities. This limitation hampers their ability to capture the complex variations of different lesions.

Lian et al. [30] proposed a Structure-Aware Relation Network (SAR-Net) by leveraging constant structures and disease relationships extracted from domain knowledge, which significantly improved the performance of the extended Mask R-CNN [39]. Lin et al. [31] introduced a lesion detection method, based on a Scalable Attention Residual CNN (SAR-CNN) [40], designing three modules to address the issues of single resolution and weak inter-layer feature communication in CXR object recognition, greatly enhancing its efficiency. Yuan et al. [32] proposed a novel detection model that enhances multi-scale lesion feature extraction capability and improves the ability to simultaneously perceive multiple lesions during a single detection, achieving relatively good results on the VinDr-CXR dataset. Nguyen et al. [33] presented a new method for detecting anomalies in CXR images, using ResNet [19] to extract lung regions and surroundings from original CXR images, followed by applying YOLOv5 on the new images of the extracted lung regions. This method showed a slight improvement in detection performance compared to using the original images. Sheng et al. [34] used the adjusted Barlow Twins algorithm [41], introducing self-supervised learning into CXR anomaly localization to address cross-domain transfer learning differences. Experimental results, obtained on the NIH CXR and VinDr-CXR datasets, showed that the proposed algorithm significantly improves the anomaly localization. Xu and Duan [35] proposed a Dual Attention Supervision module for multi-label lesion detection in CXR images, named DualAttNet. Their experimental

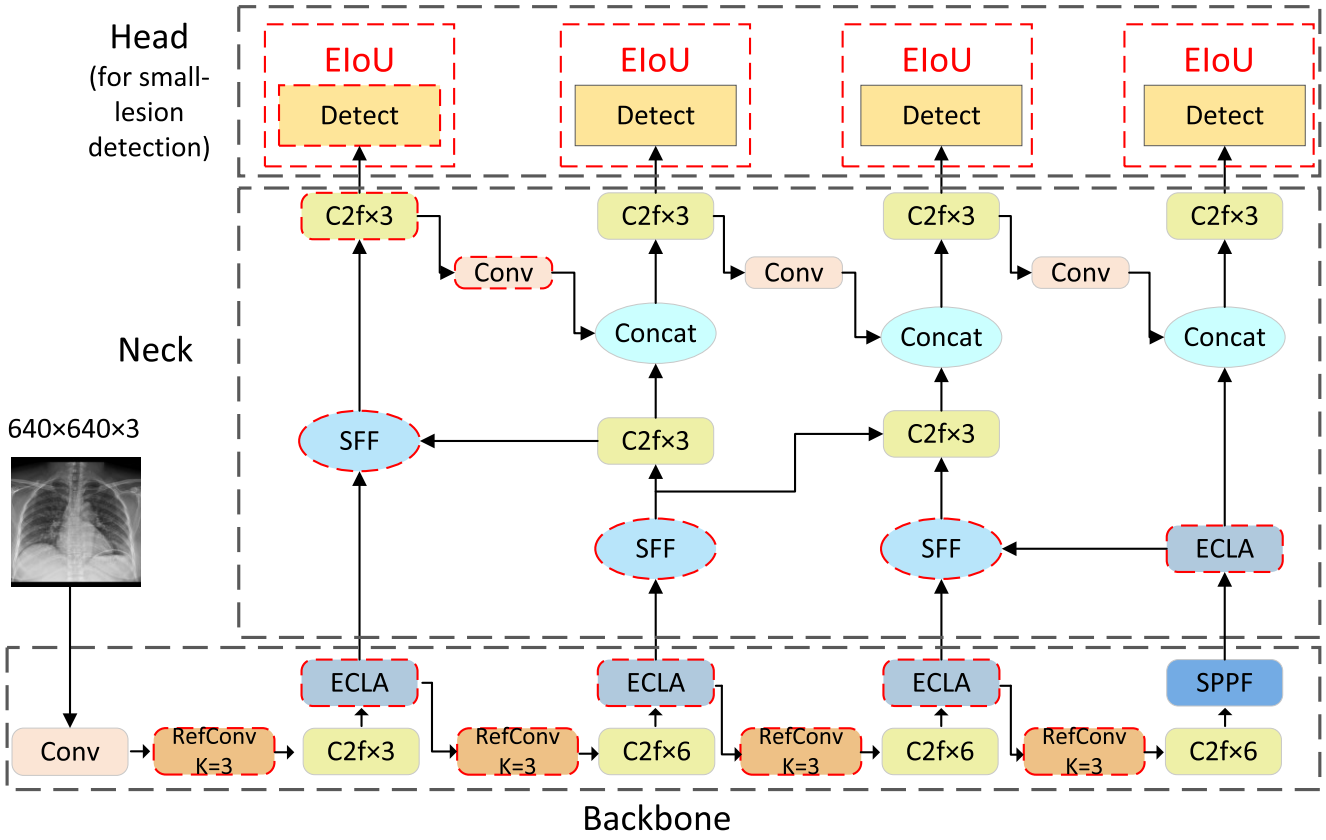


FIGURE 2. The proposed YOLO-CXR network.

results, obtained on the VinDr-CXR dataset, demonstrated effective integration of global and local lesion classification information. Lin et al. [36] introduced a new neural network for lung disease detection. By integrating the benefits of RepVGG [42] and Resblock [19] for information fusion and extraction, they developed a backbone, named RRNet, with a minimal number of parameters and robust feature extraction abilities. RRNet, combined with an enhanced version of RefineDet [43], forms a comprehensive network, referred to as CXR-RefineDet, which achieved good detection accuracy and fast inference speed. Ngo et al. [37] employed deep learning methods to identify anomalies in CXR images, enhancing the performance through data science and statistical techniques. They used ResNet in a Feature Pyramid Network (FPN) [44] backbone for the RetinaNet model, by applying data augmentation and optimization techniques, and achieved good performance.

The presented methods have improved the chest disease detection from multiple aspects, including data processing, algorithms, and model frameworks, and achieved promising results, improving the overall accuracy of detection. However, they also have some limitations. For instance, current networks still struggle with feature extraction when dealing with complex CXR images, resulting in relatively lower detection accuracy compared to other fields. Additionally, researchers have not specifically addressed the detection of small lesions, which is crucial for identifying multiple chest diseases based on CXR images. To address these limitations, in this paper, we improved the YOLOv8s backbone and proposed a new attention mechanism to enhance the network's feature extraction capabilities. Additionally, we employed a novel feature fusion method and incorporated a small-lesion detection head specifically for detecting small lesions.

III. PROPOSED NETWORK

The overall structure of the proposed YOLO-CXR network, is depicted in Figure 2. It is based on YOLOv8s by replacing the ordinary downsampling convolutional layers in its backbone with RefConv layers, introducing the newly elaborated ECLA mechanism into it, and integrating a small-lesion detection head utilizing the SFF technique for enhancing the feature fusion. More details of these are presented in the following subsections.

A. YOLOv8 FUNDAMENTALS

YOLOv8 is an advanced version of the You Only Look Once (YOLO) [45] object detection model, refined over several iterations to achieve both high inference speed and accuracy. It utilizes a unified architecture for object detection

via a single forward pass through the network, in contrast to traditional two-stage detectors.

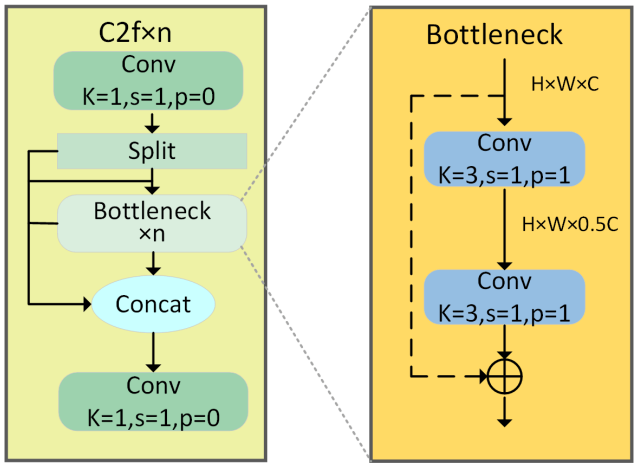


FIGURE 3. The C2f module, utilized by the YOLOv8s backbone.

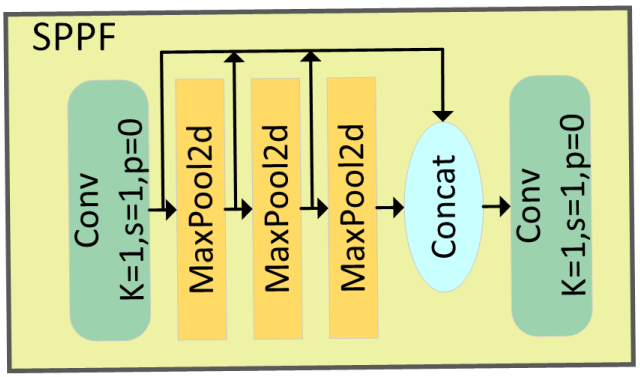


FIGURE 4. The SPPF module, utilized by the YOLOv8s backbone.

There are five versions of YOLOv8: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. Among them, YOLOv8s achieves a good balance between inference speed and detection accuracy. Therefore, in the presented study, YOLOv8s was chosen as a foundation of the proposed network for detection of chest diseases based on CXR images. The YOLOv8s network architecture is primarily divided into three parts: a backbone, a (feature enhancement) neck, and a (detection) head. Specifically, the backbone includes convolution (Conv), Faster Implementation of Cross Stage Partial Bottleneck with 2 convolutions (C2f), and Spatial Pyramid Pooling - Fast (SPPF) modules, whereby the Conv module performs feature-map downsampling, and the C2f and SPPF modules enhance feature extraction capabilities. The C2f module expands gradient branches via bottleneck modules to obtain richer information, [46]. As illustrated in Figure 3, C2f employs two convolutional layers to perform feature transformations on the data, aiding in extracting multi-level features. Additionally, a residual structure is utilized, where a branch undergoes bottleneck processing through the bottleneck modules. The bottleneck structure reduces computational complexity and the number of parameters, while improving the model's ability to handle complex data. The structure of the SPPF module is depicted in Figure 4. The input feature map is first processed by a convolutional block, which is used to reduce the amount of computation and extract the preliminary features. The three MaxPool2d layers reduce computational complexity and capture information at different scales by pooling across different regions of the feature map using specific convolutional kernel sizes.

The main function of the neck is to integrate features of different scales extracted by the backbone using a feature pyramid network (FPN) [44] and a path aggregation network (PAN) [47]. This integration facilitates the localization and identification of targets, enhancing the detection at various scales.

Within the head, predictions for objects of different sizes are implemented. YOLOv8s also integrates advanced regularization techniques, enhancing its generalization ability and robustness in diverse and challenging environments. YOLOv8s excels in object detection with its high accuracy and fast inference capabilities, making it the chosen baseline for our work.

B. IMPROVED YOLOv8s BACKBONE

As currently, convolution operations are widely applied across various computer vision tasks, enhancing their efficiency is crucial with increasing the network depth and width. In typical convolutional layers, input feature maps undergo convolution with kernels to detect and extract local features from input data. However, this approach has limitations when detecting chest diseases with variable characteristics.

RefConv [48] improves the ordinary convolution operation by re-parameterizing the input feature map. It introduces a learnable parameterized matrix that re-parameterizes the input feature map to a new feature map, which is better suited to the detection capabilities of the convolutional kernels, thus enhancing the network ability to recognize lesions in complex CXR images and allowing to extract more useful information.

Additionally, ordinary convolutional kernels only capture local information from the input feature map. In contrast, RefConv can re-parameterize the input feature map and dynamically adjust it during the network training, thereby improving the network's efficiency in feature extraction.

Specifically, the shape of the ordinary convolutional kernels is determined by the number of output channels C_{out}, the number of input channels C_{in}, the kernel size K, and the number of groups G, as follows:

$$W_o \in \mathbb{R}^{C_{out} \times \left(\frac{C_{in}}{G}\right) \times K \times K} \tag{1}$$

where W_o denotes the original weights. Next, a learnable operation, called Refocusing Transformation, is applied to the original weights W_o to generate a new convolutional kernel,

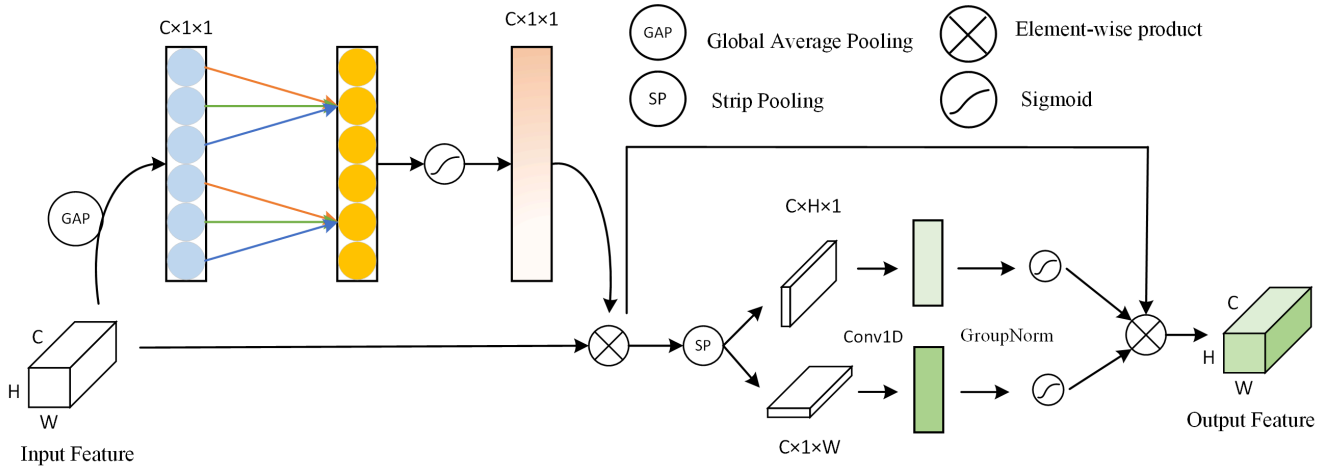


FIGURE 5. The newly designed ECLA module, utilized by the proposed YOLO-CXR network.

referred to as the transformed weights W_t:

$$W_t = W_o * W_r + W_b \tag{2}$$

where $*$ denotes the convolution operation and W_r denotes the learnable weights introduced by the Refocusing Transformation. This operation transforms the original weights W_o into new weights W_t, establishing connections between each convolutional kernel channel and the other channels. During the learning process, the focus is on optimizing the refocusing weights W_r. Finally, the transformed weights W_t are used instead of the original weights to operate on the input features. The shape of the original weights W_o remains unchanged, and the shape of the transformed weights W_t is defined as follows:

$$W_t \in \mathbb{R}^{C_{out} \times \left(\frac{C_{in}}{G}\right) \times K \times K} \tag{3}$$

The carefully designed Refocusing Transformation technique allows RefConv to correlate the parameters of specific kernel channels with those of other kernel channels. This enables the network to focus on different parts rather than just on the input features, facilitating the learning of new representations. During the network training, each kernel channel interacts with corresponding feature channels, encoding information from other channels. This process establishes connections between different feature channels, allowing for better feature representation. By replacing the ordinary convolutional layers with RefConv layers in the YOLOv8s backbone, the proposed network can better focus on different channels, reducing the channel similarity and redundancy. This enables the network to learn more diverse representations and enhances its representational capacity [49], [50].

C. EFFICIENT CHANNEL AND LOCAL ATTENTION (ECLA)

Attention mechanisms have garnered significant interest in the field of computer-aided diagnosis due to their numerous advantages, such as the low inference cost, fast computation, and the potential for enhancing the network performance. However, commonly used attention mechanisms for localizing chest diseases have certain limitations. For instance, the Squeeze-and-Excitation (SE) attention compresses the spatial dimensions after global pooling, followed by a complex process of dimensionality reduction and augmentation. This approach risks losing spatial information in the feature maps, rendering the network insensitive to disease localization. In contrast, the Coordinate Attention (CA) combines spatial position information with channel attention, enabling the network to capture broader spatial relationships. However, CA also has drawbacks, including limited generalization capability and adverse effects on channel dimension reduction.

Therefore, we propose here the ECLA mechanism for the detection of thoracic diseases. The overall structure of the corresponding ECLA module, utilized by the proposed network, is illustrated in Figure 5. In the channel dimension, it utilizes a weight-shared one-dimensional (1D) convolution to replace frequent channel dimension reductions and expansions [51]. This approach not only enables local cross-channel information exchange but also reduces information loss while also ensuring module efficiency. The convolutional kernel size k of the 1D convolution is dynamically determined by the input channel number C, as follows:

$$k = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{4}$$

where $|t|_{odd}$ denotes the nearest odd number greater than or equal to t. In the experiments, parameters γ and b were set to 2 and 1, respectively. After convolution, the output passes through a Sigmoid activation function to ensure the output values range from 0 to 1. Then, the normalized output undergoes dimensional transformation to restore its shape, and finally, the original feature map is multiplied by the channel attention weights. The computation process for the channel attention is implemented as follows:

$$M_c = CWM\left(Sigmoid\left(Conv1D\left(GAP\left(x\right)\right)\right)\right) \tag{5}$$

where M_c denotes the output feature map, *CWM* denotes the channel-wise multiplication, and *GAP* denotes the global average pooling.

In the spatial dimension, ECLA employs strip pooling both horizontally and vertically on the feature map, allowing the network to capture long-range dependencies and reduce redundant information interference from irrelevant areas, thus focusing on the positional information of the target disease in both directions. Subsequently, 1D convolutions are applied to interact locally with the two feature vectors. These 1D convolutions are adept at handling such information, enhancing the capability to capture positional information.

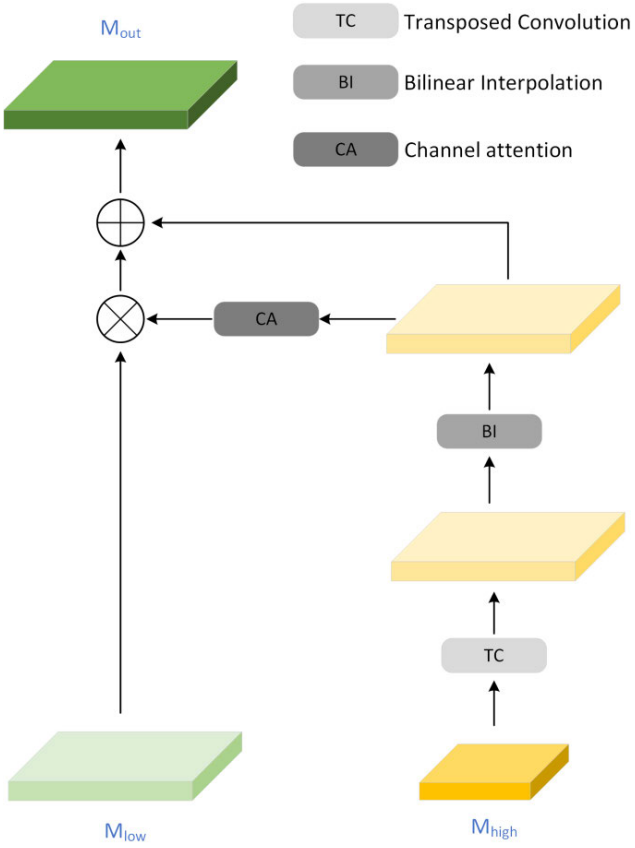


FIGURE 6. The SFF module, utilized by the proposed YOLO-CXR network.

The resultant feature vectors undergo GroupNorm and Sigmoid processing to generate positional attention predictions in two directions [52]. This method of generating an attention avoids sacrificing channel-wise information and effectively captures spatial information. The attention scores for horizontal and vertical directions S_H and S_V are computed as follows:

$$S_H = Sigmoid\left(GN\left(conv1D(SP_H(M_c))\right)\right) \quad (6)$$

$$S_V = Sigmoid\left(GN\left(conv1D(SP_V(M_c))\right)\right) \quad (7)$$

where SP_H and SP_V denote the strip pooling operations performed in the horizontal and vertical directions, respectively, and *GN* represents GroupNorm. The final features are obtained by an element-wise multiplication of the positional attentions from both directions with the original features.

D. DETECTION OF SMALL LESIONS

With advancements in deep learning, substantial progress has been achieved in the field of automatic detection of medical diseases. However, detecting small lesions remains a major challenge in medical imaging due to their minute size and complex geometric shapes.

In the field of chest disease detection, numerous instances of extremely small lesions pose significant challenges for their accurate detection and diagnosis, as these subtle abnormalities can easily be missed by both human radiologists and automated models. According to previous studies, shallow-layer features play more effective role in detecting such targets, especially when they are small and their features are not prominent [53]. Therefore, we introduced an additional small-lesion detection head into the shallow layers of the network to address this issue. This new structure aims to mitigate the adverse effects caused by inconsistent disease scales and enhance the detection of small lesions. As shown in Figure 2, we integrated the detection head after the C2f×3 module at the first high-resolution level. In addition, we implemented a new fusion strategy using the Select Feature Fusion (SFF) [10] within a multi-level feature fusion pyramid. Given the inevitable presence of some redundant information in shallow-layer features, this approach synergistically integrates deep and shallow information within the network. This integration enriches the resulting features with comprehensive semantic information, enabling the network to capture a broader range of chest disease characteristics. The structure of the corresponding SFF module, utilized by the proposed YOLO-CXR network, is illustrated in Figure 6.

SFF employs a transposed convolution and a bilinear interpolation to upsample deep-layer features, as this allows direct scaling of the feature maps, providing greater flexibility in handling feature map sizes. Specifically, given deep-layer features $M_{high} \in \mathbb{R}^{C\times H\times W}$ and shallow-layer features $M_{low} \in \mathbb{R}^{C\times H1\times W1}$, SFF performs upsampling of the deep-layer features using a transposed convolution with a stride of 2 and a kernel size of 3, resulting in $M'_{high}\in\mathbb{R}^{C\times 2H\times 2W}$. Then, a bilinear interpolation to M'_{high} is applied to ensure consistency in dimensions between deep-layer and shallow-layer features, resulting in $M_b\in\mathbb{R}^{C\times H1\times W1}$. Using a channel attention mechanism, the deep-layer features are transformed into attention weights, which are then used to remove redundant information from the shallow-layer features. Finally, the updated shallow-layer features are fused with the deep-layer features to obtain $M_{out}\in\mathbb{R}^{C\times H1\times W1}$, completing the feature fusion process, as follows:

$$M_b = BI\left(TConv\left(M_{high}\right)\right) \quad (8)$$

$$M_{out} = M_{low} * CA\left(M_b\right) + M_b \quad (9)$$

where, *TConv* denotes a transposed convolution, *BI* denotes a bilinear interpolation, and *CA* denotes a channel attention.

In the process of applying the SFF to a small-lesion detection head, considering shallow-layer features $M_{low} \in \mathbb{R}^{128\times160\times160}$ and deep-layer features $M_{high} \in \mathbb{R}^{256\times80\times80}$, first, a transposed convolution is applied to M_{high} for upsampling, resulting in $M'_{high} \in \mathbb{R}^{128\times160\times160}$. Then, a bilinear interpolation is performed on M'_{high} to ensure consistent dimensions between the deep-layer and shallow-layer features, producing $M_b \in \mathbb{R}^{128\times160\times160}$. Subsequently, M_b is fed into a channel attention module, where it is transformed into attention weights, used to refine the shallow-layer features by removing redundant information. Finally, the updated shallow-layer features are fused with the deep-layer features to obtain $M_{out} \in \mathbb{R}^{128\times160\times160}$, ensuring that the output retains the same shape as the shallow-layer features.

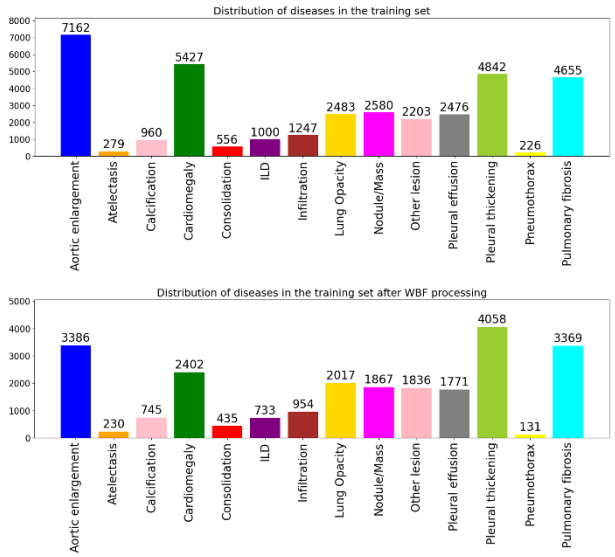


FIGURE 7. Distribution of disease types before WBF processing (top chart) and after WBF processing (bottom chart).

IV. EXPERIMENTS AND RESULTS

A. DATASET

In the experiments, the VinDr-CXR dataset [54] was used, consisting of large-scale CXR images with high-quality annotations.

The images were sourced from two prominent hospitals in Vietnam: Hospital 108 and Hanoi Medical University Hospital. The dataset comprises 18,000 images annotated for both localization of critical findings and classification of common chest diseases, with each image possibly containing multiple annotations. For the experiments, the dataset was split into a training set of 15,000 images and a test set of 3,000 images. Each image in the training set was independently labeled by three radiologists, whereas test-set annotations were derived by a consensus among five radiologists after more rigorous processing.

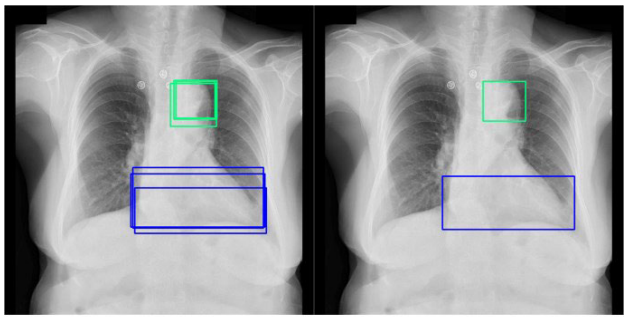


FIGURE 8. Comparison of chest X-ray images: (a) before WBF processing; (b) after WBF processing.

Due to the independent annotations by three professional radiologists, the data contained numerous overlapping bounding boxes. To address this, we used the Weighted Box Fusion (WBF) technique to reduce the overlap and redundancy.

Table 2 presents the counts and proportions of 14 diseases, images of which are contained in the VinDr-CXR dataset, before and after the WBF processing. Among these diseases, aortic enlargement and cardiomegaly have the highest counts, each exceeding 15% in proportion, while atelectasis and pneumothorax account for less than 1%. The dataset exhibits a long-tail distribution, indicating that it is imbalanced. Additionally, we created histograms of disease type distributions before and after the WBF processing to visually compare the data distribution changes, as shown in Figure 7. From Table 2 and Figure 7, it can be concluded that the WBF processing reduces the overlapping and the number of redundant bounding boxes, thus enhancing the clarity and accuracy of the annotation boxes while also minimizing the waste of computational resources [32]. Moreover, the overall data distribution remains largely unaffected.

WBF first sorts all bounding boxes in descending order based on confidence scores determining the value of the weights (lower confidence scores indicate less accurate predictions). For each target box in an image, fusion is performed to generate a new list of bounding boxes, checking for matches with the initial bounding boxes. This is achieved by determining whether the Intersection over Union (IoU) exceeds a predefined threshold. Then, new coordinates and confidence scores are calculated using a fusion formula. Figure 8a shows sample original annotations, whereas Figure 8b shows the results after the WBF processing.

B. EVALUATION METRICS

To analyze the experimental results, we employed evaluation metrics, briefly introduced below, which are commonly used in the field of object detection to assess the experimental outcomes.

Precision measures the percentage of correctly predicted samples out of all predicted samples, as follows:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

TABLE 2. Counts and proportions of disease types in the VinDr-CXR dataset before and after WBF processing.

Disease types ↓	Before WBF processing		After WBF processing	
	Counts	Proportion	Counts	Proportion
Aortic Enlargement	7162	19.84%	3386	14.15%
Cardiomegaly	5427	15.03%	2402	10.04%
Pleural Thickening	4842	13.41%	4058	16.95%
Pulmonary Fibrosis	4655	12.90%	3369	14.08%
Nodule/Mass	2580	7.15%	1867	7.80%
Lung Opacity	2483	6.88%	2017	8.43%
Pleural Effusion	2476	6.86%	1771	7.40%
Other Lesions	2203	6.10%	1836	7.67%
Infiltration	1247	3.45%	954	3.99%
Interstitial Lung	1000	2.77%	733	3.06%
Calcification	960	2.66%	745	3.11%
Consolidation	556	1.54%	435	1.82%
Atelectasis	279	0.77%	230	0.96%
Pneumo-Thorax	226	0.63%	131	0.55%

where True Positives (*TP*) denotes the number of correctly detected samples with *IoU* (between the detected and ground-truth boxes) exceeding a predefined threshold, while False Positives (*FP*) represents the number of incorrectly detected samples with *IoU* below the threshold.

Recall measures the proportion of correctly detected samples among all ground-truth objects, as follows:

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

where False Negatives (*FN*) indicates the number of missed detections with undetected ground-truth boxes.

Average precision (*AP*) is an important metric that evaluates the validity of detection networks by representing the area under the *precision-recall* curve, with higher values indicating better performance. The mean average precision (*mAP*) for all classes provides comprehensive assessment of the detection performance of each class in a task. The *mAP* is calculated as follows:

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{12}$$

where n represents the number of classes. In computer-assisted clinical diagnosis, *mAP* using 50% *IoU* threshold can not only capture the location of important lesions, but also allows deviation within a reasonable range, which has practical clinical application value [34]. For more comprehensive results, *mAP@[0.5:0.95:0.05]* was also used to evaluate the network performance at different *IoU* thresholds. This metric represents the *mAP* calculated across 10 *IoU* thresholds, ranging from 0.5 to 0.95 in increments of 0.05.

C. EXPERIMENTAL ENVIRONMENT

We use the Ubuntu operating system, Python 3.8 as the programming language, and PyTorch 1.11.0 as a deep learning framework. Training was conducted on an NVIDIA RTX 3090 GPU. The resolution of all training and testing samples was adjusted to 640 × 640. The network was trained for 300 epochs using the SGD optimizer with a momentum set to 0.937, a batch size of 16, an initial learning rate of 1e-2, and a final learning rate of 1e-4. An early stopping strategy was employed during the network training to prevent overfitting. Additionally, images were horizontally flipped with a probability of 0.5, and, in addition, the contrast, saturation, and hue were randomly varied between 90% and 110%.

D. ANALYSIS OF EXPERIMENTAL RESULTS

The experimental analysis aims to assess the network's performance by thoroughly examining the training process and the resulting values of the evaluation metrics.

During the training process, we monitored the learning curves and tracked the key performance metrics. Figure 9 presents the generated training loss and validation loss curves. The training loss measures the discrepancy between detection results and ground truth during the network training, whereas the validation loss is used to evaluate the network performance on the validation set. Typically, the training loss gradually decreases as a network optimizes its parameters to minimize the loss function. However, the validation loss may start to increase after a certain number of training epochs. This is often due to a network overfitting on a dataset, leading to a decline in its generalization ability on the validation set. During the first 200 epochs, both training

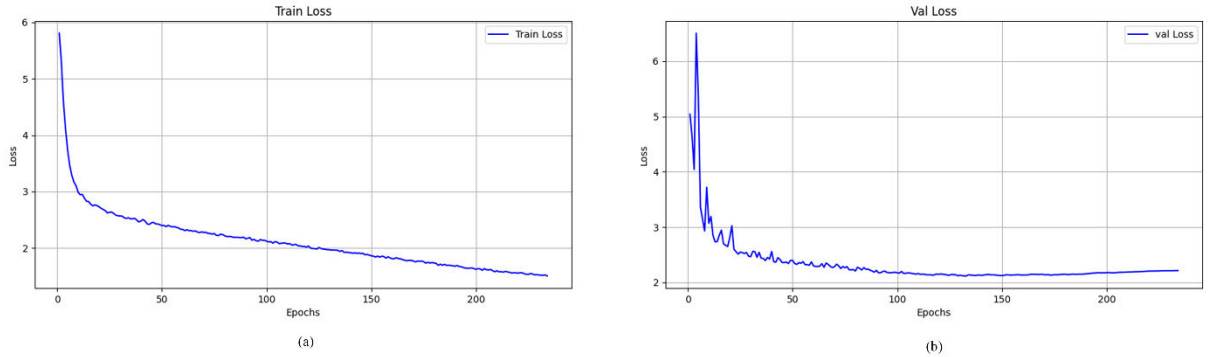


FIGURE 9. (a) The training loss curve; (b) The validation loss curve.

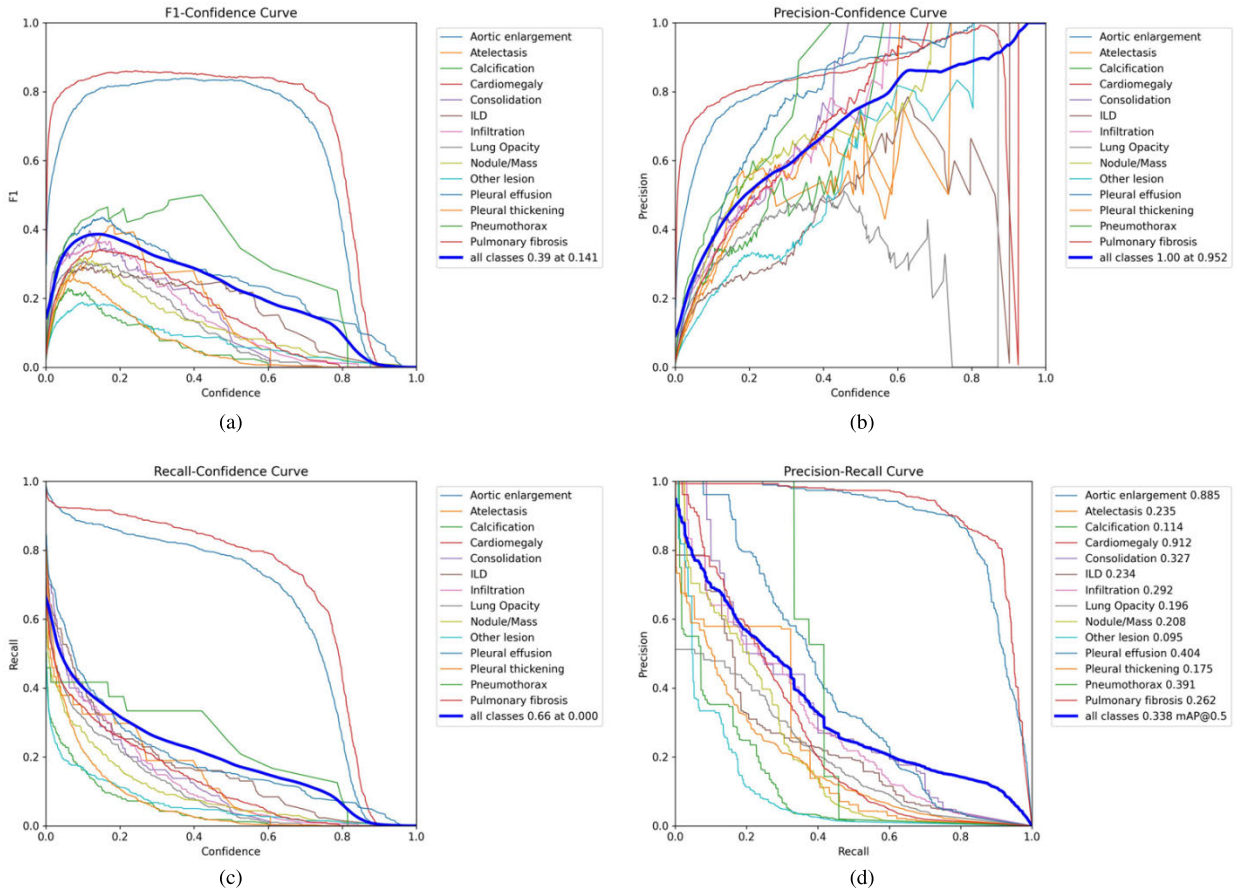


FIGURE 10. (a) The *F1* confidence curves; (b) the *precision* confidence curves; (c) the *recall* confidence curves; (d) the *precision-recall* curves.

and validation losses of the proposed network decreased as its training progressed. After 200 epochs, the validation loss converged and no longer decreased, indicating that the network training had reached an optimal point. Due to the use of an early stopping strategy, the training was terminated early.

Additionally, we performed a statistical analysis of the experimental results using the confidence curves for *F1*, *precision*, and *recall*, along with the *precision-recall* curves, illustrated in Figure 10. As can be seen in Figure 10a, the *F1* score for different conditions gradually increases at lower confidence levels, reaching a peak before declining as confidence continues to rise. This indicates that the proposed YOLO-CXR network performs better in distinguishing positive from negative samples at lower confidence thresholds. Figure 10b reflects that YOLO-CXR maintains high *precision* even at lower confidence levels, demonstrating strong accuracy in target identification. Figure 10c shows

TABLE 3. Pros and cons of the SOTA networks participated in the experiments.

Network	Pros	Cons
YOLOv8 [9]	Excellent flexibility and scalability	Lacks specific methods for detecting small lesions and feature fusion techniques
YOLOv9 [26]	More lightweight and easier to deploy	Lower average precision in detecting lesions
Sparse RCNN [55]	Not relying on dense object candidates	Insufficient support for small object detection
TridentNet [56]	Using a parallel multi-branch architecture with high precision	Limited feature fusion
EARL [57]	Reducing the negative impact of label quality	Non-end-to-end architecture with low average precision
DualAttNet [35]	Effectively integrating global and local lesion information	Poorly performance in detecting small lesions
YOLO-CXR (*proposed*)	Improved average precision, high small-lesion detection, and enhanced feature fusion	Increased model complexity and lack of a lightweight design

that the proposed network can detect more positive samples at lower confidence levels, but as confidence increases, the network tends to forgo detections on some samples, thereby reducing the *recall* rate. Figure 10d reveals significant variations in the *precision-recall* curves across different disease types. Certain diseases, such as aortic enlargement, exhibit a favorable balance between *precision* and *recall*, while others, like pulmonary fibrosis, perform poorly. Overall, each curve provides unique insights into the network's performance. From these curves, one can conclude that the proposed network generally performs well, with high detection and discrimination abilities for certain diseases. However, the network's shortcomings are evident in its unstable performance for some diseases, likely due to the influence of data distribution and the scarcity of samples.

E. COMPARISON WITH SOTA NETWORKS

To objectively evaluate the proposed YOLO-CXR network, we selected various networks for comparison, all of which represented the state-of-the-art (SOTA) in object detection at the time of their release. The pros and cons of these networks are summarized in Table 3. Table 4 shows the performance results of the proposed network in comparison to these SOTA networks, based on the VinDr-CXR dataset, using multiple metrics for more comprehensive evaluation. From this table, the following conclusions can be drawn:

(1) The proposed YOLO-CXR network achieved a significant improvement on *mAP@0.5, mAP@[0.5:0.95:0.05]*and *recall* over the SOTA networks. Moreover, there were performance enhancements across all 14 disease classes, indicating the effectiveness of YOLO-CXR.

(2) Different networks are suitable for detecting different types of chest lesions. For instance, Sparse RCNN exhibited best performance in detecting aortic enlargement, cardiomegaly, and pleural thickening. TridentNet showed top performance in detecting consolidation, ILD, and 'other lesions'. The proposed YOLO-CXR network demonstrated best performance in detecting the remaining eight diseases. It also achieved second best *AP* results for cardiomegaly (0.912) and consolidation (0.327), indicating an enhanced utilization of spatial positional information in locating chest abnormalities. In detecting small lesions, such as atelectasis and lung opacity, YOLO-CXR demonstrated the best performance. This is mainly attributed to the inclusion of a small-lesion detection head and improvements in feature fusion methods, which allowed it to effectively leverage information at different scales.

(3) The proposed network was the first runner-up in detecting 'other lesions', primarily due to the existence of only few samples in this class. The insufficient sample size negatively impacts the network training effectiveness, highlighting that YOLO-CXR lacks a bit the capability to handle imbalanced data. This underscores the need to improve the proposed network's handling of imbalanced datasets.

F. ABLATION STUDY

We conducted also an ablation study to rigorously evaluate the effectiveness of each main component of the proposed network. Additionally, we visually compared its detection performance to that of the baseline, showcasing its efficacy. We experimented with the utilization of the three main components (individually and in combinations) by the proposed network, using the following steps shown in Table 5: (0) using the baseline (YOLOv8s) without additional components; (1) replacing the ordinary convolutional layers with RefConv layers to enhance the YOLOv8s backbone; (2) adding only the ECLA mechanism to the baseline; (3) incorporating only the small-lesion detection head in the baseline, with SFF for pyramid feature fusion; (4) using RefConv and ECLA simultaneously; (5) using ECLA and SFF simultaneously; and (6) using RefConv, ECLA, and SFF simultaneously (resulting in the proposed YOLO-CXR network).

Based on results contained in Table 5, one can draw the following conclusions:

TABLE 4. AP, mAP@0.5, mAP@[0.5:0.95:0.05] and recall results of network performance comparison, performed on the VinDr-CXR dataset (the top values are highlighted in bold).

Disease types ↓	Networks						
	YOLOv8 [9]	YOLOv9 [26]	Sparse RCNN [55]	TridentNet [56]	EARL [57]	DualAttNet [35]	YOLO-CXR (*proposed*)
Aortic Enlargement	0.887	0.877	**0.917**	0.891	0.549	0.634	0.885
Atelectasis	0.220	0.181	0.076	0.094	0.149	0.153	**0.235**
Calcification	0.091	0.053	0.112	0.092	0.085	0.061	**0.114**
Cardiomegaly	0.893	0.903	**0.932**	0.904	0.566	0.671	0.912
Consolidation	0.267	0.309	0.111	**0.335**	0.139	0.199	0.327
ILD	0.247	0.203	0.207	**0.305**	0.281	0.255	0.234
Infiltration	0.289	0.282	0.173	0.239	0.185	0.232	**0.292**
Lung Opacity	0.181	0.150	0.106	0.192	0.154	0.155	**0.196**
Nodule/Mass	0.128	0.199	0.130	0.156	0.137	0.140	**0.208**
Pleural Effusion	0.309	0.343	0.331	0.362	0.252	0.295	**0.404**
Pleural Thickening	0.140	0.136	**0.194**	0.162	0.104	0.189	0.175
Pneumo-Thorax	0.332	0.336	0.028	0.328	0.123	0.142	**0.391**
Pulmonary Fibrosis	0.241	0.261	0.143	0.230	0.144	0.197	**0.262**
Other Lesions	0.063	0.055	0.026	**0.106**	0.052	0.082	0.095
mAP@0.5	0.306	0.306	0.249	0.314	0.208	0.243	**0.338**
mAP@[0.5:0.95:0.05]	0.150	0.156	-	-	0.143	0.116	**0.167**
recall	0.326	0.325	-	-	0.238	0.257	**0.365**

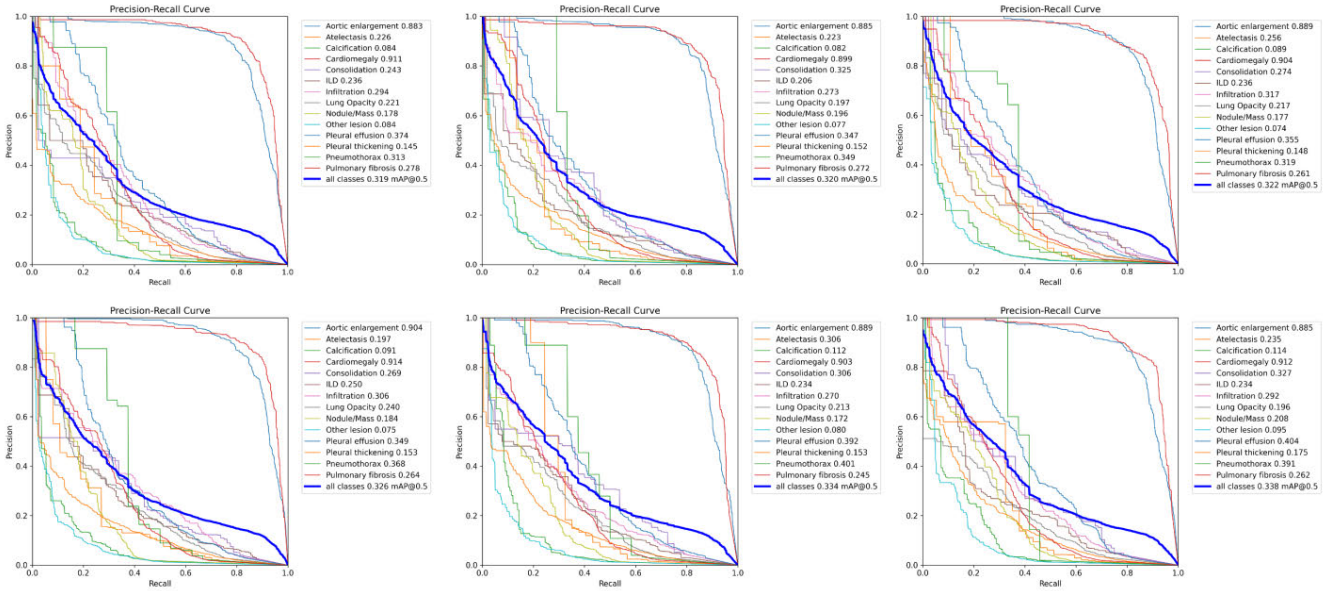


FIGURE 11. The*precision-recall* curves of the ablation study using: (a) RefConv only; (b) ECLA only; (c) only a small-lesion detection head with SFF; (d) RefConv and ECLA simultaneously; (e) ECLA and SFF simultaneously; (f) RefConv, ECLA, and SFF simultaneously (resulting in the proposed YOLO-CXR network).

(1) Adopting RefConv layers by the proposed network as a replacement of the ordinary convolutional layers in the YOLOv8s backbone allows to improve *mAP@0.5* to 0.319 and *mAP@[0.5:0.95:0.05]* to 0.157, indicating increases of 0.013 and 0.007, respectively, compared to the baseline (YOLOv8s). Thus, it can be concluded that RefConv enhances the feature extraction capability of the backbone, thereby improving the network's detection performance.

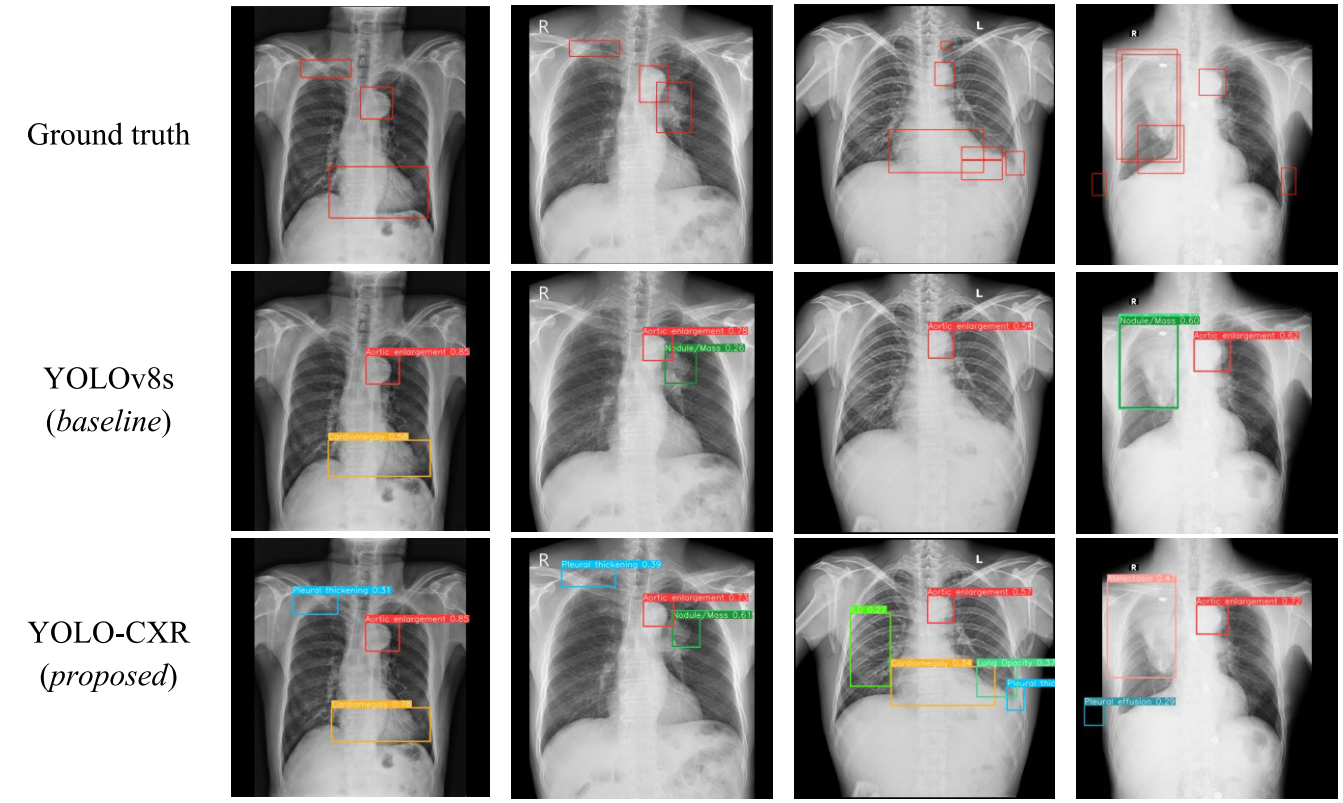


FIGURE 12. Visualization of detection results of the proposed YOLO-CXR network vs. the baseline (YOLOv8s).

TABLE 5. Results of ablation study experiments.

RefConv	ECLA	SFF	*mAP@50*	*mAP@[0.5:0.95:0.05]*
			0.306	0.150
√			0.319	0.157
	√		0.320	0.161
		√	0.322	0.162
√	√		0.326	0.158
	√	√	0.334	0.166
√	√	√	**0.338**	**0.167**

(2) The incorporation of the ECLA module into the proposed network allows to enhance its capability to capture information in both the channel and spatial dimensions, resulting in an *mAP@0.5* value of 0.320 and *mAP@[0.5:0.95:0.05]* value of 0.161, with improvements of 0.014 and 0.011, respectively, compared to the baseline.

(3) Additionally, using both RefConv and ECLA to enhance the backbone network resulted in a *mAP@0.5* value of 0.326 and a *mAP@[0.5:0.95:0.05]* value of 0.158, with increases of 0.020 and 0.008, respectively, compared to the baseline. This indicates that ECLA significantly enhances the feature extraction capability of the backbone, making the network more sensitive to spatial positional information of lesions.

(4) Integrating a small-lesion detection head into the network and using the SFF technique has yielded promising results. Specifically, *mAP@0.5* reached 0.322 and *mAP@[0.5:0.95:0.05]* reached 0.162, with increases of 0.016 and 0.012, respectively, compared to the baseline.

(5) By simultaneously using both SFF and ECLA to enhance the backbone, *mAP@0.5* reached 0.334 and *mAP@[0.5:0.95:0.05]* reached 0.166, with increases of 0.028 and 0.016, respectively, compared to the baseline. Thus, integrating a small-lesion detection head and employing the SFF technique significantly enhances small-lesion detection, reduces redundancy in information fusion, and notably improves the network overall detection performance.

Figure 11 presents the *precision-recall* curves for each step of the ablation study, with each graph corresponding to a specific experimental result. This figure illustrates the relationship between *precision* and *recall* for different disease types, alongside the calculated *mAP* score. From the figure, one can conclude that as more of the proposed components are added to the baseline during the ablation study, most of the curves become smoother and more upwardly convex. This suggests that the network becomes more effective at detecting a greater number of lesions. Furthermore, the rising *mAP* score with the inclusion of additional components reflects the effectiveness of each of them.

Figure 12 intuitively demonstrates the effectiveness of the proposed YOLO-CXR network by comparing its results

to the ground truth and to the results of the baseline. YOLO-CXR demonstrates precise localization of multiple chest abnormalities, with close alignment to the ground-truth bounding boxes, particularly in handling various lesion areas and small targets. In addition, the proposed network captures semantic information across different scales while maintaining overall coherence. Moreover, its utilization of the SFF technique enhances the feature fusion, allowing it to obtain more relevant positional information related to different diseases, thus enhancing its capability to accurately localize different lesions.

V. DISCUSSIONS

In recent years, with the advancement of deep learning, automatic detection of chest diseases has alleviated some of the diagnostic workload on radiologists. However, numerous challenges remain in this field. For example, current networks designed to simultaneously detect multiple chest diseases with varying lesion types often struggle to accurately extract small lesions, or to capture their spatial locations, or to delineate pathological features, which all negatively impact their detection performance. To address these challenges, we propose here the YOLO-CXR network, based on YOLOv8s. Comprehensive experimental results, obtained on the VinDr-CXR dataset, demonstrate that YOLO-CXR achieves top detection performance.

Specifically, its highest *mAP* score among the SOTA networks indicates that the proposed network performs better in identifying various diseases based on chest images, thereby reducing the risk of misdiagnosis. It also suggests that YOLO-CXR is more effective at identifying true positive cases, which is crucial for early detection of diseases. This also means that the proposed network more effectively reduces false positive and false negative counts, providing more reliable decision support to assist clinicians, as it allows them to focus on more complex cases or confirm diagnoses more efficiently. This can streamline workflows in busy clinical environments, leading to better resource utilization and potentially faster patient treatment. Overall, the proposed network is more accurate and reliable than the SOTA networks, potentially leading to better clinical decision making, earlier interventions, and improved patient care in real-world settings. As illustrated in Figure 12, YOLO-CXR accurately locates lesions of various sizes, demonstrating significant improvement in lesion detection compared to the baseline (YOLOv8s).

Although the proposed network achieved top results among the SOTA networks, it still has some limitations. First, its detection accuracy on the VinDr-CXR dataset fell short of the expected level, especially for diseases like pleural thickening. This is mainly due to insufficient focus on data imbalance, which resulted in suboptimal detection accuracy for less prevalent diseases. Second, compared to advanced networks, like YOLOv9 [26], the proposed YOLO-CXR network lags in parameter count and computational speed. Therefore, our future work will focus on addressing data imbalance to better handle complex real-world clinical environments and elaborate a lightweight version of the proposed network with an improved detection accuracy, while also preserving the computational resources.

VI. CONCLUSION

This paper has introduced a novel chest disease detection network, YOLO-CXR, capable of detecting multiple diseases based on CXR images. Compared to conventional single-disease detection networks, the proposed network is more suitable for assisting physicians in early diagnosis and comprehensive patient assessment. It holds significant value in reducing the workload of radiologists in underdeveloped regions and enhancing the diagnostic efficiency. To achieve this, YOLO-CXR utilizes an improved YOLOv8s backbone to enhance its feature extraction capabilities. Additionally, it uses a novel ECLA mechanism to effectively extract information at different scales, thus capturing spatial location information of lesions across different feature map sizes. Furthermore, a small-lesion detection head is utilized by the proposed network to strengthen its detection of minor abnormalities. Subsequently, using the SFF technique for feature fusion allows it to improve its detection capability of various chest lesions.

Ultimately, with the proposed YOLO-CXR network, several challenges in chest disease detection are addressed, such as the significant size and positional variations among images corresponding to different disease types, which current networks struggle to identify, especially w.r.t. small lesions. YOLO-CXR outperforms existing networks, which will have positive impact on clinical practice. Looking ahead, our focus will shift towards addressing the challenges posed by data imbalance and optimizing the network for lightweight implementation, reducing its parameter count and computational complexity, with the aim of making further advancements in this field.

REFERENCES

[1] M. J. Divo, C. H. Martinez, and D. M. Mannino, "Ageing and the epidemiology of multimorbidity," *Eur. Respiratory J.*, vol. 44, no. 4, pp. 1055–1068, Oct. 2014.

[2] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest X-ray analysis: A survey," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102125.

[3] S. Sajed, A. Sanati, J. E. Garcia, H. Rostami, A. Keshavarz, and A. Teixeira, "The effectiveness of deep learning vs. Traditional methods for lung disease diagnosis using chest X-ray images: A systematic review," *Appl. Soft Comput.*, vol. 147, Nov. 2023, Art. no. 110817.

[4] F. Yimer, A. W. Tessema, and G. L. J. I. J. Simegn, "Multiple lung diseases classification from chest X-ray images using deep learning approach," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 5, pp. 2936–2946, 2021.

[5] C. Avramescu, A. Tenescu, B. Bercean, and M. Marcu, "Bounding box supervision benefits lung pathology classification in pulmonary X-Rays," in *Proc. IEEE 17th Int. Symp. Appl. Comput. Intell. Informat. (SACI)*, May 2023, pp. 000557–000560.

[6] M. Firmino, G. Angelo, H. Morais, M. R. Dantas, and R. Valentim, "Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy," *Biomed. Eng. OnLine*, vol. 15, no. 1, p. 2, Dec. 2016.

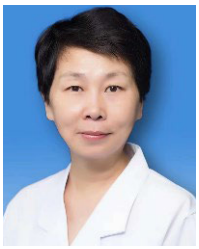
[7] N. O'Mahony, "Deep learning vs. traditional computer vision," in *Proc. Adv. Comput. Vis. Conf. (CVC)*, 2020, pp. 128–144.

[8] J. H. Lee, H. Hong, G. Nam, E. J. Hwang, and C. M. Park, "Effect of human-AI interaction on detection of malignant lung nodules on chest radiographs," *Radiology*, vol. 307, no. 5, Jun. 2023, Art. no. e222976.
[9] G. Jocher, A. Chaurasia, and J. Qiu. *Ultralytics YOLO*. Accessed: Sep. 22, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics
[10] Y. Chen, C. Zhang, B. Chen, Y. Huang, Y. Sun, C. Wang, X. Fu, Y. Dai, F. Qin, Y. Peng, and Y. Gao, "Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases," *Comput. Biol. Med.*, vol. 170, Mar. 2024, Art. no. 107917.
[11] I. W. Harsono, S. Liawatimena, and T. W. Cenggoro, "Lung nodule detection and classification from thorax CT-scan using RetinaNet with transfer learning," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 3, pp. 567–577, Mar. 2022.
[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
[13] X. Li, L. Shen, X. Xie, S. Huang, Z. Xie, X. Hong, and J. Yu, "Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection," *Artif. Intell. Med.*, vol. 103, Mar. 2020, Art. no. 101744.
[14] S. Xu, H. Lu, M. Ye, K. Yan, W. Zhu, and Q. Jin, "Improved cascade R-CNN for medical images of pulmonary nodules detection combining dilated HRNet," in *Proc. 12th Int. Conf. Mach. Learn. Comput.*, vol. 5, Feb. 2020, pp. 283–288.
[15] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[16] J. Mendoza and H. Pedrini, "Detection and classification of lung nodules in chest X-ray images using deep convolutional neural networks," *Comput. Intell.*, vol. 36, no. 2, pp. 370–401, May 2020.
[17] E.-S.-A. El-Dahshan, M. M. Bassiouni, A. Hagag, R. K. Chakrabortty, H. Loh, and U. R. Acharya, "RESCOVIDTCNnet: A residual neural network-based framework for COVID-19 detection using TCN and EWT with chest X-ray images," *Expert Syst. Appl.*, vol. 204, Oct. 2022, Art. no. 117410.
[18] J. Gilles, "Empirical wavelet transform," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3999–4010, Aug. 2013.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[20] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016, pp. 47–54.
[21] A. Tolkachev, I. Sirazitdinov, M. Kholiavchenko, T. Mustafaev, and B. Ibragimov, "Deep learning for diagnosis and segmentation of pneumothorax: The results on the kaggle competition and validation against radiologists," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1660–1672, May 2021.
[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, 2015, pp. 234–241.
[23] T. Agrawal and P. Choudhary, "ALCNN: Attention based lightweight convolutional neural network for pneumothorax detection in chest X-rays," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104126.
[24] H. Bhatt and M. Shah, "A convolutional neural network ensemble model for pneumonia detection using chest X-ray images," *Healthcare Anal.*, vol. 3, Nov. 2023, Art. no. 100176.
[25] M. F. Hashmi, S. Katiyar, A. W. Hashmi, and A. G. Keskar, "Pneumonia detection in chest X-ray images using compound scaled deep learning model," *Automatika*, vol. 62, nos. 3–4, pp. 397–406, Oct. 2021.
[26] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024, *arXiv:2402.13616*.
[27] R. Guo, K. Passi, and C. K. Jain, "Tuberculosis diagnostics and localization in chest X-Rays via deep learning models," *Frontiers Artif. Intell.*, vol. 3, Oct. 2020, Art. no. 583427.
[28] W. Fan, X. Guo, L. Teng, and Y. Wu, "Research on abnormal target detection method in chest radiograph based on YOLO v5 algorithm," in *Proc. IEEE Int. Conf. Comput. Sci., Electron. Inf. Eng. Intell. Control Technol. (CEI)*, vol. 10, Sep. 2021, pp. 125–128.
[29] S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid deep learning for detecting lung diseases from X-ray images," *Informat. Med. Unlocked*, vol. 20, Jun. 2020, Art. no. 100391.
[30] J. Lian, J. Liu, S. Zhang, K. Gao, X. Liu, D. Zhang, and Y. Yu, "A structure-aware relation network for thoracic diseases detection and segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 8, pp. 2042–2052, Aug. 2021.
[31] C. Lin, Y. Huang, W. Wang, S. Feng, and M. Huang, "Lesion detection of chest X-ray based on scalable attention residual CNN," *Math. Biosciences Eng.*, vol. 20, no. 2, pp. 1730–1749, 2023.
[32] Y. Yuan, L. Liu, X. Yang, L. Liu, and Q. Huang, "Multi-scale lesion feature fusion and location-aware for chest multi-disease detection," *J. Imag. Informat. Med.*, vol. 2, pp. 1–16, May 2024.
[33] H. T. Nguyen, M. N. Nguyen, S. C. Pham, and P. H. D. Bui, "Abnormalities detection on chest radiograph with bounding box-based lungs extraction and object detection algorithm," *Int. J. Inf. Technol.*, vol. 16, no. 4, pp. 2241–2251, Apr. 2024.
[34] H. Sheng, L. Ma, J.-F. Samson, and D. Liu, "BarlowTwins-CXR: Enhancing chest X-ray abnormality localization in heterogeneous data with cross-domain self-supervised learning," *BMC Med. Informat. Decis. Making*, vol. 24, no. 1, p. 126, May 2024.
[35] Q. Xu and W. Duan, "DualAttNet: Synergistic fusion of image-level and fine-grained disease attention for multi-label lesion detection in chest X-rays," *Comput. Biol. Med.*, vol. 168, Jan. 2024, Art. no. 107742.
[36] C. Lin, Y. Zheng, X. Xiao, and J. Lin, "CXR-RefineDet: Single-shot refinement neural network for chest X-ray radiograph based on multiple lesions detection," *J. Healthcare Eng.*, vol. 2022, pp. 1–11, Jan. 2022.
[37] N. Ngo, T. Vo, and L. Ngo, "Application of deep learning in chest X-ray abnormality detection," *Ministry Sci. Technol., Vietnam*, vol. 65, no. 4, pp. 84–93, 2023.
[38] G. Jocher. *YOLOv5 By Ultralytics*. Accessed: Sep. 22, 2024. [Online]. Available: https://github.com/ultralytics/yolov5
[39] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[40] G. Chierchia, D. Cozzolino, G. Poggi, and L. Verdoliva, "SAR image despeckling through convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5438–5441.
[41] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.
[42] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making VGG-style Convnets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, May 2021, pp. 13733–13742.
[43] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.
[44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[45] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[46] Y. Li, Q. Fan, H. Huang, Z. Han, and Q. Gu, "A modified YOLOv8 detection network for UAV aerial image recognition," *Drones*, vol. 7, no. 5, p. 304, May 2023.
[47] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
[48] Z. Cai, X. Ding, Q. Shen, and X. Cao, "RefConv: Re-parameterized refocusing convolution for powerful ConvNets," 2023, *arXiv:2310.10563*.
[49] X. Wang and X. Y. Stella, "Tied block convolution: Leaner and better CNNs with shared thinner filters," in *Proc. AAAI Conf. Artif. Intell.*, 2021, no. 11, pp. 10227–10235.
[50] Y. Zhou, Y. Zhang, Y.-F. Wang, and Q. Tian, "Accelerate CNN via recursive Bayesian pruning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3305–3314.
[51] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
[52] W. Xu and Y. Wan, "ELA: Efficient local attention for deep convolutional neural networks," 2024, *arXiv:2403.01123*.
[53] M. Zhang, S. Xu, W. Song, Q. He, and Q. Wei, "Lightweight underwater object detection based on YOLO v4 and multi-scale attentional feature fusion," *Remote Sens.*, vol. 13, no. 22, p. 4706, Nov. 2021.
[54] H. Q. Nguyen et al., "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Sci. Data*, vol. 9, no. 1, p. 429, Jul. 2022.

[55] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14449–14458.
[56] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-aware trident networks for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6054–6063.
[57] K. H. Le, T. V. Tran, H. H. Pham, H. T. Nguyen, T. T. Le, and H. Q. Nguyen, "Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis," *IEEE Access*, vol. 11, pp. 14105–14114, 2023.

SHENGNAN HAO received the B.S. degree from the North China University of Science and Technology, China, in 1996, and the M.S. degree from Beijing University of Technology, China, in 2009. She joined the North China University of Science and Technology, in 1996, and became an Associate Professor, in 2009. Her current research interests include complex systems, impulsive systems, and stochastic control.

XINLEI LI was born in 1999. He received the bachelor's degree from Hebei Geo University, in 2022. He is currently pursuing the master's degree with the North China University of Science and Technology. His research interests include machine vision and graphic image processing.



WEI PENG received the B.S. degree from Hebei Medical University, in 1991, and the M.S. degree from the North China Coal Medical College, in 2005. Her current research interests include machine vision and medical image processing.

ZHU FAN was born in Baishan, Jilin. She received the bachelor's degree from the Clinical Department, North China University of Science and Technology, and the master's degree from the Respiratory Medicine Department, Tianjin Medical University. She was with the Respiratory Medicine Ward, Affiliated Hospital, North China University of Science and Technology, for five years. Currently, she is engaged in clinical treatment and teaching in the Respiratory Medicine Department. She is proficient in the diagnosis and treatment of common respiratory diseases, frequent diseases, and critical illnesses.

ZHANLIN JI (Member, IEEE) received the Ph.D. degree from the University of Limerick, Ireland, in 2010. Currently, he is a Professor with Zhejiang Agriculture and Forestry University, China; and an Associate Researcher with the Telecommunications Research Centre (TRC), University of Limerick. He was a recipient of the Irish Research Council for Science, Engineering and Technology (IRCSET) Post-Graduate Research Scholarship, in 2008; and an IRC Postdoctoral Fellowship, in 2013. He has authored/co-authored more than 100 research papers in refereed journals and conferences. His research interests include the ubiquitous consumer wireless world (UCWW), Internet of Things (IoT), cloud computing, big data management, and artificial intelligence (AI)-based image processing.

IVAN GANCHEV (Senior Member, IEEE) received the Engineering (summa cum laude) and Ph.D. degrees from Saint-Petersburg University of Telecommunications, in 1989 and 1995, respectively. He is an International Telecommunications Union (ITU-T) Invited Expert and an Institution of Engineering and Technology (IET) Invited Lecturer, currently affiliated with the University of Limerick, Ireland, University of Plovdiv "Paisii Hilendarski," Bulgaria, and Institute of Mathematics−Bulagarian Aacademy of Sciences, Bulgaria. He was involved in more than 40 international and national research projects. He has served on the TPC of more than 400 prestigious international conferences/symposia/workshops and has authored/co-authored one monographic book, three textbooks, four edited books, and more than 300 research papers in refereed international journals, books, and conference proceedings. He is on the editorial board of and has served as a guest editor for multiple prestigious international journals.