Multimedia Tools and Applications (2024) 83:39983–40001
https://doi.org/10.1007/s11042-023-17024-y

CPSNet: a cyclic pyramid-based small lesion detection network

Yan Zhu[1] · Zhe Liu[2] · Yuqing Song[2] · Kai Han[2] · Chengjian Qiu[2] · YangYang Tang[2] · Jiawen Zhang[3] · Yi Liu[2]

Received: 18 May 2023 / Revised: 7 August 2023 / Accepted: 11 September 2023 /
Published online: 6 October 2023

Abstract

The presence of small lesions is an important marker for determining whether a patient will develop malignant tumors. Clinical practitioners could easily overlook the presence of small lesions, meaning automated approaches are essential for screening test results. The use of deep learning-based detectors for this purpose has so far been suboptimal as small lesions easily lose the spatial information during the convolution operation, resulting in unsatisfactory detection accuracy and limited application in clinical decision making. In this paper, we propose a Cyclic Pyramid-based Small lesion detection Network (CPSNet), which iteratively enhances the features in the parallel layer of the Feature Parallel Network (FPN), the features learned in the loop are fused again with the initial FPN to compensate for the inadequacy problem in the initial training. In addition, we propose an aggregated dilation block (ADB) to capture small variations at different scales and a global attention block (GAB) to adaptively recalibrate the channel-based feature responses while focusing on the target spatial information and highlighting the most relevant feature channels. Extensive experiments on eight organs included in the DeepLesion dataset show that our method has a high detection accuracy(mAP=60.4) and a high overall sensitivity(80.5%), which is superior to the state-of-art methods.

Keywords Deep lesion detection · Cyclic feature pyramid · Attention mechanism · Dilated convolution · CNN

1 Introduction

The number of deaths caused by malignant tumors has increased year by year (https://www.ncc.go.jp/en/contact/index.html). Consequently, it is important that potential tumors are discovered at an early stage so that effective diagnosis and treatment can occur. Computer-aided diagnosis such as that available through imaging is an effective and common method for detecting cancer. Clinical practitioners use images from tests such as Computed Tomography

✉ Yi Liu
ly@ujs.edu.cn

Extended author information available on the last page of the article

(CT) and Magnetic Resonance Imaging (MRI) to judge the existence of tumors based on their medical knowledge and relevant laboratory report datar [1, 2]. However, it is not only onerous for clinical practitioners to conduct the massive task of analyzing CT and MRI images to detect potentially cancerous lesions, but the accuracy is affected by the experience and ability of the individuals who are performing this task. To improve this situation, previous research has focused on effectively automating and standardizing the detection of lesions in CT images to alleviate the burden on clinical staff and achieve a higher diagnostic accuracy. However, it is still difficult to effectively identify potentially cancerous small lesions in the human body.

While there have been many attempts in the field of computer vision and image recognition to use object detection techniques to detect potentially cancerous lesions [3, 4]. Recently, prior methods have been developed for lesion detection. Tao et al. [5] introduced a dual-attention mechanism to utilize 3D contextual information. A deep anchor-free one-stage volumetric lesion detector (VLD) [6] incorporates pseudo-3D convolution to recycle the architectural configurations and pre-trained weights from the off-the-shelf 2D networks. Li et al. [7] designed a Slice Attention Transformer (SATr) block that can be easily embedded into backbone to form hybrid network structures. Although these 3D methods have been greatly improved, they bring huge computing costs due to utilizing 3D volumes or multiple consecutive slices. Besides, it is difficult to obtain a high-quality 3D lesion detection dataset annotated by veteran radiologists. By contrast, the applicability of 2D detectors is more flexible. CenterNet++ [8] is a bottom-up detection method, which detects each object as a triplet keypoint, which can locate objects with arbitrary geometry and perceive global information. By decomposing the features into different frequency bands using learnable wavelets, FEDER [9] can solve the problem of intrinsic similarity between foreground and background. However, the above methods easily ignore small lesions due to feature interaction is insufficient. Although Liu et al. [10] improved small lesion detection performance by deepening the backbone network and selecting more size anchors, the performance is still suboptimal for small lesion detection in terms of insufficient feature representation.

To fuse more semantic information upon the feature pyramid, we deepened the network based on ResNet-101. After up-sampling, the output of residual blocks on each layer was fused with the high-resolution topographic map to preserve as much spatial and semantic information as possible at different scales. Subsequently, the backbone was improved by iteratively enhancing the features of the parallel layers of the FPN as part of the input features, and the features learned in the loop were fused with the initial Feature Parallel Network (FPN) again. A Multi-scale Response (MSR) block was implemented to facilitate lesion detection across fine granularity. In the MSR, an Aggregate Dilation Block (ADB) and Global Attention Block (GAB) were combined to further increase the receptive fields of top-down paths in the feature pyramid using regional correlations in each pyramidal feature generation block, which were focused on different lesion responses in the feature map. Experiments show that the accuracy of our network significantly improves o the original two-stage network. The main contributions of this work can be summarized as follows:

1) A cyclic learning method is proposed to address the problem of inadequate training of focal feature information in one-way learning during the training process of object detection network.
2) An Aggregated Dilation Block (ADB) is proposed to alleviate the shortcomings of the low-resolution feature layer in the network due to the large receptive field and fuzzy features of small lesions.

3) The Global Attention Block (GAB) is designed to reduce the influence of background noise and highlight the foreground features, which is effective for detecting obscure objects at different scales.

2 Related work

In this section, we introduce several representative deep object detection frameworks and their characteristics. Then, works about dilated convolution are introduced to enlarge receptive field. Finally, we list attention mechanism methods and summarize their drawbacks.

2.1 Object detection

The first two-stage detector R-CNN was proposed by Girshick et al. [11] by integrating segmentation algorithms [12, 13] into AlexNet [14], which classified each candidate location on the generated region proposal. Yang et al. [15–17] designed fusion operations and improved the ability of context-aware by incorporating 3D adjacent slice information. To better model long-distance feature dependency, Li et al. [7] introduced a plugand-play transformer block to form hybrid backbones. Meanwhile, one-stage detectors [18, 19] have been proposed to implement real-time object detection in recent years. Zhao et al. [20] proposed the MVP-Net, which is a multi-view FPN with a position-aware attention mechanism to assist universal lesion detection. Liu et al. [10] made improvements to the original YOLOv3 by data augmentation, feature attention enhancement and feature complementarity enhancement. CenterNet++ [8] detected each object as a triplet keypoints, which enjoys the ability in locating objects with arbitrary geometry and to perceive the global information within objects. In summary, two-stage detectors show excellent performance while single-stage network has a great advantage in speed.

2.2 Dilated convolution

Networks [14, 21, 22] reduce the resolution of input images by successive down-sampling layers as a solution to obtain global predictions with sufficient semantic representation. However, tasks like object detection require higher resolution output in order to identify and localize object instances. Recently, methods involving dilated convolution [23, 24] have shown better performance in the object detection task with the aim of extracting image features over a larger perceptual field without loss of resolution. Yu et al. [23] developed an extended convolution-based module that combines multi-scale contextual information for semantic segmentation. DeepLabs [25, 26] proposed a module using cascaded or parallel extended convolution to further improve the performance of target segmentation on multiple scales. By setting different dilation rates, contextual information at multiple scales could be captured. In order to expand the perceptual field while maintaining the spatial dimension of the feature map, the backbone network of DetNet [27] uses expanded convolution to significantly improve the detection accuracy of large objects. Although the methods mentioned above all use dilated convolution to increase the size of the perceptual field, our proposed ADB module makes full use of multiple branching dilated convolution outputs to enhance the detection of multiple scale lesions better.

2.3 Attentional mechanism

In artificial neural networks, the attention mechanism generally refers to focused attention to improve the efficiency of the network. Bahdanau et al. [25] utilized an attention-like mechanism to simultaneously translate and align machine translation tasks, allowing the application of attention mechanisms to the field of natural language processing. Cheng et al. [26] proposed intra-attention to focus on all positions in a sequence to obtain a response at a position in the sequence. Vaswani et al. [28] further argued that machine translation models can achieve superior performance through self-attentiveness. Nonlocal neural network (NLNet) [29] was designed to model pixel-level pairwise relations with an attention mechanism. Based on NLNet, Zhang et al. [30] proposed a Self-Attentive Generative Adversarial Network (SAGAN) that allows attention-driven remote dependency modeling for image generation tasks. Additionally, [31] obtained the feature weights of each channel in the feature map by global average pooling, which enables the model to give different attention to each channel in the feature map. Most of the above approaches recalibrate the feature maps by assigning attention weights or focus only on the location information of the object, without a comprehensive combination to increase the semantic attention and location attention to small objects.

3 Methods

The structure of proposed framework is illustrated in Fig. 1. We respond to the difficulty of small lesion detection with a series of improvements. Many small lesions are not obvious in CT images, we thus make them clearer through mathematical morphology. Additionally, since the network is one-way learning during training, there is inadequate training of lesion feature information. By using a cyclic pyramid structure, the network can recursively learn feature information (Section 3.2). Since the network is unable to actively learn the lesion region, we then feed the output features of each Res-Block layer into the MSR module (including ADB, GAB) to enhance the network's ability to actively focus on the lesion region (Sections 3.3 and 3.4). The output of the MSR is subsequently fed into the RPN, and the network undergoes Softmax loss (detection of classification probability) and Smooth L1 loss (detection of frame regression) to train classification probability and bounding box regression.

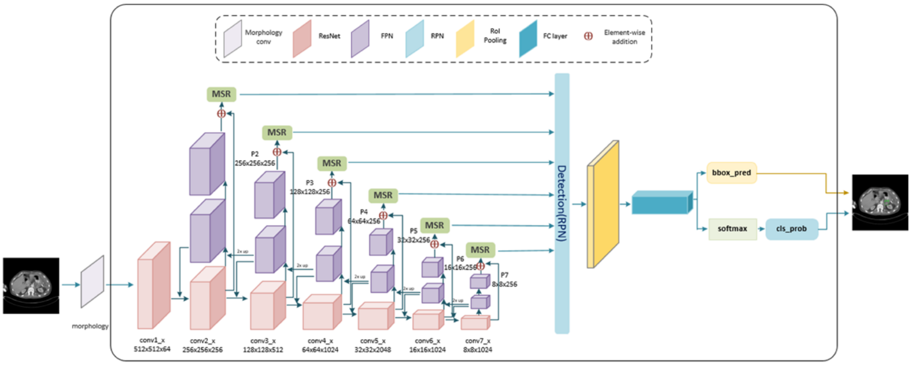


Fig. 1 The framework of the proposed lesion detection method

3.1 Feature extraction network

Faster R-CNN [32] is a classic two-stage object detection algorithm. Our algorithm, based on Fast R-CNN introduces the RPN, a region proposal box generation algorithm, which greatly improves object detection speed. The detection part is divided into four steps: i) input the whole image into CNN for feature extraction; ii) generate the anchor box by RPN; iii) the RoI pooling layer makes each RoI generate a feature map with a fixed size; iv) Softmax Loss and Smooth L1 Loss are used for classification and Bounding box regression, respectively.

In a traditional Convolution Neural Network (CNN), the feature map goes through multiple down-sampling operations from the input network to the output process, and the spatial information reduces as the network deepens, which means that the features are no longer recognizable and thus leads to poor detection of small objects by the traditional CNN.

To address this issue, the backbone is deepened to enhance the network's ability to extract richer semantic information. Conv2_x, conv3_x, conv4_x, conv_x, conv6_x, and conv7_x blocks are used to build the feature pyramid, the corresponding layers of which are P2, P3, P4, P5, P6 and P7. The corresponding bottom-up feature maps are convolved by 1 × 1 kernel to reduce the number of channels. Feature maps from P2-P4 can help the network find and locate small lesions. With the deepening of the network, information from small objects will be dismissed because of the down-sampling operation. Conv6_x and onv7_x help to bring deep semantic information into the higher resolution feature map in the upper layer, by improving the feature extraction capability of the backbone, and fuse with the upper layer feature map after up-sampling when constructing FPN.

3.2 Cyclic feature pyramid network

Since the one-way learning of the object detection network in the training process provides inadequate training of focal feature information, a cyclic learning approach is proposed to provide secondary learning of the image feature map. Through this repeated learning approach effective features that have not been fully learned are extracted. The Cyclic Feature Pyramid (CFP) architecture consists of three parts: a principal backbone and two auxiliary FPNs (F1, F2), as shown in Fig. 2. Among them, the backbone consists of ResNet with each stage consisting of several convolutional layers, and the two auxiliary FPNs are composed of the same Feature Pyramid built based on the backbone.

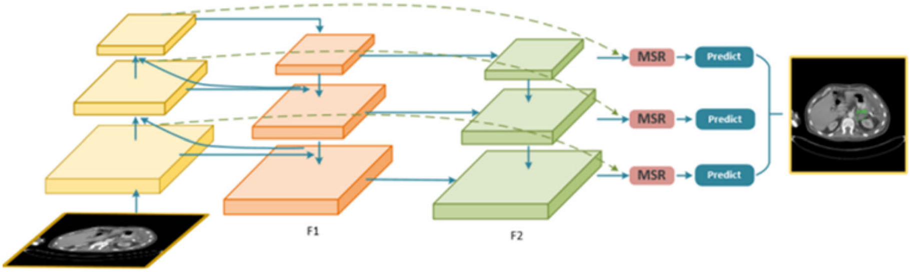


Fig. 2 The structure of CFP based on the proposed MSR module

In a traditional convolution network with only one backbone, the i_th stage takes the output (x_i) of the previous ith stage as input, and we denote the bottom-up i th stage by H_i, this process can be expressed as formula (1):

$$x_i = H_i(x_{i-1}), \quad i \geq 2 \tag{1}$$

In contrast, in CFP, we have innovatively adopted an auxiliary FPN (F1) to strengthen the backbone by iteratively feeding the output features of the feature pyramid (F1) as part of the input features to the backbone of the network in a stage-by-stage manner. More specifically, the input of the i-th stage of the backbone is a fusion of the output of the previous $i-1$ th stage of the backbone and the i th stage features of the parallel FPN. Where, ∀i ∈ $\{1, \ldots, I\}$, f_i denotes the i th feature output of RFP is shown in the following equation:

$$\begin{aligned} f_i &= F_i(f_{i+1}, x_i) \\ x_i &= H_i(x_{i-1}, f_i) \end{aligned} \tag{2}$$

where F_i denotes the up-sampling and feature fusion operations in the i th stage of the backbone corresponding to the feature pyramid. CFP is a network architecture implemented based on this recursive iterative operation, ∀i ∈ $\{1, \ldots, I\}$,$\forall i = \in \{1, \ldots, S\}$, $c = 1, \ldots, C$:

$$\begin{aligned} f_{c,i} &= F_{c,i}(f_{c,i+1}, x_{c,i}) \\ x_{c,i} &= H_{c,i}(x_{c,i-1}, f_{c-1,i}) \end{aligned} \tag{3}$$

where C is the count of unfolding iterations, and we use the superscript c to denote the c-th operation in the unfolding step. Based on this, we improve the ResNet backbone to allow it to use both the front layer input x and the parallel layer feedback f of the FPN as input features.

For the object detection task, we first build feature pyramids F1 and F2 based on the P2-P7 layers of the backbone. The features in each layer of the F1 pyramid are fused from the previous feature layer after up-sampling and the backbone parallel layer features, we feed it into the original backbone parallel layer to learn again, so that the wrong information in the backpropagation of the object detection can be relearned and adjusted in time when it is passed in the second cycle, while the features of the lesion area are relearned through the cyclic pyramid structure to strengthen the sensitivity of the network to lesion features.

3.3 Aggregated dilation block

In the process of generating feature pyramids based on residual blocks, the imbalanced problem between spatial and semantic information appears. To this end, we build a feature pyramid network constructed by multiple scale output of res-block in the top-down pathway. Dilated convolution is introduced in the ADB module by using a multi-branch structure to adapt to the receptive field of feature maps with multi-scales through different dilation rates. In each parallel dilated convolution branch, the feature map is enhanced by the cascade convolution kernels with different dilation rates. After the convolution of each layer, the output is non-linearized by the activation function to prevent gradient explosion and bring more differential representations for feature transformation. To some extent, weighted combinations in the multi-branch dilated convolution process could eliminate the noise left behind in low-resolution images. Then the output features of each branch with the original image are concatenated and get an aggregated feature map. The feature map output by the ADB module has a larger receptive field.

In ADB module, $f \in R^{W \times D}$ is used to describe the architecture of this module, where W and D represent the width and depth of ADB, respectively. The dilation rate of specific layer in ADB is expressed as f_{ij}, where $i = 1, 2..., W$ and $j = 1, 2, ..., D$ represent the index of width and depth, respectively. The aggregated dilated operation is shown as follows:

$$\mathcal{F}(x) = \sum_{i=1}^{W} \mathcal{T}_i \left(x \mid f_{i1}, f_{i2}, \ldots, f_{iD}\right) \tag{4}$$

where $T_{i(x)}$ represents the cascade-transformation.

As shown in Fig. 3, the parallel structure branch inside the ADB module is connected in series with convolution kernels with different dilation rates, and the output multi-scale feature map restores more detailed spatial information of the instance. It also provides more long-range context information for the construction of a feature pyramid. The receptive field of each layer is expressed as follows:

$$\mathcal{A}_{i0} = 1 \tag{5}$$

$$\bar{k}_{ij} = k_{ij} + \left(k_{ij} - 1\right) \times \left(f_{ij} - 1\right) \tag{6}$$

$$\mathcal{A}_{ij} = r_{i,j-1} \times k_{ij} - \left(k_{ij} - 1\right) \times \left(\mathcal{A}_{i,j} - \prod_{k=1}^{j-1} s_k\right) \tag{7}$$

where A_{ij} denotes the receptive field, k_{ij} denotes the kernel size and represents s_k the stride.

From the formulas, the size of the receptive field extracted by the convolution kernel with different dilation rates is also different. Usually in the feature extraction network, that is, in the backbone, dilated convolution helps us identify the large object from the enlarged receptive field [27]. However, we add dilated convolution follows the output of the feature pyramid, expecting to provide more context spatial information to improve the detection accuracy of small lesions.

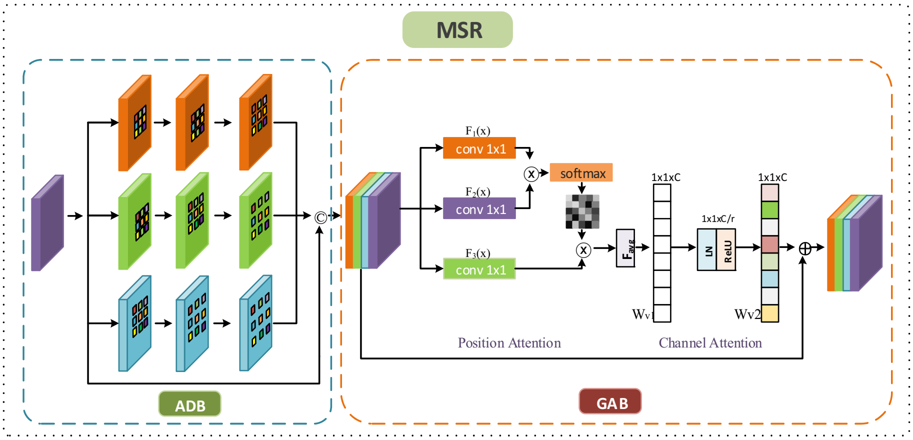


Fig. 3 The structure of MSR module. The detailed architecture of the Multi-Scale Response (MSR) module consists of two parts: Aggregated dilated block (ADB) in the blue box and Global attention Block (GAB) in the orange box

3.4 Global attention block

The traditional attention mechanism allows the network to focus on the object region during the training process, but during the learning process for small objects, the spatial information of the object will be lost at a deeper level with the down-sampling effect of the network, and the effect of the attention mechanism disappears at that time.

Based on this deficiency, we propose a Global Attention Block (GAB), which allows the network to focus more on spatial features as well as channel features. Each GAB consists of two parts: a spatial attention block and a channel attention block. Object detection needs to be extremely sensitive to changes in spatial location, so our proposed spatial attention block uses a self-attentive mechanism to model remote dependencies which enhances the network's global understanding of the visual scene. In addition, inspired by SENet [31], a channel attention block is introduced, which aims to focus on the feature information we need. The input feature map x is converted into three paths: F_1, F_2 and F_3, where$F_h(x) = W_h x$, ∀h $\in \{1, 2, 3\}$. Firstly, obtaining the attention map of the long-range correlation between each position in the feature map through S_{ij}, where $S_{ij} = F_1\left(x_i\right)^T \otimes F_2\left(x_j\right)$. S_{ij} is transformed into A_{ij} by $softmax$, where $A_{ij} = softmax(s_{ij})$ represents the relationship between the position of i and j in the feature map, and then A_{ij} and $F3$ are multiplied to query the response relationship between pixels on the feature map.

$$z_i = \sum_{j=1}^{H \times W} A_{i,j} \otimes F_3\left(x_j\right) \tag{8}$$

where i and $\otimes$ denote the index of query position and matrix multiplication, respectively.

After the spatial attention block, we compress the global information into channels through global average pooling, the main difference between the SE block and GAB is the fusion module, which reflects the goals. The SE block uses re-adjustment to re-calibrate the importance of the channel, but it does not fully simulate the long-range correlation. The long-range correlation is captured by using addition to aggregate the global context to all positions. The detailed architecture of the GAB is formulated as follows:

$$y_i = x_i + W_{v2}\,\mathrm{ReLU}\left(LN\left(W_{v1} z\right)\right) \tag{9}$$

where, $W_{v1} \in R^{\frac{C}{r} \times C}$, $W_{v2} \in R^{C \times \frac{C}{r}}$ In order to obtain the lightweight attribute of the channel attention block, the parameters of the module are changed from C to C/r. Where r is the bottleneck ratio, setting r too large will lose feature information, while too small will consume a lot of computation, so it is necessary to strike a balance between two costs, and we found when $r = 4$, the model performs best.

4 Experiments

4.1 Dataset

The DeepLesion dataset is the largest open dataset of multi-category, lesion-level labeled clinical medical CT images ever published by the NIH Clinical Center. By training deep neural networks on this dataset, it will be possible to obtain a large-scale universal lesion detector that can more accurately and automatically measure the size of all lesions in the patient's body, allowing initial assessment of cancer system-wide. The dataset contains 32,735 labeled lesion instances from 4,427 independent, anonymous patients. The dataset covers a wide range of lesions involving the lung (LU), liver (LV), mediastinum (ME), kidney (KD), pelvis (PV), bone (BN), abdomen (AB) and soft tissues (ST). We used 70% samples of the dataset for training, 15% for validation, and 15% for testing.

Traditional image pre-processing operations only flip, offset and crop the image, but this does not help for detection of small objects. The features of small lesion regions must become larger and sharper to be more easily detected. Therefore, we took advantage of binary image processing. By constructing corresponding mathematical morphological structure elements suitable for the DeepLesion dataset, the maximum and minimum areas can be efficiently found in the image, and the noise in the CT scan can be reduced making the lesion more visible. The DeepLesion dataset, is roughly divided into two categories: dark background with light lesion areas and the opposite, of which we selected example CT images of the lung and liver organs. Figure 4 shows the images obtained after different morphological processing methods on light and dark representations of tumors. In CT images with a dark background, the expansion operation enlarges the lesion area and facilitates network detection of small lesions, while in images with a light background, the tumor part of the image after erosion processing not only becomes larger, but also retains a large amount of texture information.

4.2 Evaluation metric

We chose two evaluation metrics in our subsequent ablation and comparison experiments. One metric is mean Average Precision (mAP) when Intersection over the Union threshold = 0.5, which is used to measure object detection accuracy. Another metric is the average sensitivity values at different false positive rates (FROC) of the whole testing set.

4.2.1 Mean average precision

The Average Precision (AP) is defined as the approximate area under the precision-recall (PR) curve of a certain class. Hence, mean Average Precision (mAP) is the mean value of APs added up by each class. For single-class detection tasks, mAP is equal to AP. The all-points interpolation method, suggested by PASCAL VOC 2012 and extensively adopted will be used

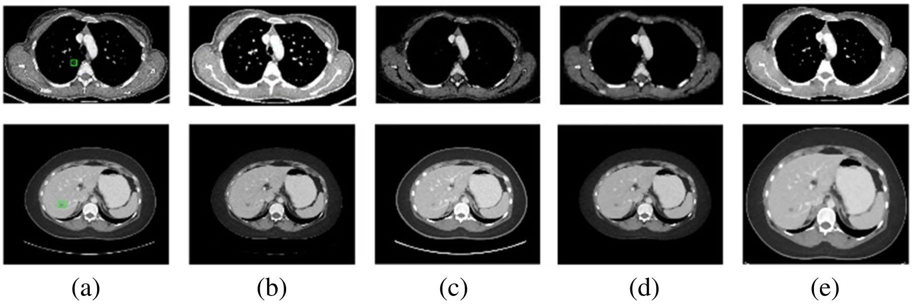


Fig. 4 Output of different colors tumors after morphological operations. Organs with representative light and dark colors of lesions in the dataset are shown. Where, (a)-(e) represent the labeled image, erosion, dilation, opening and closing respectively. First row is lung CT image and second row is liver CT image

to scatter and finally draw the PR curve, the progressive recall value and its corresponding precision value which should be calculated, according to all positive predicted boxes with confidence score sorted from high to low. The paired recall and precision values are calculated from the current number of true positives (TPs) and false positives (FPs).

$$\begin{aligned} \text{precision } &= \frac{TPs}{TPs + FPs} = \frac{TPs}{\text{All Detections}} \\ \text{recall } &= \frac{TPs}{\text{All Ground Truths}} \\ \text{with } P &= \begin{cases} TP, \text{if IOU}(p, GT) \geq \text{ threshold} \\ FP, \text{if IOU } (p, GT) < \text{ threshold} \end{cases} \end{aligned} \tag{10}$$

where, $IOU(p, GT)$ stands for the intersection over union between certain predicted box p and ground truth box GT, while the threshold is set to 0.5 by default.

$$\begin{aligned} AP &= \sum_{r=0}^{1} (r_{n+1} - r_n) * p_{\text{interp }} (r_{n+1}) \\ \text{with } \quad p_{\text{interp}} (r_{n+1}) &= \max_{\hat{r}:\hat{r} \geq r_{n+1}} p(\hat{r}) \\ mAP &= \frac{1}{N_c} \sum_{c=0}^{N_c-1} AP_c \end{aligned} \tag{11}$$

where, the middle equation indicates searching for the precision envelop at the right side of recall point, to gradually obtain the final estimated area under the PR curve.

4.2.2 Sensitivity at various FPs per image

Different from the constant AP metric, another stricter metric we employ is the sensitivity (recall) at various FPs per image. As its name implies, it is a metric to evaluate the capability of detector under various strict levels. To implement this, one should set a different confidence threshold, to distinguish positive samples from negative ones before doing non-maximum suppression (NMS). For example, if the confidence threshold is set from 0.001 to 0.01, there would be an increase of recall and a reduction in precision due to a change in TPs and FPs (10). Hence, by continuously changing the threshold, we can obtain recall values under different FPs per image. The metric fixes FPs per image (usually 0.5, 1, 2, 4 and 8), to see whether the detector could find more true positives under the same fault tolerance.

4.3 Implementation details

Experiments were conducted on a Workstation with IntelCore i7, 2.7GHz CPU, 8GB RAM under Ubuntu 18.4, and an NVIDIA GTX 2080 video processing card with 11GB memory. We set training learning rate to 0.008 and training momentum to 0.9; the training batch was 128, the mini batch was 2; and the learning process was 12 epochs. The initialization weights for P1-P5 are set according to the ImageNet pre-trained model. For the deepened network part, we randomly initialize parameters. The input images are resized to 512 × 512. The optimization algorithm was stochastic gradient descent (SGD), which took about 60 hours to train our detector.

4.4 Experimental results

4.4.1 Sensitivity at various FPs per image

In this section, we performed a comprehensive experiment to analyze the effectiveness of the proposed method. Table 1 compared the detection results of Faster-RCNN [32], Faster-RCNN+CFP, Faster-RCNN+CFP+ADB, Faster-RCNN+CFP+GAB and Faster-RCNN+CFP+MSR on the DeepLesion dataset. As seen for the detection task on eight organs in the dataset, Faster-RCNN+CFP with a cyclic pyramid structure gains some accuracy on top of Faster-RCNN, with 6% improvement in mAP, which reflects that the idea of repeated learning of features is necessary. On this basis, we added the ADB module, and Tables 1 and 2 show that the mAP improved by 3.8 percentage points based on the recall improvement, especially for the AP improvement of small nodal organs is the most obvious, in order to prove the effectiveness of ADB from several indicators, in Table 3 we have improved the Map by the average sensitivity values at different false positive rates to demonstrate the effectiveness of our module. From Figs. 5 and 6, by introducing PR Curve and FROC Curve, we can visualize the effectiveness of our proposed CFP, ADB, and GAB module. Last, we show the experiment of adding Global Attention Block (GAB), see the last row of Table 1, where the fusion module of ADB and GAB are collectively called Multi-Scale Response (MSR), it increased by more than 2%. In summary, the proposed method improves by 12% compared to the original Faster R-CNN.

While the traditional detection method has a high accuracy for general detection, each module presented here addresses the problem of insufficient detection of small objects in the original model. The ADB module improves the sensitivity of multiple scales to the object based on the global attention mechanism to focus on learning the small object features captured in the previous stage. The increase of mAP in Table 1, we can observe the effectiveness of the MSR module, which helps improve the original network. From Fig. 7, it can be seen that our proposed method has better results for all sizes of lesions under the premise of better detection for small lesions.

The proposed network consists of four main components: Faster R-CNN, CFP, ADB and GAB. In Tables 1, 2 and 3, Baseline denotes the original Faster R-CNN model, and Baseline + CFP denotes the method of circular pyramids mentioned in Section 3.2. To assess the validity of each module, we performed ablation studies on the DeepLesion dataset. From Table 1, we can see that the original model was improved by more than 12% on mAP. Based on the coarse lesion types provided by DeepLesion for each CT slice, we calculated the AP for each lesion type. Besides, the table shows that the AP values were increased by different magnitudes for different sites. Table 3 shows the average sensitivity values for the entire test

Table 1 mAP and AP of each lesion type on the official split test set of DeepLesion

Methods	Total	BN	AB	ME	LV	LU	KD	ST	PV
Baseline	0.484	0.524	0.391	0.512	0.549	0.582	0.419	0.436	0.368
Baseline+ CFP	0.541	0.525	0.508	0.544	0.556	0.589	0.435	0.445	0.411
Baseline+ CFP + ADB	0.579	0.529	0.531	0.559	0.571	**0.667**	0.522	0.465	0.527
Baseline+ CFP + GAB	0.560	0.527	0.534	0.554	0.569	0.644	0.520	0.451	0.509
Baseline+ CFP + MSR	**0.604**	**0.542**	**0.539**	**0.575**	**0.577**	0.665	**0.538**	**0.476**	**0.541**

Bold entries indicate the best performance

Table 2 Recall and mAP on the official split test set of DeepLesion

Methods	recall	mAP
Baseline	0.836	0.484
Baseline+ CFP	0.849	0.541
Baseline+ CFP + ADB	0.851	0.579
Baseline+ CFP + GAB	0.841	0.520
Baseline+ CFP + MSR	**0.869**	**0.604**

Bold entries indicate the best performance

set at different false positives rates. Through comparison between different configurations, the proposed method achieves the highest sensitivity at different false positives rates. We plotted the FROC curves to make the results more intuitive, see Fig. 6.

4.4.2 Comparison with other methods

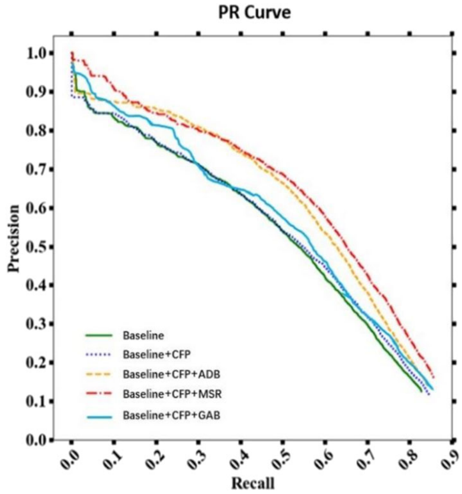
In this section, we compared our proposed detector with other state-of-the-art detectors by comprehensively analyzing the detection accuracy tradeoff on the DeepLesion dataset. Table 4 shows the detection sensitivity of each method at different FPs for each image. Typical two-stage detectors, i.e., Mask R-CNN [34], which have completely inferior detection sensitivity per FP per image than ours. Other single-stage methods like RetinaNet [19] and YOLOv3 [33], outperform our method in terms of detection efficiency, but our accuracy is far superior to them. The most challenging counterparts are ULDor [35] and 3DCE [36], which also use a two-stage pipeline in their network. ULDor [35] uses pseudo-masking and hard negative example mining strategies to improve the accuracy. Unfortunately, it does not give much attention to the issue of scale imbalance as well as attention mechanism, and they cannot detect lesions well when the proposal is very similar in appearance to its surrounding tissue, thus the result is still not as good as ours. In 3DCE [36], 3D contextual information is introduced during training and testing than us. Most of these sliced feature maps can be cached into memory and reused for the next inference. 3DCE [36] achieved the best results among the different methods using 27 input slices. Its sensitivity at 8 FP per image (a common comparison criterion) is 89.1%, while ours is 89.4%, indicating that the proposed detector still performs better than this method, although we do not take the strategy of 3D context enhanced strategy. Figure 8 shows the visual comparison of our proposed deep learning network with the state-of-the-art proposed detection methods. To make the comparison more intuitive, we draw the FROC curves of several methods at the top of Fig. 9, from which we can see clearly that the sensitivity of our method is better than other methods.

Table 3 Ablation w.r.t. Sensitivity(%) at various FPs per image on the official split test set of DeepLesion

Components	0.5	1	2	4	8
Baseline	0.531	0.635	0.730	0.814	0.850
Baseline+ CFP	0.546	0.657	0.739	0.820	0.867
Baseline+ CFP + ADB	0.614	0.682	0.767	0.841	0.876
Baseline+ CFP + GAB	0.630	0.701	0.752	0.857	0.866
Baseline+ CFP + MSR	**0.661**	**0.754**	**0.831**	**0.883**	**0.894**

Bold entries indicate the best performance

Fig. 5 Ablation PR Curve on test set of official split test set of DeepLesion



The comparison of mAP and APs is shown in Table 5, and similar phenomena and trends can be found. In particular, paper [36] proposed a detection method based on 3D slicing, although it sacrificed a lot of time and resources in the training process, our mAP (60.4%) was still much higher than its (54.4%). Secondly, our detector performed slightly worse than 3DCE with 9 slices in detecting lesions of mediastinum, but again had excellent realizations on other sites. In conclusion, compared with other state-of-the-art detectors, our proposed

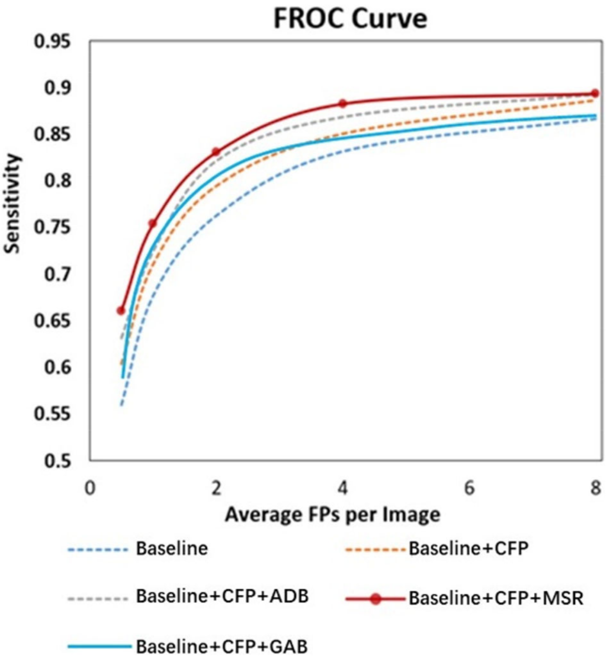


Fig. 6 Ablation FROC Curve on test set of official split test set of DeepLesion

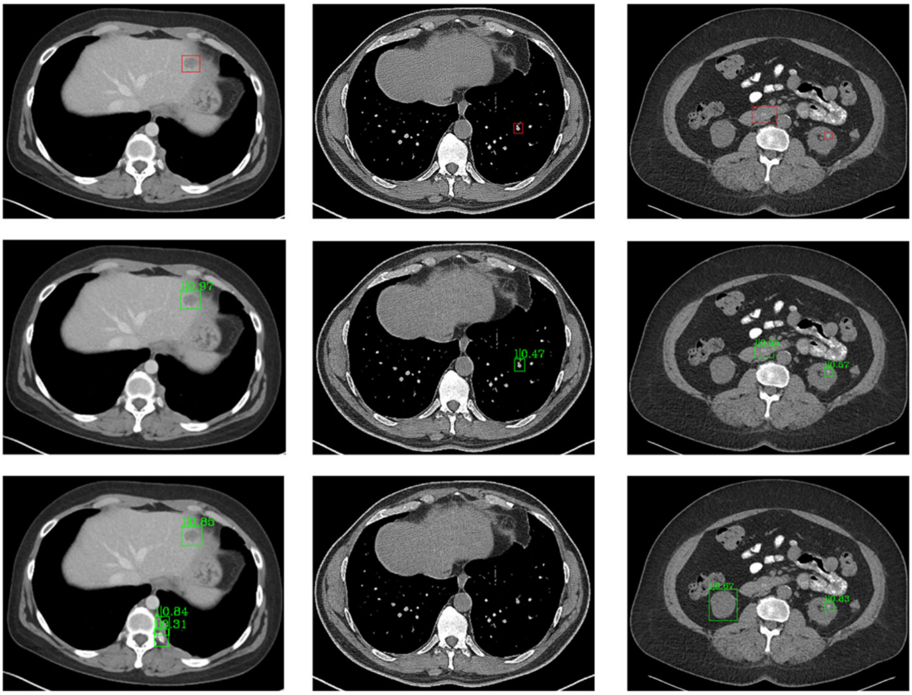


Fig. 7 Lesion detection results for sample CT images of various methods. Each row from top to bottom represents the label image, our proposed method, Faster R-CNN, respectively

Table 4 Comparison of the proposed method with state-of-the-art methods on the DeepLesion test set

Methods	FPs per image					Average
	0.5	1	2	4	8	
RetinaNet [19]	0.458	0.542	0.625	0.698	0.742	0.595
YOLOv3 [33]	0.520	0.626	0.719	0.795	0.843	0.700
Mask R-CNN [34]	0.398	0.527	0.656	0.777	0.852	0.642
ULDor [35]	0.529	0.648	0.748	0.844	0.861	0.726
3DCE, 3, slices [36]	0.569	0.673	0.756	0.816	0.858	0.734
3DCE, 9, slices [36]	0.593	0.707	0.791	0.843	0.878	0.762
3DCE, 27, slices [36]	0.625	0.733	0.807	0.857	0.891	0.782
Liu et al. [10]	0.633	0.704	0.787	0.851	-	0.744
CenterNet++ [8]	0.648	0.739	0.799	0.868	0.883	0.787
DN-DETR [37]	0.652	0.732	0.808	0.869	**0.898**	0.792
ours	**0.661**	**0.754**	**0.831**	**0.883**	0.894	**0.805**

Lesion detection sensitivity values are reported at different false positive (FP) rates
Bold entries indicate the best performance

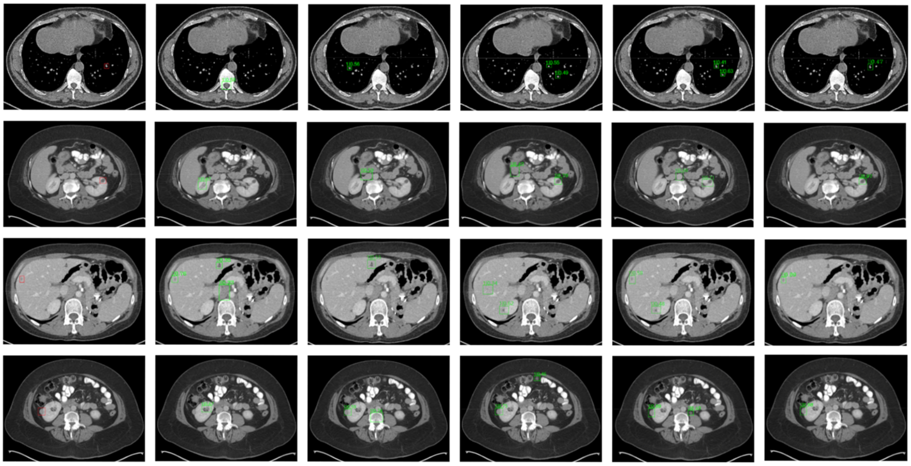


Fig. 8 Visual comparison of our proposed deep learning network with state-of-the-art detection methods, experimental results on DeepLesion dataset. From left to right, Ground Truth, RetinaNet, YOLOv3, Mask R-CNN, ULDor and our proposed method

method has advantages in terms of detection accuracy and sensitivity, with the premise that it performs well in detecting normal size lesions, it also meets the need for better detection of small lesions in the medical field.

Our CPSNet shows better performance for medical lesion detection, especially for the detection of small lesions. However, our method embraces one shortcoming: the MSR module introduces multi-branch dilatation convolution operators and attention mechanisms, which undoubtedly increase the computing costs and reasoning time.

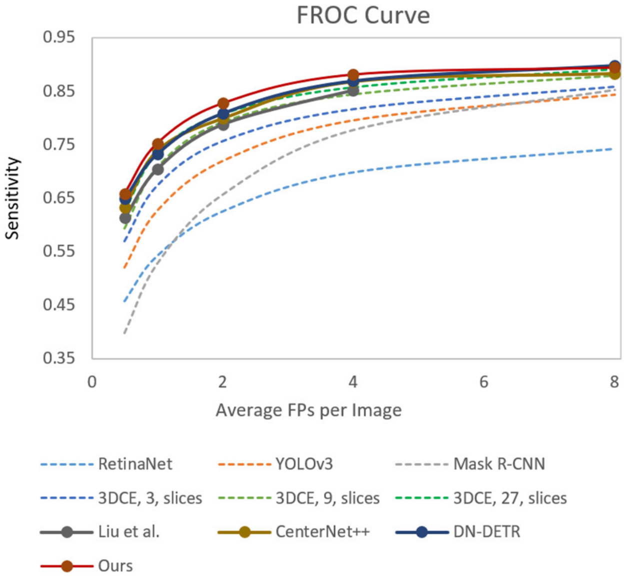


Fig. 9 FROC Curves of various methods

Table 5 Comparison of the proposed method with state-of-the-art methods on the DeepLesion test set

Methods	Total	BN	AB	ME	LV	LU	KD	ST	PV
YOLOv3 [33]	0.468	0.371	0.372	0.587	0.524	0.562	0.373	0.403	0.435
RetinaNet [19]	0.510	0.539	0.430	0.555	0.524	0.612	0.424	0.455	0.421
3DCE, 3 slices [36]	0.506	0.434	0.424	0.522	0.543	0.633	0.426	0.421	0.423
3DCE, 9 slices [36]	0.544	0.492	0.468	**0.577**	0.564	0.663	0.480	0.441	0.470
Ours	**0.604**	**0.542**	**0.539**	0.575	**0.577**	**0.665**	**0.538**	**0.476**	**0.541**

Lesion detection sensitivity values are reported at different false positive (FP) rates
Bold entries indicate the best performance

5 Conclusion

In this paper, we proposed a cyclic pyramid-based small lesion detection network to enhance the detection of lesions on feature maps of different sizes. This encompasses an ADB module to augment the detector's awareness of feature map scale variation - providing finer size estimates of the feature map to capture the response to scale under different receptive fields, and a GAB module to effectively choose meaningful responses. Extensive experiments on the DeepLesion dataset showed that: our CPSNet has a 60.4 mAP value and a 80.5% overall sensitivity, which is superior to the state-of-art methods. Due to the fact that multi-branch dilatation convolution operators and attention mechanisms are introduced into our framework, which decreases the reasoning speed. Thus, our future work is to designed a more lightweight network through knowledge distillation. Besides, we also utilize Neural Architecture Search NAS to explore a more suitable backbone architecture for lesion detection.

Acknowledgements The authors would like to thank radiologists of the Medical Imaging Department of Affiliated Hospital of Jiangsu University. This work was supported by the National Natural Science Foundation of China (61976106); China Postdoctoral Science Foundation (2017M611737); Six talent peaks project in Jiangsu Province (DZXX-122); Key RESEARCH and development program for social development (SH2021056); Zhenjiang city key research and development plan (SH2020011); Jiangsu province emergency management science and technology project (YJGL-TG-2020-8).

Declarations

This paper is the expanding version of conference paper that was published in International Conference on Artificial Intelligence and Security. The corresponding reference is Tang, Y., Liu, Z., Song, Y., Han, K., Su, J., Wang, W., ... & Zhang, J. Automatic CT Lesion Detection Based on Feature Pyramid Inference with Multi-scale Response In International Conference on Artificial Intelligence and Security, pp. 167-179, Springer, Cham, 2021.

Conflict of interest The authors declare no confict of interest.

References

1. Guo K, Chen T, Ren S, Li N, Hu M, Kang J (2022) Federated learning empowered real-time medical data processing method for smart healthcare. IEEE/ACM Trans Comput Biol Bioinforma
2. Guo K, Shen C, Hu B, Hu M, Kui X (2022) Rsnet: relation separation network for few-shot similar class recognition. IEEE Trans Multimed
3. Lee S-g, Bae JS, Kim H, Kim JH, Yoon S (2018) Liver lesion detection from weakly-labeled multi-phase ct volumes with a grouped single shot multibox detector. In: Medical image computing and computer

assisted intervention-MICCAI 2018: 21st international conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11, pp 693–701. Springer

4. Chiao J-Y, Chen K-Y, Liao KY-K, Hsieh P-H, Zhang G, Huang T-C (2019) Detection and classification the breast tumors using mask r-cnn on sonograms. Medicine 98(19)
5. Tao Q, Ge Z, Cai J, Yin J, See S (2019) Improving deep lesion detection using 3d contextual and spatial attention. In: Medical image computing and computer assisted intervention-MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, pp. 185–193. Springer
6. Cai J, Yan K, Cheng C-T, Xiao J, Liao C-H, Lu L, Harrison AP (2020) Deep volumetric universal lesion detection using light-weight pseudo 3d convolution and surface point regression. In: Medical image computing and computer assisted intervention-MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23, pp 3–13. Springer
7. Li H, Chen L, Han H, Kevin Zhou S (2022) Satr: slice attention with transformer for universal lesion detection. In: International conference on medical image computing and computer-assisted intervention, pp 163–174. Springer
8. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2022) Centernet++ for object detection. arXiv preprint arXiv:2204.08394
9. He C, Li K, Zhang Y, Tang L, Zhang Y, Guo Z, Li X (2023) Camouflaged object detection with feature decomposition and edge reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 22046–22055
10. Liu Z, Han K, Xue K, Song Y, Liu L, Tang Y, Zhu Y (2022) Improving ct-image universal lesion detection with comprehensive data and feature enhancements. Multimedia Systems 28(5):1741–1752
11. Girshick R, Donahue J, Darrell T, Malik, J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
12. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. Int J Comput Vis 104:154–171
13. Zitnick CL, Dollár P (2014) Edge boxes: locating object proposals from edges. In: Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, pp 391–405. Springer
14. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
15. Yang J, He Y, Huang X, Xu J, Ye X, Tao G, Ni B (2020) Alignshift: bridging the gap of imaging thickness in 3d anisotropic volumes. In: Medical image computing and computer assisted intervention-MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23, pp 562–572. Springer
16. Yang J, He Y, Kuang K, Lin Z, Pfister H, Ni B (2021) Asymmetric 3d context fusion for universal lesion detection. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, pp 571–580. Springer
17. Yang J, Huang X, He Y, Xu J, Yang C, Xu G, Ni B (2021) Reinventing 2d convolutions for 3d images. IEEE J Biomed Health Inform 25(8):3009–3018
18. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
19. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
20. Zhao P, Li H, Jin R, Zhou SK (2023) Diffuld: diffusive universal lesion detection. arXiv preprint arXiv:2303.15728
21. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
22. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
23. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122
24. Liu S, Huang D et al (2018) Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision (ECCV), pp 385–400
25. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
26. Cheng J, Dong L, Lapata M (2016) Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733

27. Li Z, Peng C, Yu G, Zhang X, Deng Y, Sun J (2018) Detnet: A backbone network for object detection. arXiv preprint arXiv:1804.06215
28. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
29. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
30. Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. In: International conference on machine learning, pp 7354–7363. PMLR
31. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
32. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28
33. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767
34. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
35. Tang Y-B, Yan K, Tang Y-X, Liu J, Xiao J, Summers RM (2019) Uldor: a universal lesion detector for ct scans with pseudo masks and hard negative example mining. In: 2019 IEEE 16th International symposium on biomedical imaging (ISBI 2019), pp 833–836. IEEE
36. Yan K, Bagheri M, Summers RM (2018) 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. In: Medical image computing and computer assisted intervention-MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part I, pp 511–519 . Springer
37. Li F, Zhang H, Liu S, Guo J, Ni LM, Zhang L (2022) Dn-detr: accelerate detr training by introducing query denoising. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13619–13627
38. Yan K, Wang X, Lu L, Summers RM (2017) Deeplesion: automated deep mining, categorization and detection of significant radiology image findings using largescale clinical lesion annotations. arXiv preprint arXiv:1710.01766

Authors and Affiliations

Yan Zhu[1] · Zhe Liu[2] · Yuqing Song[2] · Kai Han[2] · Chengjian Qiu[2] · YangYang Tang[2] · Jiawen Zhang[3] · Yi Liu[2]

Yan Zhu
salary_hi@126.com

Zhe Liu
lzhe@ujs.edu.cn

Yuqing Song
yqsong@ujs.edu.cn

Kai Han
2112108003@stmail.ujs.edu.cn

Chengjian Qiu
2111908005@stmail.ujs.edu.cn

YangYang Tang
2221908037@stmail.ujs.edu.cn

Jiawen Zhang
zhangjw2000@126.com

[1] Department of Imaging, Affiliated Hospital of Jiangsu University, Zhenjiang, China

[2] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China

[3] Department of Radiology, Fudan University, Shanghai, China