Computers in Biology and Medicine 163 (2023) 107149

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

EFPN: Effective medical image detection using feature pyramid fusion enhancement

Zhenghua Xu [a,*], Xudong Zhang [a], Hexiang Zhang [a,*], Yunxin Liu [a], Yuefu Zhan [b,*], Thomas Lukasiewicz [c,d]

[a] State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, Tianjin, China
[b] Department of Radiology, Hainan Women and Children's Medical Center, Haikou, China
[c] Institute of Logic and Computation, TU Wien, Vienna, Austria
[d] Department of Computer Science, University of Oxford, Oxford, United Kingdom

ARTICLE INFO

Keywords:
Medical image detection
Enhanced feature pyramid network
Deep and diverse multi-scale feature
Weighted multi-scale feature fusion

ABSTRACT

Feature pyramid networks (FPNs) are widely used in the existing deep detection models to help them utilize multi-scale features. However, there exist two multi-scale feature fusion problems for the FPN-based deep detection models in medical image detection tasks: insufficient multi-scale feature fusion and the same importance for multi-scale features. Therefore, in this work, we propose a new enhanced backbone model, EFPNs, to overcome these problems and help the existing FPN-based detection models to achieve much better medical image detection performances. We first introduce an additional top-down pyramid to help the detection networks fuse deeper multi-scale information; then, a scale enhancement module is developed to use different sizes of kernels to generate more diverse multi-scale features. Finally, we propose a feature fusion attention module to estimate and assign different importance weights to features with different depths and scales. Extensive experiments are conducted on two public lesion detection datasets for different medical image modalities (X-ray and MRI). On the mAP and mR evaluation metrics, EFPN-based Faster R-CNNs improved 1.55% and 4.3% on the PenD (X-ray) dataset, and 2.74% and 3.1% on the BraTs (MRI) dataset, respectively. EFPN-based Faster R-CNNs achieve much better performances than the state-of-the-art baselines in medical image detection tasks. The proposed three improvements are all essential and effective for EFPNs to achieve superior performances; and besides Faster R-CNNs, EFPNs can be easily applied to other deep models to significantly enhance their performances in medical image detection tasks.

1. Introduction

With the fast development of artificial intelligence, medical image analysis technologies based on deep learning have been increasingly applied in clinical computer-aided diagnosis (CAD) [1–5]. Deep-learning-based medical image detection is one of the most important tasks in CAD [6,7], which aims to recognize the locations and classes of lesions in medical images using deep models. The vanilla deep detection models, e.g., vanilla Faster R-CNN [8] and vanilla YOLO [9], simply use classic convolutional networks as their backbones. To use multi-scale features at different depths of convolutional networks, feature pyramid networks (FPNs) [10] are proposed to fuse multi-scale features at bottom-up and top-down pyramids using lateral connections. Consequently, FPNs are widely used as the backbones of deep models for many medical image detection works, e.g., FPN-based Faster R-CNNs for spinal cord injury detection [11], and FPN-based RetinaNet is adopted in [12] to detect lesions in CT images.

However, there exist two multi-scale feature fusion problems for the FPN-based deep detection models in medical image detection tasks: (i) **Insufficient fusion problem** [13–15]: Although multi-scale feature fusion has been achieved in FPNs, their performances for some medical image detection tasks are still limited, because the detection objects in some medical images are relatively small and highly similar to the background. So, the backbones of medical image detection models should fuse deeper and more diverse semantic information to enhance their feature learning capabilities. (ii) **Equal importance problem** [16,17]: FPNs treat all multi-scale features with equal importance in feature fusion. However, features with different scales at different depths should have different importance for the deep model's feature learning, so different weights should be assigned during feature fusion.

Therefore, in this work, a new detection backbone, called enhanced feature pyramid network (EFPN), is proposed to overcome the above

* Corresponding authors.
E-mail addresses: zhenghua.xu@hebut.edu.cn (Z. Xu), 202222901005@stu.hebut.edu.cn (H. Zhang), zyfradiology@hainmc.edu.cn (Y. Zhan).

https://doi.org/10.1016/j.compbiomed.2023.107149
Received 3 January 2023; Received in revised form 15 May 2023; Accepted 7 June 2023
Available online 10 June 2023

problems of FPNs and to achieve more accurate medical image detection. Compared to FPNs, EFPNs mainly have three improvements: an additional top-down pyramid, scale enhancement (SE) modules, and feature fusion attention (FFA) modules. Specifically, besides the original top-down pyramid in FPNs, the first improvement of EFPNs is to introduce an additional top-down pyramid to help the deep detection model generate and fuse multi-scale features at deeper layers and with deeper semantics. Then, EFPNs propose to integrate scale enhancement (SE) modules onto the new lateral connections between the original and additional top-down pyramids, which use parallel-arranged convolution kernels with different kernel sizes to generate and fuse more diverse multi-scale features. Consequently, EFPNs overcome the insufficient multi-scale feature fusion problem of FPNs by using additional top-down pyramid and scale enhancement modules to fuse multi-scale features with deeper and more diverse semantic information. Finally, since different multi-scale features generated at different depths or using kernels with different sizes should have different importance for the model's feature learning, novel feature fusion attention (FFA) modules are proposed in EFPNs to estimate the importance weights for multi-scale features using an attention mechanism.

In this work, EFPNs are used as the backbone of Faster-RNNs for medical image detection. However, similarly to FPNs, EFPNs is a general backbone model that can be used in almost all existing deep-learning-based detection models (e.g., Cascade R-CNNs, ATSS, RetinaNet, Grid R-CNNs) to improve their feature learning capabilities.

The contributions of this paper are briefly summarized as follows:

- We identify two multi-scale feature fusion problems (i.e., insufficient fusion and equal importance problems) for the FPN-based deep detection models in medical image detection tasks, and then propose a new enhanced backbone model, EFPNs, to overcome these problems and help the existing FPN-based detection models to achieve much better medical image detection performances.
- An additional top-down pyramid module is first proposed in EFPNs to help the model fuse deeper multi-scale information; then scale enhancement (SE) modules are applied to new lateral connections to generate more diverse multi-scale information; consequently, both improvements work together to solve the insufficient fusion problem. Finally, feature fusion attention (FFA) modules are proposed to estimate and assign different importance weights to multi-scale features with different depths and scales, which thus resolve the equal importance problem.
- Extensive experiments are conducted on two public lesion detection datasets for different medical image modalities (x-ray and MRI). The results show that (i) EFPN-based Faster R-CNNs achieve much better performances than the state-of-the-art baselines in medical image detection tasks; (ii) the proposed three improvements are all essential and effective for EFPNs to achieve superior performances; and (iii) besides Faster R-CNNs, EFPNs can be easily applied to other deep models to significantly enhance their performances in medical image detection tasks.

Section 2 introduces related work in the field of object detection and analyzes the advantages and shortcomings by comparing other methods with EFPN. Section 3 introduces the network structure and functional implementation of EPFN. Section 4 describes the experimental environment details and presents the main experimental design and the experimental results of EPFN. Section 5 summarizes the experimental results to conclude the structural and functional characteristics of EFPN.

2. Related work

2.1. Automatic medical image detection

Current deep-learning-based object detection models can be classified into two categories: (i) two-stage models (e.g., Faster R-CNNs [8]) and (ii) one-stage models (e.g., YOLO series [9,18,19] and RetinaNet [12]). R-CNNs [20] is the first deep-learning-based two-stage detection model, which is then improved by Fast-RCNNs [21]. Faster R-CNNs [8], Cascade R-CNNs [22] and Grid R-CNNs [23] are the state-of-the-art two-stage models: the former proposes a region proposal network (Denotes RPN) to extract candidate boxes and achieves end-to-end training for the first time, and the latter proposes a multi-stage structure for better detection performances. YOLO [9,18,19] is the best-known one-stage model, which directly uses the feature map for prediction without extracting candidate boxes. SSD in [24] is proposed to use feature maps at each depth to achieve better object detection. RetinaNet first uses focal loss to solve the problem of data imbalance between different classes [12], and the idea of the focal loss function is also widely used in other deep learning fields. While ATSS [25] focuses on solving the impact of positive and negative samples on detection performance, it proposes a method to improve the performance of target detection by automatically selecting a suitable anchor box as a positive sample based on the statistical features associated with the true value of the label.

The above models have been widely used for automatic medical image detection [26–28], where FPNs are used as the backbone. Y. Yan et al. [29] use YOLO to extract regions of interest to build a multi-stage breast nodule detection network and achieved improved detection results. M. Zeng et al. [30] build a multi-stage network using cascaded convolutional networks for automatic cephalometric landmark detection. There are also works [31–33] that improve on classical target detection networks and achieve breakthroughs in different medical image domains. Most of the above methods are based on classical target detection networks with specific improvements to adapt them to the characteristics of different medical image types. Our approach is to improve a generic network module by making it more suitable for the detection of medical lesions during the feature fusion phase. Table 4 shows that EFPN can be directly applied to different classical target detection networks and has improved lesion detection. Differently from these works, EFPNs propose an additional top-down pyramid and scale-enhancement modules to generate and fuse deeper and more diverse multi-scale features, while feature fusion attention modules are used to estimate the importance of multi-scale features.

2.2. Multi-scale solutions

Multi-scale solutions [34] have been widely used in deep-learning-based automatic detection tasks. In [15], the bottom-up channel is incorporated into FPNs to improve detection capabilities. [35] proposes an atrous spatial pyramid pooling (ASPP) module to obtain multi-scale features using multiple parallel-arranged atrous filters with different sampling rates. [36] analyzes the relationships between feature scale and model pretraining, and then proposes a multi-scale training method, scale normalization for image pyramids (SNIP). Furthermore, there are also many multi-scale related research works in the field of medical image segmentation [37–39]. In this work, given the additional top-down pyramid, we first add new lateral connections between the layers in original top-down pyramids and the corresponding layers in additional top-down pyramids, and then incorporate the proposed scale-enhancement modules into these new lateral connections to obtain diverse feature maps with different scales.

However, there exist two problems for the existing multi-scale feature fusion solutions, i.e., insufficient fusion problem and equal importance problem. Specifically, the insufficient fusion problem is also identified by [15], which thus adds additional feature fusion paths to the FPN network to make the fusion of feature maps at different scales more adequate, and achieves significant result improvement in the field of image segmentation. Moreover, [13] also tries to overcome the insufficient fusion problem and make FPN capable of detecting tiny targets, where a new concept fusion factor is proposed to control the information passed from the deep layer to the shallow layer. Recently, [14] proposes a more effective fusion method to overcome

this problem and to improve the accuracy of most detectors, where an improved feature pyramid network (ImFPN) is introduced to accommodate the loss of information for instances of different sizes and top-level features. Different from these works, our proposed EFPN first import an additional top-down pyramid module to help the model fuse deeper multi-scale information, then propose scale enhancement (SE) modules on new lateral connections to generate more diverse multi-scale information, which thus overcomes the insufficient fusion problem by fusing multi-scale features with deeper and more diverse semantic information.

In addition, the equal importance problem is also identified by [17], where a consistent supervision solution is proposed to introduce a supervised signal for each layer of features prior to fusion. Moreover, [16] also notice that a simple direct fusion of features with different scales may lead to underutilization of important features; therefore, a novel channel enhancement feature pyramid network (CEFPN) is proposed to alleviate channel information loss and the aliasing effects caused by hybridized fusion feature maps. Differently, in this work, we propose feature fusion attention (FFA) to provide importance weights for the different multi-scale feature maps, which thus provides greater weights for important features and improves the detection capability of the model.

2.3. Attention mechanism

Attention mechanisms have also been used in many recent works to improve the performances of deep-learning-based image processing models. SENet [40] proposes to estimate the weights of channels. In SSA-CNNs [41], a semantic self-attention is applied to suppress the background and improve the detection results. [42] proposes an Attention CoupleNet to combine attention-related information with global and local information of the object to improve detection performances. In [43], they propose a scale-attention deep learning network (SA Net) that extracts features at different scales in the residual module and uses the attention module to enhance the scale-attention capability. [44] proposes an improved U-Net with residual connections, adding channel attention (CA) blocks and hybrid dilated attention convolution (HDAC) layers to improve the accuracy of medical image segmentation. SKNet [45] uses an attention mechanism to dynamically select different convolutional kernel sizes to focus on important features at different spatial scales to improve the performance of image classification. The attention mechanism is also incorporated with Transformer-based target detection algorithms, such as DETR [46], to help the network focus on comprehensive feature information, which is particularly useful in object detection tasks with objects of different sizes and shapes in an image.

Differently from the existing works, whose attention modules are used to estimate channel or region weights on a particular feature map, our attention modules are specially designed to estimate the different importance weights of feature maps with different scales.

3. Methods

Although the existing FPN-based models have achieved good results in many automatic medical image detection tasks, FPN still encounters two problems: insufficient multi-scale feature fusion and equal importance for multi-scale features. Therefore, we propose enhanced feature pyramid networks (EFPNs) as a new detection backbone to overcome the identified problems of FPNs and achieve more accurate medical image detection. As shown in Fig. 1, compared to FPNs, EFPNs mainly consist of three improvements: an additional top-down pyramid, scale enhancement (SE) modules, and feature fusion attention (FFA) modules. Specifically, by adding an additional top-down pyramid, EFPNs have more and deeper convolution layers than FPNs; so, they can generate features with deeper and more powerful semantics, which are then fused together. Furthermore, scale enhancement modules are integrated into the new lateral connections between the original and additional top-down pyramids to introduce more diverse multi-scale features by parallel-arranged convolution kernels with different kernel sizes. Finally, feature fusion attention modules are proposed to estimate the importance weights of multi-scale features, adding the capability to highlight the important features while depressing the useless ones during model training.

3.1. Overall structure and learning procedure of EFPNs

Fig. 1 shows the EFPN-based Faster R-CNNs (abbreviated as EFPNs). EFPNs mainly consists of three feature learning pyramids. For a given input medical image, it is first sent into the bottom-up pyramid to learn more and more abstract but semantically stronger features; then, the most abstract features generated in the highest layers are fed into the top-down pyramid, where higher resolution feature maps are generated by upsampling features from higher pyramid levels; these features are fused (by element-wise summation) with features with the same size in the bottom-up pyramid using lateral connections to generate multi-scale features. To fuse deeper and richer semantic features, we introduce an additional top-down pyramid, which takes the outputs of the top-down pyramid as inputs, where new lateral connections are imported to merge feature maps of the same size from the top-down pyramid and the additional top-down pyramid. However, differently to FPNs, scale enhancement modules are added onto the new lateral connections to generate more diverse multi-scale features, while feature fusion attention modules are introduced to assign different weights to the incoming different multi-scale features with the same size before feature fusion. Finally, the fused multi-scale features are sent into a region proposal network (RPN) to obtain the candidate boxes, and the features in the candidate boxes are classified and regressed to obtain the final bounding boxes.

3.2. Additional top-down pyramid

In order to fuse more accurate and deeper features from the complex background of medical images, we first integrate FPNs with an additional top-down pyramid. In this module, a 3×3 convolution is firstly added to the lateral connection between the original and additional top-down pyramids to obtain deeper semantic features, and then an additional top-down pyramid is used to fuse these deeper semantic features. The 3×3 convolution can filter out more efficient features from the features of the initial fusion of FPNs, so as to obtain deeper features. Specifically, the output of the highest layer of the additional top-down pyramid ($\mathbf{P_5}$) is generated by performing 3×3 convolutions on $\mathbf{H_5}$. The outputs of the other layers ($\mathbf{P_1}$ to $\mathbf{P_4}$), are obtained by summing the up-sampled features with the 3×3 convolution results of same-size features ($\mathbf{H_1}$ to $\mathbf{H_4}$) on new lateral connections. Formally,

$$\mathbf{P_i} = Up(\mathbf{P_{i+1}}) + Conv_{3\times3}(\mathbf{H_i}), \ (i \in 1, 2, 3, 4) \tag{1}$$

$$\mathbf{P_5} = Conv_{3\times3}(\mathbf{H_5}) \tag{2}$$

where $\mathbf{P_i}$ (resp., $\mathbf{H_i}$) are the features generated at the ith layer of the additional (resp., original) top-down pyramid, $Up(\cdot)$ is double upsampling, and $Conv_{3\times3}(\cdot)$ is a 3×3 convolution.

3.3. Scale enhancement (SE) modules

To generate and fuse more diverse multi-scale features, SE modules are then proposed to be added on the new lateral connections between the original and additional top-down pyramids to extend the original 3×3 convolution operation to two parallel-arranged convolution operations with kernel sizes of 3×3 and 5×5. From another level, while extracting deeper features with the 3×3 convolution, the participation of the convolution of 5×5, a larger convolution kernel, can obtain a larger range of feature information from the original features. In a

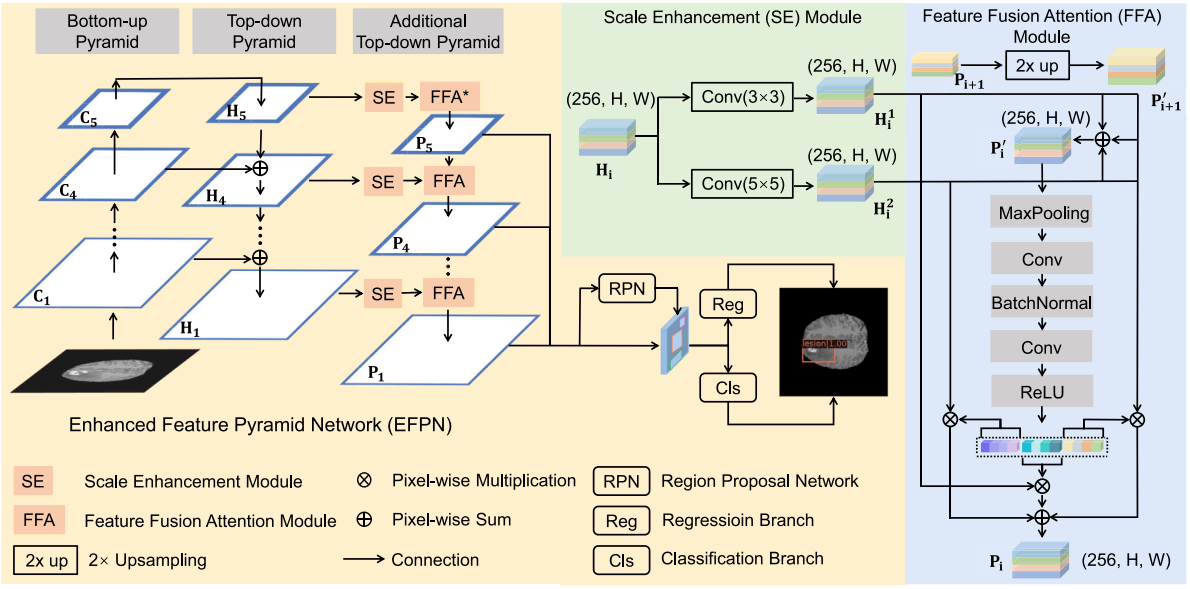

Fig. 1. Overall structure of EFPN-based Faster R-CNNs, where the feature fusion attention modules (FFA) have three inputs (i.e., $\mathbf{H}_i^1$, $\mathbf{H}_i^2$, and $\mathbf{P'}_{i+1}$), and FFA* represents a special case of FFA at the top layer of the additional top-down pyramid with only two inputs (i.e., $\mathbf{H}_5^1$, $\mathbf{H}_5^2$).

nutshell, The SE module can obtain a feature map that integrates a larger range of features while obtaining deeper features. Consequently, with SE, for a given feature map ($\mathbf{H_i}$) generated at the ith layer of the original top-down pyramid, SE generates two feature maps, $\mathbf{H}_\mathbf{i}^\mathbf{1}$ and $\mathbf{H}_\mathbf{i}^\mathbf{2}$, with different scales. Formally,

$$\mathbf{H}_\mathbf{i}^\mathbf{1} = Conv_{3\times3}(\mathbf{H_i}) \tag{3}$$

$$\mathbf{H}_\mathbf{i}^\mathbf{2} = Conv_{5\times5}(\mathbf{H_i}) \tag{4}$$

where $\mathbf{H}_\mathbf{i}^\mathbf{1}$ and $\mathbf{H}_\mathbf{i}^\mathbf{2}$ are two feature maps with the same size but different scales generated by the new lateral connection at the ith layer after applying SE, and $Conv_{5\times5}(\cdot)$ is the convolution with a 5×5 kernel.

3.4. Feature fusion attention (FFA) modules

After using the scale enhancement module to obtain diverse multi-scale feature maps H_i^1 and H_i^2 from H_i, we have three multi-scale feature maps at the ith layer of additional top-down pyramid, i.e., H_i^1, H_i^2, and P'_{i+1} (P'_{i+1} is obtained by double up-sampling of the feature map P_{i+1} from the $i+1^{th}$ layer). When fusing these three feature maps to obtain the hybrid feature map P_i, the conventional way of FPN directly addS them up, i.e., $P_i = H_i^1 + H_i^2 + P'_{i+1}$, where the importance coefficient of each feature map is the same (i.e., 1). However, we believe these three multi-scale feature maps should have different importance for the deep model's feature learning; so each feature map should be multiplied by a different importance coefficient when fusing them up. We call this an equal importance problem.

To assign different importance weights to the feature maps of different scales, new feature fusion attention (FFA) modules are proposed. Besides the one at the highest layers (denoted FFA*), FFA generally has three inputs: two feature maps with different scales generated by SE at the corresponding lateral connection ($\mathbf{H}_\mathbf{i}^\mathbf{1}$, $\mathbf{H}_\mathbf{i}^\mathbf{2}$), and a feature map ($\mathbf{P'}_{\mathbf{i+1}}$) generated by double up-sampling of the feature maps from the higher layer ($\mathbf{P}_{\mathbf{i+1}}$). FFA* only has two inputs: two feature maps with different scales generated by SE at the corresponding lateral connection ($\mathbf{H}_\mathbf{5}^\mathbf{1}$, $\mathbf{H}_\mathbf{5}^\mathbf{5}$), and other operations are the same as FFA. FFA first fuses the inputs into a new multi-scale feature map ($\mathbf{P'}_\mathbf{i}$) using element-wise summation. Then, a series of operations (i.e., max pooling, convolution, batch normalization, convolution, and ReLU operations in order) are conducted to estimate an importance weight for each channel of three input feature maps; since the number of channels for each input is 256, the size of the importance weight vector $\mathbf{W_i}$ is $3 \times 256 = 768$. $\mathbf{W_i}$ is further divided into three parts evenly ($\mathbf{W}_\mathbf{i}^\mathbf{1}$, $\mathbf{W}_\mathbf{i}^\mathbf{2}$, and $\mathbf{W}_\mathbf{i}^\mathbf{3}$), each of which is multiplied with an input feature map to highlight the important features and depress the irrelevant ones. Finally, the input feature maps are fused again by weighted element-wise summation to obtain the weighted hybrid multi-scale features ($\mathbf{P_i}$). Formally,

$$\mathbf{P'}_\mathbf{i} = \mathbf{H}_\mathbf{i}^\mathbf{1} + \mathbf{H}_\mathbf{i}^\mathbf{2} + \mathbf{P'}_{\mathbf{i+1}}, \tag{5}$$

$$\mathbf{W_i} = ReLU(Conv_{3\times3}(BN(Conv_{3\times3}(MP(\mathbf{P'_i}))))), \tag{6}$$

$$\mathbf{P_i} = \mathbf{W}_\mathbf{i}^\mathbf{1} \cdot \mathbf{H}_\mathbf{i}^\mathbf{1} + \mathbf{W}_\mathbf{i}^\mathbf{2} \cdot \mathbf{H}_\mathbf{i}^\mathbf{2} + \mathbf{W}_\mathbf{i}^\mathbf{3} \cdot \mathbf{P'}_{\mathbf{i+1}}, \tag{7}$$

where $ReLU(\cdot)$ is a ReLU-based activation function, $BN(\cdot)$ is batch normalization, and $MP(\cdot)$ is max pooling.

4. Experiments

Extensive experiments have been conducted to evaluate our proposed EFPNs. In this section, we first introduce the information of datasets, baselines, implementation details, and evaluation metrics (average precision and recall). Then, in order to prove the effectiveness of our method, we have conducted extensive experimental studies to compare the performance of EFPNs with seven state-of-art baselines: SSDs, YOLOv3s, RetinaNets, ATSS, Faster R-CNNs, Grid R-CNNs and Cascade R-CNNs. After that, in order to validate the effectiveness and necessity of three proposed advanced modules in EFPNs, ablation studies are further conducted. Finally, to prove the scalability of EFPNs, i.e., that EFPNs can be used in other deep models to improve their detection performances, we further compare the performances of EFPN-based RetinaNet, EFPN-based ATSS, EFPN-based Cascade R-CNNs and EFPN-based Grid R-CNNs, with those of FPN-based models.

4.1. Description of dataset

To show the performances of EFPNs for medical image detection, extensive experiments are conducted to compare the performances of EFPN-based Faster R-CNNs (abbreviated as EFPNs) with the state-of-the-art baselines on two common public datasets in different medical

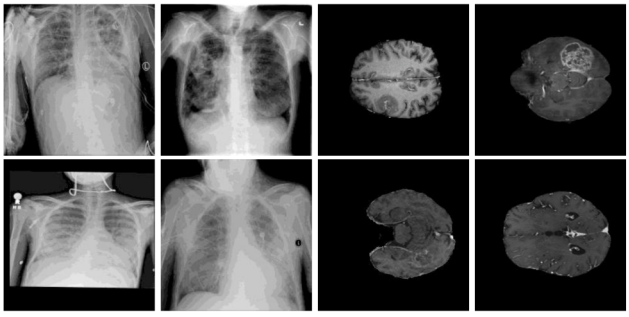

Fig. 2. Dataset examples. The left two columns are example images of PenD, and the right two columns are those of BraTs.

Table 1
The statistic information of datasets.

Dataset	Training set	Validation set	Testing set	Total
PenD	4208	601	1203	6012
BraTs	182	26	51	259

fields, Pneumonia Detection (PenD)[1] and Brain Tumor Segmentation (BraTs)[2] [47–49].

PenD is a public chest pneumonia detection dataset, which includes a total of 6,012 medical images taken from real clinical chest X-rays; in our experiments, these images will be randomly divided into training set (taking up 70% of all data), validation set (taking up 10% of all data) and test set (taking up 20% of all data). BraTs is a public magnetic resonance imaging (MRI) dataset that aims to segment the glioma of the head and has a total of 259 cases; similarly, in our experiments, we also randomly divide the dataset into training set (taking up 70% of all data), validation set (taking up 10% of all data) and test set (taking up 20% of all data). Since the original dataset is 3D images, to achieve lesion target detection, we cut the 3D images into slices according to the Z-axis, with an average of 52 slices per case. Therefore, the number of images used for training and testing is sufficient for the demand of the network for lesion target detection. To verify the effectiveness of our model on medical images of different modalities, we get the bounding box of the lesion by the segmentation labels of the dataset and use it as the label for detection. The statistic information of PenD and BraTs datasets is shown in Table 1, and Fig. 2 shows example images for these two datasets.

4.2. Baselines and implementation details

In order to evaluate the performances of the proposed EFPNs, seven state-of-the-art deep learning based detection solutions, SSD [50], YOLOv3 [51], RetinaNet [52], ATSS [25], FasterR-CNN [53], Cascade R-CNN [54] and Grid R-CNN [23], are selected as baselines for the medical image detection tasks. Please note that, although some of these baselines are first proposed in earlier years for the object detection in nature images, their vanilla versions cannot achieve satisfactory performances in medical image detection tasks because, compared to natural images, medical images have their own lesion detection difficulties, such as complex lesion texture features and tiny objects. Therefore, to show the superior performances of our proposed EFPN, instead of using their vanilla version, we select their advanced versions that are specifically proposed for medical image detection in recent years as state-of-the-art baselines.

[1] link: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
[2] link: https://www.med.upenn.edu/cbica/brats2019/data.html

The reason for selecting these seven methods as the baselines is as follows. (1) Faster R-CNNs, Cascade R-CNNs and Grid R-CNNs are state-of-the-art two-stage detection models, with relatively high detection precision. (2) SSDs, YOLOv3s, RetinaNets, and ATSS are selected as the baselines for existing excellent one-stage detection models. SSDs uses multi-scale features for detection, which effectively improves the detection accuracy of small targets; YOLOv3s has been widely used in reality due to its rapid detection; RetinaNets, as an excellent one-stage detection model, not only has a fast detection speed but also a high detection accuracy. ATSS networks use a positive sample strategy to optimize the candidate target box assignment process and thus achieve high performance in the field of natural images. (3) Among them, RetinaNets, ATSS, Faster R-CNNs, Cascade R-CNNs and Grid R-CNNs use FPNs as the backbone. Although SSDs and YOLOv3 do not adopt FPNs, they both use multi-scale features for detection.

All models are implemented using PyTorch and run on an NVIDIA GeForce GTX 2080Ti GPU. ResNet50 is adopted in FPNs and EFPNs for feature extraction. All models are trained by the SGD optimizer with a mini-batch size of 2, where the weight decay parameter is set to 0.0001. The learning rate is set to 0.002. And the threshold of score and IoU are both 0.5 at train time.

4.3. Evaluation metrics

In order to evaluate the detection performances of our proposed EFPNs and the state-of-art baselines, two widely used detection evaluation metrics [55], recall (R) and average precision (AP) are adopted. At the same time, these metrics are also applied to RSNA Pneumonia Detection Challenge (PenD). The formula for calculating R is shown in Eq. (8), it can be seen from Eq. (8) that R characterizes the ability of the model to detect lesions.

$$Recall = \frac{TP}{TP + FN}, \tag{8}$$

where TP (True Positive) means that the positive sample is correctly predicted to be the positive sample, and FN (False Negative) means that the positive sample is incorrectly predicted to be the negative sample.

AP, as the most commonly used evaluation metric for object detection, is also adopted in this work, which comprehensively considers (R) and precision (P), where the formulas of AP and P are Eq. (9) and Eq. (10).

$$AP = \int_0^1 P(r)\, dr, \tag{9}$$

$P(r)$ represents the curve composed of precision and recall.

$$P = \frac{TP}{TP + FP}, \tag{10}$$

where FP (False Positive) means that the negative sample is incorrectly predicted to be the positive sample.

To further evaluate the model effectively, we use IOU thresholds to limit the AP and R metrics. The formula of IoU is as Eq. (11). S_{Pre} denotes the area of the prediction box and S_{GT} represents the area of the ground truth box. The ratio of intersection and concatenation between the two boxes is used to calculate the distance between the predicted box and the ground truth box, and further evaluate the performance of the model. In this work, we use three different IoU thresholds, 0.5, 0.6, and 0.7, to obtain three AP results and three R results. Taking the IOU threshold of 0.5 as an example, a prediction box is considered a valid prediction when the IOU between the prediction box and the ground truth box is more than 0.5. Therefore, AP_{50}, AP_{60}, AP_{70}, R_{50}, R_{60}, and R_{70} and their mean mAP and mR are used as the final evaluation metrics.

$$IoU = \frac{S_{Pre} \bigcap S_{GT}}{S_{Pre} \bigcup S_{GT}}. \tag{11}$$

Table 2
The results of the state-of-the-art baselines (SSD, YOLOv3, RetinaNet (ReNet), ATSS, Faster R-CNN (Faster), Cascade R-CNN (Cascade) and Grid R-CNN (Grid) and our proposed method EFPN in two datasets (PenD and BraTs). The results are the average values of 3 repeated experiments.

BraTs dataset											
Method	AP_{50}	AP_{60}	AP_{70}	mAP	mAP_{CV}	R_{50}	R_{60}	R_{70}	mR	mR_{CV}	FPS
Grid (2019) [23]	0.6098	0.4918	0.3772	0.4929	0.757%	0.694	0.610	0.489	0.598	0.935%	19.26
Faster (2020) [53]	0.6057	0.5013	0.3849	0.4973	1.287%	0.681	0.590	0.458	0.576	3.523%	27.43
Cascade (2020) [54]	0.5982	0.4970	0.3953	0.4968	0.847%	0.678	0.589	0.474	0.580	1.131%	20.24
SSD (2020) [50]	0.5029	0.3861	0.2746	0.3878	0.876%	0.627	0.566	0.423	0.539	1.143%	31.44
ATSS (2020) [25]	0.6028	0.4331	0.3420	0.4594	0.706%	0.706	0.573	0.458	0.579	1.245%	27.83
YOLOv3 (2022) [51]	0.5557	0.4545	0.3411	0.4504	1.144%	0.667	0.565	0.448	0.560	0.473%	**32.91**
ReNet (2022) [52]	0.5985	0.4965	0.3787	0.4912	1.171%	0.683	0.598	0.471	0.584	0.514%	27.58
EFPN	**0.6353**	**0.5482**	**0.4041**	**0.5292**	1.062%	**0.705**	**0.615**	**0.496**	**0.605**	1.288%	24.34
PenD dataset											
Method	AP_{50}	AP_{60}	AP_{70}	mAP	mAP_{CV}	R_{50}	R_{60}	R_{70}	mR	mR_{CV}	FPS
Grid (2019) [23]	0.4748	0.2474	0.1569	0.2931	1.489%	0.636	0.410	**0.341**	0.462	2.458%	18.54
Faster (2020) [53]	0.4793	0.2869	0.1237	0.2966	2.394%	0.632	0.466	0.260	0.452	1.670%	26.34
Cascade (2020) [54]	0.4714	0.3008	0.1636	0.3108	0.792%	0.626	0.460	0.294	0.460	2.471%	19.92
SSD (2020) [50]	0.3394	0.2092	0.0644	0.2043	2.417%	0.458	0.351	0.221	0.343	2.774%	31.41
ATSS (2020) [25]	0.4665	0.2862	0.1094	0.2873	3.183%	0.569	0.447	0.301	0.439	6.618%	26.76
YOLOv3 (2022) [51]	0.3919	0.2407	0.0938	0.2421	1.169%	0.493	0.387	0.251	0.377	2.297%	**34.37**
ReNet (2022) [52]	0.4301	0.2838	0.1429	0.2839	3.110%	0.527	0.403	0.241	0.390	1.539%	28.64
EFPN	**0.4932**	**0.3339**	**0.1780**	**0.3350**	2.007%	**0.668**	**0.508**	0.336	**0.504**	2.090%	22.47

In order to ensure the effectiveness of the time, we select the FPS commonly used in object detection as a metric to evaluate the model's processing efficiency. Formally, FPS is defined as follows.

$$FPS = \frac{1}{\frac{PT}{N}}, \tag{12}$$

where PT denotes the processing time and N denotes the total number of test images.

We also use the coefficient of variation (CV) metric to measure the models' performance deviations; formally, CV can be defined as follows.

$$CV = \frac{SD}{Mean} \times 100\%, \tag{13}$$

where SD represents the standard deviation, and $Mean$ represents the mean value.

Hedge's g statistic (Hg) is a standardized indicator to measure the degree of difference between two groups, which can be used to compare the effect size [56,57] between two models. The formal definition of Hg is as follows.

$$Hg = \frac{(M1 - M2)}{SD}, \tag{14}$$

where $M1$ and $M2$ represent the average performance metrics of the two models respectively, and SD is the corrected value of the standard deviation of the performance metric of the two models. Correction values take into account the effects of sample size and biased estimates. A larger value of Hg indicates a larger difference between the two models.

4.4. Main results

To investigate the effectiveness of our proposed EFPNs, we conduct experiments on two datasets and compare the performance of EFPNs with seven state-of-the-art baselines (SSDs, YOLOv3s, RetinaNets, ATSS, Faster R-CNNs, Cascade R-CNNs and Grid R-CNNs). All experiments in this part were repeated three times under the same configurations and parameter settings. The results of EFPNs and seven state-of-the-art baselines on the two datasets in all metrics are shown in Table 2. Overall, although the detection speed of the EFPN network is slower than that of single-stage object detection networks due to the higher complexity of the two-stage architecture, within the allowed performance deviation range, EFPN outperforms other baselines in other metrics. This demonstrates that our proposed EFPN achieves more accurate medical image detection. In more detail, EFPN has higher recall compared to other baselines, being able to reach 49.4% (PenD) and 61.7% (BraTs) in the mR, which indicates that EFPNs can make the detection model more capable of detecting lesions. And it is also higher than the baselines in the AP indicators, with mAP reaching 32.35% (PenD) and 52.45% (BraTs) which shows that EFPNs are not only more able to detect lesions but also more accurate. Compared to the FPN-based Faster R-CNN network, EFPN also has a 2%–4% improvement in mAP metrics and a 3%–4% improvement in mR metrics. Furthermore, we also find that two-stage models (EFPNs, Faster R-CNNs, Cascade R-CNNs and Grid R-CNNs) are generally better than one-stage models (YOLOv3s, SSDs, RetinaNets and ATSS) on two datasets of the metrics. This is because the two-stage detection models have a network that specifically selects candidate frames compared to the one-stage detection models. Although the model requires more parameters and slows down the speed of the model, it allows the model to have more accurate positioning capabilities.

Moreover, we further calculate the coefficient of variation (CV) of EFPN and all seven SOTA baselines based on the general performance indicators mAP and mR (denoted mAP_{CV} and mR_{CV}) in Table 2. We can observe that the mAP_{CV} value of EFPN is 1.062% (resp., 2.007%) on BraTs (resp., PenD), while those of seven SOTA baselines range from 0.706% to 1.287% (resp., from 0.792% to 3.183%); similarly, the mR_{CV} value of EFPN is 1.288% (resp., 2.090%) on BraTs (resp., PenD), while those of seven SOTA baselines range from 0.473% to 3.523% (resp., 1.539% to 6.618%). Consequently, we can assert that the performance deviation of EFPN is within an acceptable level.

Table 2 also exhibits the models' processing efficiency in terms of FPS. The single-stage methods (i.e., SSD, ATSS, YOLOv3, ReNet) generally have better processing efficiency than the two-stage solutions, while the two-stage methods generally have better detection accuracy than the single-stage methods. Our proposed not only EFPN achieves the best detection accuracy but also achieves the second best processing efficiency among two-stage methods (slightly slower than Faster RCNN), which proves the applicability of EFPN in real-world scenarios.

According to the definition of Hedge's g statistic (Hg) in Eq. (14), we further calculate the effect size of the proposed model EFPN w.r.t. seven state-of-the-art baselines based on the general performance indicators (mAP and mR) in Table 2. Specifically, the effect size of our work w.r.t. the Grid RCNN is 1.723 (resp., 1.731) in mAP and 0.814 (resp., 1.680) in mR on the BraTs (resp., PenD) dataset; similarly, the effect size values of our work w.r.t. Faster RCNN and Cascade RCNN are 1.670 and 1.695 (resp., 1.668 and 1.625) in mAP, and are 1.339 and 1.576 (resp., 1.755 and 1.689) in mR on the BraTs (resp., PenD) dataset. In addition, as for the sing-stage models, the effect size values of our

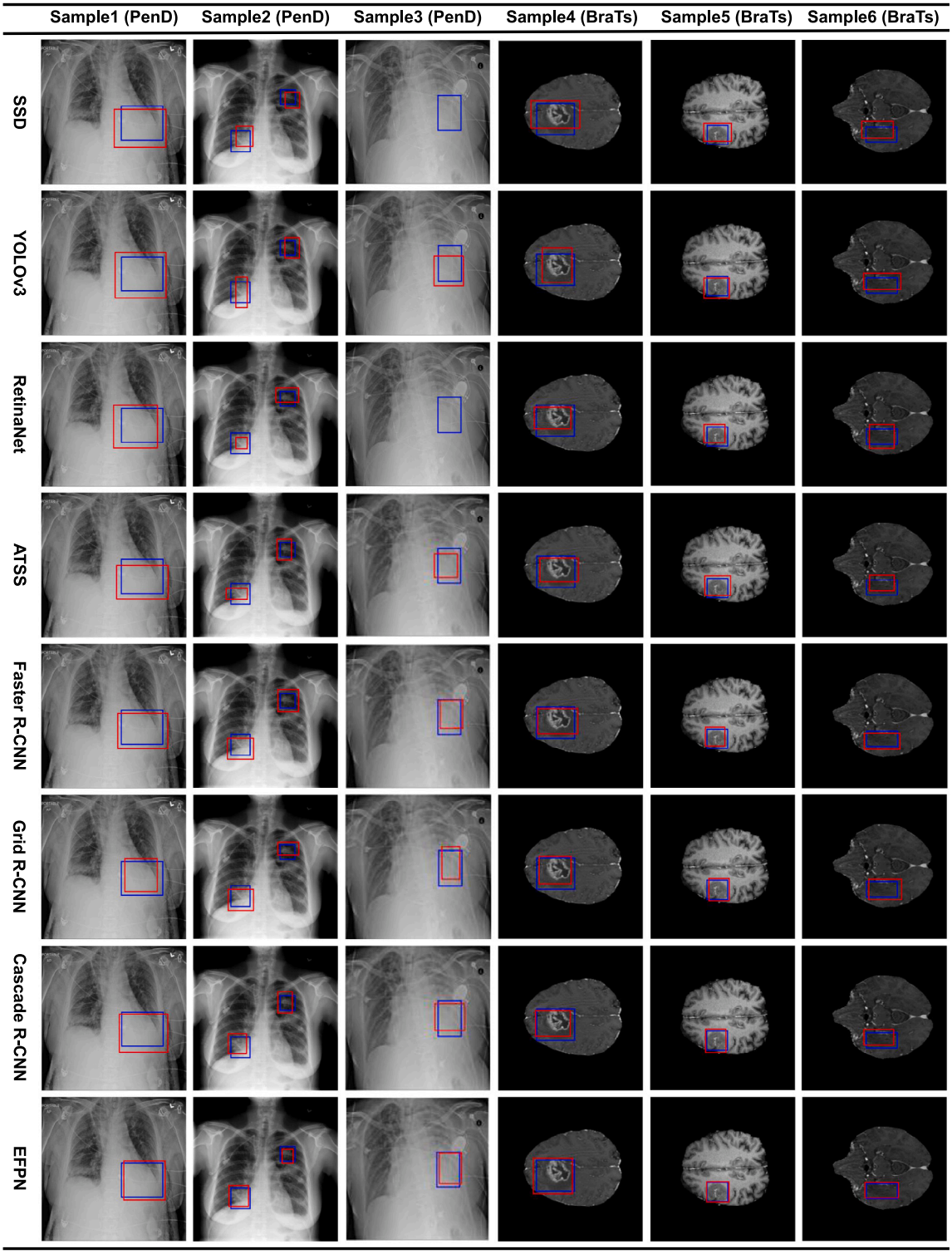

Fig. 3. Visualized examples of EFPNs and seven state-of-the-art baselines on two datasets, where blue boxes are ground truths and red boxes are predictions.

work w.r.t. Faster SSD and Cascade RCNN are 1.819 and 1.797 (resp., 1.818 and 1.685) in mAP, and are 1.767 and 1.581 (resp., 1.694 and 1.600) in mR on the BraTs (resp., PenD) dataset. Similarly, those w.r.t. YOLOv3 and RetinaNet are 1.800 and 1.717 (resp., 1.809 and 1.720) in mAP, and are 1.755 and 1.554 (resp., 1.810 and 1.812) in mR on the BraTs (resp., PenD) dataset. Consequently, since all the Hg based size effect values are larger than 0.5, it is sufficient to prove that, comparing to those of the state-of-the-art detection baselines, the performances of EFPN have significant differences, i.e., the improvements achieved by EFPN are significant w.r.t. all SOTA baselines.

Furthermore, some examples of visualized detection results of EFPNs and the seven state-of-the-art baselines are shown in Fig. 3. It shows that EFPNs can always predict more accurate bounding boxes for lesions on medical images than the baselines. Specifically, in Sample 3 (PenD), SSD and RetinaNet do not detect the lesion, and the rest of the models detect the lesion and EFPNs' prediction is closest to the

Table 3
Results of ablation studies on PenD and BraTs.

Dataset	Method	AP				R			
		AP_{50}	AP_{60}	AP_{70}	mAP	R_{50}	R_{60}	R_{70}	mR
PenD	FPN	0.4774	0.2937	0.1523	0.3078	0.626	0.463	0.265	0.451
	FPN-ATDP	0.4816	0.3012	0.1579	0.3136	0.632	0.469	0.284	0.462
	FPN-ATDP&SE	0.4872	0.3098	0.1603	0.3191	0.637	0.474	0.317	0.476
	EFPN	0.4927	0.3114	0.1664	0.3235	0.641	0.480	0.360	0.494
BraTs	FPN	0.6041	0.5040	0.3832	0.4971	0.689	0.599	0.471	0.586
	FPN-ATDP	0.6106	0.5075	0.3871	0.5017	0.698	0.606	0.480	0.595
	FPN-ATDP&SE	0.6229	0.5187	0.3952	0.5122	0.705	0.621	0.492	0.606
	EFPN	0.6317	0.5226	0.4194	0.5245	0.713	0.633	0.505	0.617

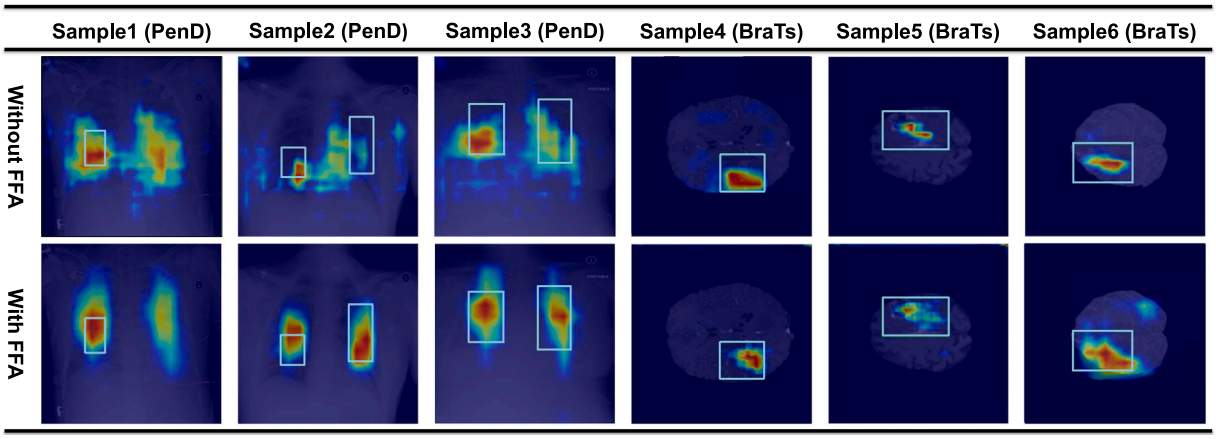

Fig. 4. Feature heatmaps of EFPNs and FPN-ATDP&SE (i.e., with or without FFA modules) on two datasets. The blue boxes are ground truths.

ground truth label. In Sample 5 (BraTs), the images of the ground truth box account for a relatively small proportion of the images and contain distracting elements of other brain-body features. One-stage networks are subject to shifted or oversized prediction frames, while two-stage networks are generally able to precisely localize to the lesion area. By comparing FPN-based Faster R-CNN and EFPN-based Faster R-CNN networks, EFPN-based Faster R-CNN is able to detect brain lesion regions more accurately. In all other samples, all models detect the lesions, and it also can be found that the predictions of EFPNs are closer to the ground truth than all baselines. This again demonstrates the superior performances of EFPNs in medical image detection.

The reason why our model can achieve better results than baselines may have the following reasons: (i) EFPNs strengthens the fusion of deeper semantic information by an additional top-down pyramid, which improves the model's ability to extract more accurate features in the complex background of medical images; (ii) The scale enhancement modules that added in the lateral connection between original and additional top-down pyramid can obtain more scale information which helps the model to obtain richer features; and (iii) In the process of feature fusion, feature fusion attention modules obtain weights according to the importance of features at different scales for weighted fusion, which improves the detection ability of models.

4.5. Ablation studies

To show the effectiveness and necessity of the proposed three improvements (additional top-down pyramid (ATDP), scale enhancement (SE) and feature fusion attention (FFA) modules), ablation studies are conducted by incrementally removing these three components from the EFPNs. Specifically, we first remove the FFA modules, resulting in an intermediate model that has ATDP and SE modules with FPN (denoted FPN-ATDP&SE), then the SE modules are removed, resulting in an FPN with ATDP model (denoted FPN-ATDP); finally, ATDP is removed, resulting in the vanilla FPN. The results of the ablation studies are in Table 3.

As shown in Table 3, we have several observations as follows: (1) It is easy to see from the Table 3 that all models (FPN-ATDP, FPN-ATDP&SE and EFPNs) are higher than the FPN on the indicators. (2) The results of EFPNs are higher than that of FPN-ATDP&SE in the indicators, which fully proves that the FFA modules are more favorable for detection by fusing features of different scales according to their importance. (3) When FPN-ATDP&SE removes the SE modules to obtain FPN-ATDP, the results of FPN-ATDP on indicators are lower than that of FPN-ATDP&SE, which proves that using the SE module to generate more diverse scale features to participate in fusion can make The model acquires richer features to aid in detection. (4) The results of FPN-ATDP on the indicators are higher than FPN, which fully reflects the importance and necessity of ATDP. The possible reason why ATDP is effective is that ATDP allows the model to fuse deeper features, which is beneficial for the model to extract accurate features from the complex background of medical images. This thus proves that the three proposed improvements are all effective and necessary for EFPNs to achieve superior detection performances in medical image detection.

Also, we show the feature heatmaps of EFPNs and FPN-ATDP&SE (i.e., without FFA) in Fig. 4 to visualize how the FFA modules optimize the learning processes of EFPNs. As shown in Fig. 4, samples 1 to 3 are three example images from the PenD dataset, and samples 4 to 6 are three example images from the BraTs dataset. By comparison, it can be found that the prominent part of the feature map with the FFA modules in the second row is closer to the green real coordinate boxes than the first row without the FFA modules. The visualization results through the feature map once again proved that the learning process of EFPNs pays much more attention to the really interesting areas of the medical images (i.e., the areas of lesions) than FPN-ATDP&SE (without FFA) does, with the help of FFA.

Last but least, although, as shown in Fig. 4, FFA works in the majority of the testing cases to help them generate accurate feature

Table 4
Results of AP and R when other lesion target detection networks adopt FPN and EFPN on PenD and BraTs, respectively. In Method and Dataset (M and D), PD stands for PenD dataset, BD represents the BraTs dataset. The simple representations of detection algorithms are RetinaNet (ReNet), Cascade R-CNN (CR-CNN) and Grid R-CNN (GR-CNN). W in the table indicates the use with the EFPNs and Wo indicates the use without EFPNs (i.e., with FPNs).

M and D	EFPN	AP_{50}	AP_{60}	AP_{70}	mAP	R_{50}	R_{60}	R_{70}	mR
ReNet and PD	w/o	0.4294	0.2765	0.1109	0.2723	0.527	0.407	0.255	0.396
	w	0.4526	0.2837	0.1134	0.2832 (+1.09%)	0.534	0.445	0.276	0.418 (+2.2%)
ReNet and BD	w/o	0.5920	0.4977	0.3839	0.4912	0.686	0.598	0.467	0.584
	w	0.6134	0.5029	0.5077	0.5080 (+1.68%)	0.718	0.609	0.455	0.594 (+1.0%)
ATSS and PD	w/o	0.4544	0.2790	0.0909	0.2748	0.524	0.438	0.265	0.409
	w	0.4661	0.2649	0.1146	0.2818 (+0.7%)	0.556	0.447	0.267	0.423 (+1.4%)
ATSS and BD	w/o	0.6008	0.4393	0.3324	0.4575	0.688	0.568	0.457	0.571
	w	0.6163	0.4463	0.3513	0.4713 (+1.38%)	0.676	0.566	0.481	0.574 (+0.3%)
CR-CNN and PD	w/o	0.4732	0.2951	0.1543	0.3075	0.614	0.475	0.281	0.457
	w	0.4892	0.3178	0.1681	0.3250 (+1.75%)	0.635	0.488	0.314	0.479 (+2.2%)
CR-CNN and BD	w/o	0.5970	0.4948	0.3866	0.4928	0.696	0.594	0.470	0.587
	w	0.6202	0.5237	0.4015	0.5151 (+2.23%)	0.710	0.622	0.480	0.604 (+1.7%)
GR-CNN and PD	w/o	0.4652	0.2442	0.1542	0.2878	0.625	0.396	0.328	0.449
	w	0.4767	0.3055	0.1520	0.3114 (+2.36%)	0.638	0.432	0.339	0.470 (+2.1%)
GR-CNN and BD	w/o	0.5989	0.4929	0.3779	0.4899	0.698	0.612	0.496	0.602
	w	0.6178	0.4941	0.3819	0.4979 (+0.8%)	0.707	0.629	0.503	0.613 (+1.1%)

maps, it cannot always guarantee accurate feature maps. There also exists some cases where FFA can only have marginal improvements or even have negative effects. As we know, since the deep learning models aim to study the generic feature distribution of the dataset, it is inevitable for them to encounter some special outliers; and we do believe that these marginal improvement cases or disadvantageous results are these kinds of outliers. Fortunately, the number of outliers tends to be small in deep learning tasks, and similarly in our work, these kinds of marginal improvement cases or disadvantageous results only account for a very small proportion of our testing set so their existence does not affect the effectiveness of FFA: as shown in Table 3, with the help of FFA, EFPN constantly outperforms the intermediate model FPN-ATDP&SE in terms of all evaluation metrics on both datasets.

4.6. Applying EFPNs to other lesion target detection networks

To prove the scalability of EFPNs, i.e., that EFPNs can be used in other deep models to enhance their detection performances, we further compare the performances of EFPN-based RetinaNet, EFPN-based ATSS, EFPN-based Cascade R-CNN and EFPN-based Grid R-CNN with those of FPN-based RetinaNet, FPN-based ATSS, FPN-based Cascade R-CNN and FPN-based Grid R-CNN. The results of all metrics for the above mentioned lesion target detection model on both datasets are presented in Table 3. The results in Table 3 show that EFPN-based models greatly outperform the FPN-based models on both datasets in the mAP and mR metrics, which thus proves that EFPNs are also applicable in other deep models to achieve better medical image detections. Compared to a one-stage lesion target detection network, EFPNs is more effective when applied to the two-stage target detection network.

Specifically, the two-stage target detection algorithm is exemplified by the Cascade R-CNN network. On the PenD and BraTs datasets, the recall (R) of EFPN-Cascade is higher than Cascade R-CNNs at all IoU thresholds. Taking their averages as an example, the mR of EFPN-Cascade on the PenD dataset is 47.9%, and the mR on the BraTs dataset is 60.4%. Cascade R-CNNs has an mR of 45.7% on the PenD dataset and 58.7% on the BraTs dataset, and EFPN-Cascade outperforms Cascade R-CNNs by 2.2% and 1.7% on both PenD and BraTs datasets, respectively. This also shows that Cascade R-CNNs using EFPNs can detect more lesions and fewer missed detection than using FPN.

In addition, on the two datasets, the average precision (AP) of EFPN-Cascade is also higher than Cascade R-CNNs at all IoU thresholds. Taking their mean mAP as an example, the mAP results of EFPNs on PenD and BraTs are 32.50% and 51.51%, respectively, and the mAP results of Cascade R-CNNs on PenD and BraTs are 30.75% and 49.28%, respectively. It can be seen that Cascade R-CNNs using EFPNs are improved by 1.75% and 2.23% respectively compared to using FPN, which shows that Cascade R-CNNs can not only obtain higher recall but also have higher precision by using EFPNs.

Similarly for other lesion target detection networks, although some network evaluation metrics may perform generally at high IOU thresholds, EFPNs networks are able to show some performance improvement for average evaluation metrics (mAP and mR). In the mAP metric, the EFPN-based lesion target detection network shows better improvement on the BraTS dataset, while in the mR metric it performs better on the PenD dataset. Various pathological image data have different lesion characteristics and image frame features, so the improvement of network detection performance by EFPN is not consistent across different datasets. But the improvement can be achieved compared to the FPN network. The above experiments show that the EFPN can replace the FPN module and achieve performance improvements in different networks, both for single-stage and two-stage lesion target detection networks. It thus proves the effectiveness and widespread use of the EFPN module.

Meanwhile, some samples of visualized detection results of the EFPN-based and the FPN-based lesion target detection network on the two datasets are shown in Fig. 5. From Fig. 5, in Sample 1, the prediction box based on the EFPN model is closer to the ground truth box than the prediction box based on the FPN model in the case of target detection for the same lesion images. Among them, the enhancement effect of the Cascade R-CNN network effect is more obvious. In Sample 2, the overall features of the image are more blurred, making the boundaries of the target region of the lesion not obvious. The prediction box based on the EFPN network is more capable of including brain lesion areas, thus providing more accurate lesion information for doctors' condition diagnosis.

4.7. Comparison with the existing multi-scale feature fusion solutions

In this subsection, additional experimental studies are conducted to compare our proposed EFPN with the existing multi-scale feature fusion solutions for the insufficient fusion and the equal importance problems to prove that the proposed EFPN can better resolve these two problems than the existing solutions. Specifically, the state-of-the-art solution for the insufficient fusion problem, PAN [15], and the state-of-the-art solution for the equal importance problem, CEFPN [16], are selected and the results are shown in Table 5.

As shown in Table 5, we first notice that PAN and CEFPN constantly outperform FPN in terms of all metrics on both datasets, which thus proves the existence of the insufficient fusion and the equal importance problems in FPN, and also proves our argument that by overcoming

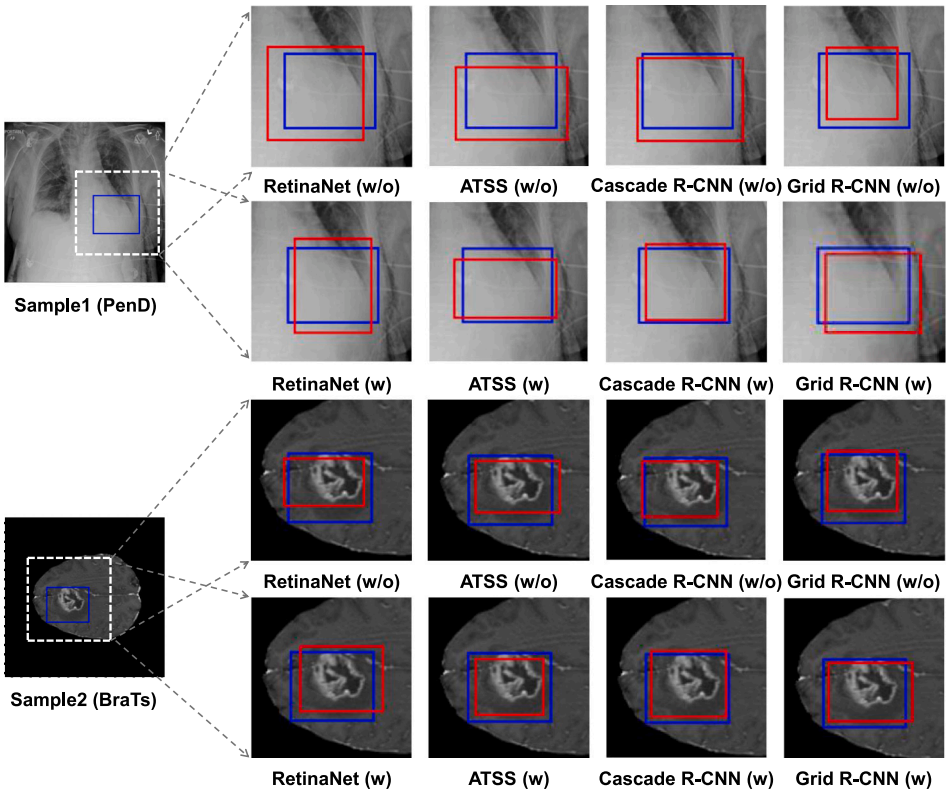

Fig. 5. Visualized examples of other lesion target detection networks on two datasets with FPNs and EFPNs, respectively, where blue boxes are ground truths and red boxes are predictions.

Table 5
Compare EFPN with the state-of-the-art solutions for the two multi-scale fusion problems on PenD and BraTs.

Dataset	Model	AP				R				FPS
		$AP_{0.5}$	$AP_{0.6}$	$AP_{0.7}$	mAP	$R_{0.5}$	$R_{0.6}$	$R_{0.7}$	mR	
BraTs	FPN (2020) [53]	0.6057	0.5013	0.3849	0.4973	0.681	0.590	0.485	0.576	27.43
	PAN (2018) [15]	0.6198	0.5120	0.3904	0.5041	0.674	0.596	0.491	0.587	25.74
	CEFPN (2021) [16]	0.6259	0.5141	0.4033	0.5144	0.692	0.624	0.502	0.606	20.58
	EFPN	0.6317	0.5226	0.4194	0.5245	0.713	0.633	0.505	0.617	24.34
PenD	FPN (2020) [53]	0.4793	0.2869	0.1237	0.2966	0.632	0.466	0.260	0.452	26.34
	PAN (2018) [15]	0.4756	0.2986	0.1447	0.3062	0.644	0.472	0.328	0.481	24.14
	CEFPN (2021) [16]	0.4831	0.3021	0.1503	0.3118	0.631	0.468	0.315	0.471	18.25
	EFPN	0.4927	0.3114	0.1664	0.3235	0.641	0.480	0.360	0.494	22.47

these two problems, the FPN-based models can achieve better performances. Furthermore, we also notice that EFPN is always better than PAN and CEFPN in detection accuracies and effectiveness, while the efficiency is also similar to them. Consequently, we can assert that the proposed EFPN is a better choice to resolve the two multi-scale fusion problems than the state-of-the-art solutions.

4.8. Comparison with the existing attention mechanisms

In this subsection, additional experimental studies are conducted to compare our proposed EFPN with the state-of-the-art attention mechanisms to demonstrate that the proposed feature fusion attention (FFA) in EFPN is better than the state-of-the-art attention mechanisms. Specifically, we first incorporate the same FPN backbone with the SOTA attention mechanisms, i.e., attention in DeTR [46] (denoted DeTR att.), channel attention [58] (denoted CA) and attention in SK-Net [45], and then compare the resulting models with EFPN. Consequently, with the same backbone, we are able to fairly evaluate the performance differences of different attention mechanisms in medical image detection tasks. The results are shown in Table 6.

As shown in Table 6, we can observe that, with the same backbone, EFPN constantly outperforms the state-of-the-art attention mechanisms in terms of all AP and R related metrics. As for the processing efficiency, the FPS values of EFPN are very close to those of CA and SK-Net, and greatly outperform those of DeTR attention. Consequently, these observations sufficiently prove that the feature fusion attention (FFA) in EFPN is better than the state-of-the-art attention mechanisms in medical image detection tasks.

5. Conclusions and future works

This work proposed an enhanced feature pyramid network (EFPN) to overcome the problems of FPNs and work as a better backbone in deep-learning-based medical image detection models. EFPNs had three improvements on FPNs: an additional top-down pyramid, scale enhancement modules, and feature fusion attention modules. Extensive experimental results proved that (i) EFPNs achieved better performances in medical image detection than the state-of-the-art baselines, (ii) the three improvements were all effective and essential for EFPNs, and (iii) EFPNs were applicable in other deep models to achieve better performances in medical image detection.

Despite achieving generally superior performance in medical image detection tasks, recent researches have proved that Transformer-based

Table 6
Compare EFPN with the state-of-the-art attention mechanisms on PenD and BraTs.

Dataset	Model	AP				R				FPS
		$AP_{0.5}$	$AP_{0.6}$	$AP_{0.7}$	mAP	$R_{0.5}$	$R_{0.6}$	$R_{0.7}$	mR	
BraTs	FPN+DeTR att.(2020) [46]	0.6005	0.4908	0.3697	0.4870	0.653	0.565	0.435	0.551	11.45
	FPN+CA (2020) [58]	0.4864	0.4021	0.3005	0.3963	0.565	0.497	0.385	0.482	25.62
	FPN+SK-Net (2019) [45]	0.5954	0.4543	0.3231	0.4576	0.641	0.543	0.414	0.533	25.15
	EFPN	0.6317	0.5226	0.4194	0.5245	0.713	0.633	0.505	0.617	24.34
PenD	FPN+DeTR att. (2020) [46]	0.4802	0.3201	0.1222	0.3008	0.614	0.468	0.289	0.457	13.32
	FPN+CA (2020) [58]	0.3569	0.2085	0.1074	0.2242	0.460	0.337	0.168	0.322	25.28
	FPN+SK-Net (2019) [45]	0.4431	0.2923	0.1181	0.2845	0.591	0.486	0.285	0.452	24.56
	EFPN	0.4927	0.3114	0.1664	0.3235	0.641	0.480	0.360	0.494	22.47

backbone has a more powerful feature learning capability than the CNN-based backbone [59]; therefore, an interesting future research is to try to incorporate the proposed EFPN with transformer-based detection models, e.g., Detection Transformer [46], to further improve the detection performances. Furthermore, Due to the high cost of annotation of medical images, we plan to extend EFPNs to semi-supervised or weakly-supervised learning models in the future. Finally, within the overall framework of multiple models, we will continue to explore the impact of information transfer and action between the feature fusion layers.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China under the grants 62276089 and 61906063, by the Natural Science Foundation of Hebei Province, China, under the grant F2021202064, by the Key Research and Development Project of Hainan Province, China, under the grant ZDYF2022SHFZ015, and by the Natural Science Foundation of Hainan Province, China, under the grant 821RC1131. This work was also supported by the Hainan Province Clinical Medical Center, China, under the grant QWYH202175, and by the AXA Research Fund, France.

References

[1] Zilong Hu, Jinshan Tang, Ziming Wang, Kai Zhang, Ling Zhang, Qingling Sun, Deep learning for image-based cancer detection and diagnosis- A survey, Pattern Recognit. 83 (2018) 134–149.

[2] Zhenghua Xu, Chang Qi, Guizhi Xu, Semi-supervised attention-guided cyclegan for data augmentation on medical images, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, 2019, pp. 563–568.

[3] Zhenghua Xu, Shijie Liu, Di Yuan, Lei Wang, Junyang Chen, Thomas Lukasiewicz, Zhigang Fu, Rui Zhang, ω-Net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution, Neurocomputing 500 (2022) 177–190.

[4] Shuo Zhang, Jiaojiao Zhang, Biao Tian, Thomas Lukasiewicz, Zhenghua Xu, Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation, Med. Image Anal. 83 (2023) 102656.

[5] Di Yuan, Yunxin Liu, Zhenghua Xu, Yuefu Zhan, Junyang Chen, Thomas Lukasiewicz, Painless and accurate medical image analysis using deep reinforcement learning with task-oriented homogenized automatic pre-processing, Comput. Biol. Med. 153 (2023) 106487.

[6] Rekka Mastouri, Nawres Khlifa, Henda Neji, Saoussen Hantous-Zannad, Deep learning-based CAD schemes for the detection and classification of lung nodules from CT images: A survey, J. X-Ray Sci. Technol. 28 (4) (2020) 591–617.

[7] Di Yuan, Zhenghua Xu, Biao Tian, Hening Wang, Yuefu Zhan, Thomas Lukasiewicz, μ-Net: Medical image segmentation using efficient and effective deep supervision, Comput. Biol. Med. (2023) 106963.

[8] Shaoqing Ren, Kaiming He, Ross B. Girshick, Jian Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, Roman Garnett (Eds.), Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, 2015, pp. 91–99.

[9] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, Ali Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[10] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, Serge J. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 936–944.

[11] Shaolong Ma, Yang Huang, Xiangjiu Che, Rui Gu, Faster RCNN-based detection of cervical spinal cord injury and disc degeneration, J. Appl. Clin. Med. Phys. 21 (9) (2020) 235–243.

[12] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, Piotr Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2999–3007.

[13] Yuqi Gong, Xuehui Yu, Yao Ding, Xiaoke Peng, Jian Zhao, Zhenjun Han, Effective fusion factor in FPN for tiny object detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1160–1168.

[14] Linxiang Zhu, Feifei Lee, Jiawei Cai, Hongliu Yu, Qiu Chen, An improved feature pyramid network for object detection, Neurocomputing 483 (2022) 127–139.

[15] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, Jiaya Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.

[16] Yihao Luo, Xiang Cao, Juntao Zhang, Jingjuan Guo, Haibo Shen, Tianjiang Wang, Qi Feng, CE-FPN: Enhancing channel information for object detection, Multimedia Tools Appl. 81 (21) (2022) 30685–30704.

[17] Chaoxu Guo, Bin Fan, Qian Zhang, Shiming Xiang, Chunhong Pan, Augfpn: Improving multi-scale feature learning for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12595–12604.

[18] Joseph Redmon, Ali Farhadi, YOLO9000: Better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6517–6525.

[19] Joseph Redmon, Ali Farhadi, YOLOv3: An incremental improvement, 2018, ArXiv Preprint, arXiv:1804.02767.

[20] Ross B. Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[21] Ross B. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[22] Zhaowei Cai, Nuno Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.

[23] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, Junjie Yan, Grid R-CNN, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7363–7372.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C Berg, SSD: Single shot multibox detector, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 21–37.

[25] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, Stan Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2020, pp. 9756–9765.

[26] Asma Baccouche, Begonya Garcia-Zapirain, C Castillo Olea, Adel S Elmaghraby, Breast lesions detection and classification via YOLO-based fusion models, Comput. Mater. Contin. 69 (2021) 1407–1425.

[27] Zhenghua Xu, Tianrun Li, Yunxin Liu, Yuefu Zhan, Junyang Chen, Thomas Lukasiewicz, PAC-Net: Multi-pathway FPN with position attention guided connections and vertex distance IoU for 3D medical image detection, Front. Bioeng. Biotechnol. 11 (2023) 1049555.
[28] Hexiang Zhang, Zhenghua Xu, Dan Yao, Shuo Zhang, Junyang Chen, Thomas Lukasiewicz, Multi-head feature pyramid networks for breast mass detection, 2023, arXiv preprint arXiv:2302.11106.
[29] Yutong Yan, Pierre-Henri Conze, Mathieu Lamard, Gwenolé Quellec, Béatrice Cochener, Gouenou Coatrieux, Towards improved breast mass detection using dual-view mammogram matching, Med. Image Anal. 71 (2021) 102083.
[30] Minmin Zeng, Zhenlei Yan, Shuai Liu, Yanheng Zhou, Lixin Qiu, Cascaded convolutional networks for automatic cephalometric landmark detection, Med. Image Anal. 68 (2021) 101904.
[31] Hongyuan Huang, Zhijiao You, Huayu Cai, Jianfeng Xu, Dongxu Lin, Fast detection method for prostate cancer cells based on an integrated ResNet50 and YoloV5 framework, Comput. Methods Programs Biomed. 226 (2022) 107184.
[32] Yongye Su, Qian Liu, Wentao Xie, Pingzhao Hu, YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms, Comput. Methods Programs Biomed. (2022) 106903.
[33] Zhenggong Han, Haisong Huang, Qingsong Fan, Yiting Li, Yuqin Li, Xingran Chen, SMD-YOLO: An efficient and lightweight detection method for mask wearing status during the COVID-19 pandemic, Comput. Methods Programs Biomed. (2022) 106888.
[34] Lei Wang, Bo Wang, Zhenghua Xu, Tumor segmentation based on deeply supervised multi-scale U-net, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, 2019, pp. 746–749.
[35] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[36] Bharat Singh, Larry S. Davis, An analysis of scale invariance in object detection SNIP, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3578–3587.
[37] Xue Wang, Zhanshan Li, Yongping Huang, Yingying Jiao, Multimodal medical image segmentation using multi-scale context-aware network, Neurocomputing 486 (2022) 135–146.
[38] Sahadev Poudel, Sang-Woong Lee, Deep multi-scale attentional features for medical image segmentation, Appl. Soft Comput. 109 (2021) 107445.
[39] Abhishek Srivastava, Debesh Jha, Sukalpa Chanda, Umapada Pal, Håvard D Johansen, Dag Johansen, Michael A Riegler, Sharib Ali, Pål Halvorsen, Msrf-net: A multi-scale residual fusion network for biomedical image segmentation, IEEE J. Biomed. Health Inf. 26 (5) (2021) 2252–2263.
[40] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[41] Chengju Zhou, Meiqing Wu, Siew-Kei Lam, SSA-CNN: Semantic self-attention CNN for pedestrian detection, 2019, ArXiv Preprint, arXiv:1902.09080.
[42] Yousong Zhu, Chaoyang Zhao, Haiyun Guo, Jinqiao Wang, Xu Zhao, Hanqing Lu, Attention couplenet: Fully convolutional attention coupling network for object detection, IEEE Trans. Image Process. 28 (1) (2018) 113–126.
[43] Jingfei Hu, Hua Wang, Jie Wang, Yunqi Wang, Fang He, Jicong Zhang, SA-Net: A scale-attention network for medical image segmentation, PLoS One 16 (4) (2021) e0247388.
[44] Zekun Wang, Yanni Zou, Peter X. Liu, Hybrid dilation and attention residual U-net for medical image segmentation, Comput. Biol. Med. 134 (2021) 104449.
[45] Xiang Li, Wenhai Wang, Xiaolin Hu, Jian Yang, Selective kernel networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.
[46] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai, Deformable detr: Deformable transformers for end-to-end object detection, 2020, arXiv preprint arXiv:2010.04159.
[47] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024.
[48] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, Christos Davatzikos, Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features, Sci. Data 4 (1) (2017) 1–13.
[49] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, 2018, ArXiv Preprint.
[50] Gang Sha, Junsheng Wu, Bin Yu, Detection of spinal fracture lesions based on SSD, in: Proceedings of the 2020 International Conference on Aviation Safety and Information Technology, 2020, pp. 539–542.
[51] Farman Ali, Sadia Khan, Arbab Waseem Abbas, Babar Shah, Tariq Hussain, Dongho Song, Shaker EI-Sappagh, Jaiteg Singh, A two-tier framework based on GoogLeNet and YOLOv3 models for tumor detection in MRI, Comput. Mater. Contin. 72 (2022) 73.
[52] Ivan William Harsono, Suryadiputra Liawatimena, Tjeng Wawan Cenggoro, Lung nodule detection and classification from Thorax CT-scan using RetinaNet with transfer learning, J. King Saud Univ.-Comput. Inform. Sci. 34 (3) (2022) 567–577.
[53] Tingxi Wen, Hanxiao Wu, Yu Du, Chuanbo Huang, Faster R-CNN with improved anchor box for cell recognition, Math. Biosci. Eng. 17 (6) (2020) 7772–7786.
[54] Shihuai Xu, Huijuan Lu, Minchao Ye, Ke Yan, Wenjie Zhu, Qun Jin, Improved cascade R-CNN for medical images of pulmonary nodules detection combining dilated HRNet, in: Proceedings of the 2020 12th International Conference on Machine Learning and Computing, 2020, pp. 283–288.
[55] Yi Ding, Qiqi Yang, Guozheng Wu, Jian Zhang, Zhiguang Qin, Multiple instance segmentation in brachial plexus ultrasound image using BPMSegNet, 2020, ArXiv Preprint.
[56] Gail M. Sullivan, Richard Feinn, Using effect size—or why the P value is not enough, J. Graduate Med. Educ. 4 (3) (2012) 279–282.
[57] Shinichi Nakagawa, Innes C. Cuthill, Effect size, confidence interval and statistical significance: A practical guide for biologists, Biol. Rev. 82 (4) (2007) 591–605.
[58] Yudong Liu, Yongtao Wang, Siwei Wang, TingTing Liang, Qijie Zhao, Zhi Tang, Haibin Ling, Cbnet: A novel composite backbone network architecture for object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, no. 07, 2020, pp. 11653–11660.
[59] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, Dustin Tran, Image transformer, in: International Conference on Machine Learning, PMLR, 2018, pp. 4055–4064.

Zhenghua Xu received a M.Phil. in Computer Science from The University of Melbourne, Australia, in 2012, and a D.Phil in computer Science from University of Oxford, United Kingdom, in 2018. From 2017 to 2018, he worked as a research associate at the Department of Computer Science, University of Oxford. He is now a professor at the Hebei University of Technology, China, and a awardee of "100 Talents Plan" of Hebei Province. He has published more than thirty papers in top AI or database conferences and journals, e.g., NeurIPS, AAAI, IJCAI, ICDE, IEEE TNNLS, Medical Image Analysis, etc. His current research focuses on intelligent medical image analysis, deep learning, reinforcement learning and computer vision.

Xudong Zhang is currently a master's student at the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China. He received B.Eng. degree from Hebei University of Engineering University, China, in 2019. His research interests lie in medical image analysis using deep learning methods.

Hexiang Zhang is currently a master s student at Hebei University of Technology, China. He received B.Eng. degree in Automation from Yanshan University, China, in 2022. His research interests lie in medical image processing using deep learning methods.

Yunxin Liu is currently a PhD student at Hebei University of Technology, China. He received his master degree in computer science from Jiangxi Normal University, Nanchang, China, in 2021. His research interests lie in medical image analysis using deep reinforcement learning.

Yuefu Zhan is currently an Associate Chief Physician at Department of Radiology, Hainan Women and Children's Medical Center. He received a Doctor of Medicine degree from the West China Medical School, Sichuan University, China, in 2022, and a Master of Medicine degree from the Xiangya School of Medicine, Central South University, China, in 2010. His current research interests mainly focus on imaging diagnosis, minimally invasive intervention, and AI-based medical image analysis.

Thomas Lukasiewicz is a Professor at Institute of Logic and Computation, TU Wien, Vienna, Austria, and Department of Computer Science, University of Oxford, UK. He currently holds an AXA Chair grant on "Explainable Artificial Intelligence in Healthcare" and a Turing Fellowship at the Alan Turing Institute, London, UK, which is the UK's National Institute for Data Science and Artificial Intelligence. He received the IJCAI-01 Distinguished Paper Award, the AIJ Prominent Paper Award 2013, the RuleML 2015 Best Paper Award, and the ACM PODS Alberto O. Mendelzon Test-of-Time Award 2019. He is a Fellow of the European Association for Artificial Intelligence (EurAI) since 2020. His research interests are especially in artificial intelligence and machine learning.