IAFI-FCOS: Intra- and across-layer feature interaction FCOS model for lesion detection of CT images

Qiu Guan[1], Mengjie Pan[1], Feng Chen[2], Zhiqiang Yang[1], Zhongwen Yu[1],
Qianwei Zhou[1], Haigen Hu[1]*

[1] *College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China*
[2] *Department of Radiology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China*
{gq, 211122120107, zqw, hghu}@zjut.edu.cn,
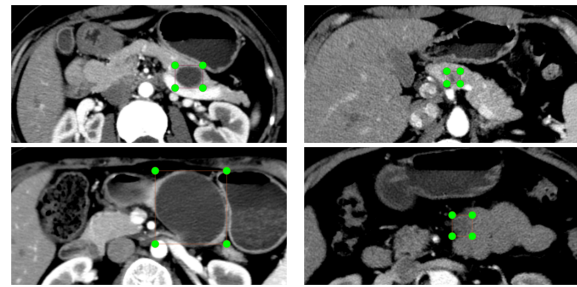chenfenghz@zju.edu.cn, 874125760@qq.com, vincentyu67373@gmail.com

Abstract—Effective lesion detection in medical image is not only rely on the features of lesion region, but also deeply relative to the surrounding information. However, most current methods have not fully utilize it. What's more, multi-scale feature fusion mechanism of most traditional detectors are unable to transmit detail information without loss, which makes it hard to detect small and boundary-ambiguous lesion in early stage disease. To address the above issues, we propose a novel intra- and across-layer feature interaction FCOS model (IAFI-FCOS) with a multi-scale feature fusion mechanism ICAF-FPN, which is a network structure with intra-layer context augmentation (ICA) block and across-layer feature weighting (AFW) block. Therefore, the traditional FCOS detector is optimized by enriching the feature representation from two perspectives. Specifically, the ICA block utilizes dilated attention to augment the context information in order to capture long-range dependencies between the lesion region and the surrounding. The AFW block utilizes dual-axis attention mechanism and weighting operation to obtain the efficient across-layer interaction features, enhancing the representation of detailed features. Our approach has been extensively experimented on both the private pancreatic lesion dataset and the public DeepLesion dataset, with AP_{50} of 62.2% and 60.0%, respectively, and these results are 6.4% and 2.3% higher than the FCOS. Additionally, our model achieves SOTA results on the pancreatic lesion dataset.

Index Terms—medical images, computer aided diagnosis, lesion detection, deep learning, object detection.

I. INTRODUCTION

Cancer is a major global public health concern, with 10 million people worldwide succumbing to cancer by the year 2020. The chances of survival would significantly increase if cancer is detected early [1]. However, there is currently a lack of simple and efficient lesion detection methods, resulting in the discovery of most cancers at later stages. For example, pancreatic Ductal Adenocarcinoma (PDAC) is often treated only after the appearance of metastatic symptoms, which leads to an increase in mortality rate [2].

In the early stages of disease diagnosis, radiologists would normally screen for tumors using medical imaging, such as Computed Tomography (CT). However, the accuracy of

* Corresponding author: Haigen Hu

(a) easy to detect   (b) hard to detect

Fig. 1. CT scan visualization of the pancreas dataset. (a) easy to detect: tumor features are distinct and have clear boundaries. (b) hard to detect: tumor boundaries are fuzzy and the target is small, a situation that is often difficult for the network to identify.

diagnose is often relies on the experience of the medical professionals, and when screening a large number of CT scan images, it can consume a considerable amount of their time and energy, leading to the possibility of errors and omissions [3]. Utilizing automated Computer-Aided Diagnosis (CAD) systems can assist doctors automatically identifying suspected lesion locations and reduce the workload for physician, enabling faster and more accurate early cancer screening.

Computer-Aided Detection (CADe) is a component of CAD, which is aim to detect lesion areas by object detection methods. Directly applying these methods to lesion detection in CT scan images may not guarantee the satisfactory performance since there's a difference between medical images and natural images. To the best knowledges we know, there exist several challenges as follows:

Firstly, there is often a correlation between the type of lesion and its location. Such as mucinous cystic neoplasms (MCNs), a type of pancreatic cystic neoplasms (PCNs), most commonly occurs in the pancreatic body and tail [5]. Existing methods [6] mainly focus on the area of the lesion and do not fully utilize the information surrounding the lesion. Secondly, the traditional detectors [8], [10], [16] has two drawbacks

when performing multi-scale feature fusion: (1) direct fuse of different layers of features reduces the representation of multi-scale features, (2) top-down transfer of features leads to loss of information. When the lesion area occupies only a small fraction of the pixels in CT scan image, or when the difference in features between the lesion area and the non-lesion area is not obvious, the loss of detail information makes it difficult to identify and localize the lesion. Fig. 1 illustrate visualization of CT scan with distinctive and less distinctive lesion features.

To address the above challenges, this paper proposes a novel intra- and across-layer feature interaction FCOS model (IAFI-FCOS). The main design a multi-scale feature fusion mechanism ICAF-FPN with intra-layer context augmentation (ICA) block and across-layer feature weighting (AFW) block. This model enriches the lesion-related features and improves the accuracy of early cancer detection. The main contributions of this paper are as follows:

- The proposed ICA block, at each intra-layer, utilizes dilated attention transformer to increase the receptive field, supplement contextual information in the lesion area, and learn long-range dependencies with surrounding.
- The proposed AFW block utilizes dual-axis attention to aggregate across-layer features, then the aggregated features adaptively filter redundant and conflicting information through weighting when fused with features of each layer. This complement the texture information and position information of small targets from low-level feature maps, alleviating the issue of information loss from traditional methods.
- The proposed IAFI-FCOS model is train and validate on both the private pancreatic lesion dataset and the public DeepLesion dataset, results in better performance compared to the other methods, which proves its performance and its robustness on different datasets.

The rest of the paper is organized as follows. We provide an overview of object detection methods and its applications on the field of medical imaging in Section II. In Section III, the method proposed in this paper is specifically described. In Section IV, the results are reported and analyzed. Finally, conclusions are drawn in Section V.

II. RELATED WORK

Object Detectors. Object detection is a fundamental task in the field of computer vision, and numerous research achievements have propelled its advancement. The object detectors are generally classified into three categories: two-stage detectors [8]–[12], one-stage detectors [13]–[19] and transformer-based detectors [20]–[22]. The two-stage detectors follow the traditional object detection pipeline, by generating lots of candidate regions and then classifying the objects present in each candidate box into different object classes. The one-stage detectors operate as either a regression or classification problem, directly mapping from image pixels to bounding box coordinates and class confidences. Transformer-based detectors are end-to-end structures, eliminating the need for manually designed components (e.g., non-maximum suppression).

For the lesion detection task, we choose the one-stage detector FCOS [16] as our baseline model. This is because its simplifies the whole target detection process and performs better when dealing with small targets.

Feature Pyramid Networks. Object detection tasks commonly utilize Feature Pyramid Networks (FPN) [10] to enable models to effectively detect objects at different scales. SSD [13] first attempts to predict the location and class of targets with multi-scale features. FPN [10] introduces a top-down and lateral connection mechanism, effectively fusing features at different scales. Subsequently, PANet [23] further proposes a bottom-up path, which combined with FPN to enable high-level features to capture detailed information in low-level features. NAS-FPN [24] optimizes the Feature Pyramid Network through neural architecture search to achieve automatic search and design and enhance the performance of object detection models. FCOS uses the traditional FPN structure with simple top-down fusion of multi-scale features, which can lead to information loss.

Medical Images lesion Detection. Medical image lesion detection, which is used of computer-aided detection (CADe) to identify the location and category of lesions [9]. With the help of CADe, the time and computational cost required can be reduced, assisting physicians to improve efficiency. For example, Ding et al. [25] and Zhu et al. [26] improve the Faster RCNN [8] and combine it with 3D convolution for the detection of lung nodules, which enhance the accuracy of nodule identification by strengthing improve the fine-grained representation and capturing more unique features. Fan et al. [27] develops a framework for Computer-Aided Diagnosis (CAD) system based on Mask Region Convolutional Neural Network [11] for the large-scale detection and segmentation of breast cancer. These frameworks all utilize 3D CNN to enhance the two-stage detector, improving detection performance, but they require substantial computational resources and extensive manual annotation of 3D bounding boxes. Subsequently, Liu et al. [6] further explores the one-stage method YOLOv3 to improve the detection performance in universal lesion areas. It significantly strengthened the accuracy of the lesion detector using only a 2D structure. However, there still remain some challenges to recognize those tough lesions, such as the smaller lesions and those with ambiguous borders.

To address above issues, we propose a novel IAFI-FCOS method to capture effective information for CT scan images, improving the accuracy and robustness of lesion detection.

III. METHOD

In this section, we present the details of the proposed the IAFI-FCOS. This study utilizes the FCOS as our baseline, and combines it with a novel proposed Neck called ICAF-FPN. The overall architecture of the network is illustrated in Fig. 2. It mainly consists of the following components: (a) Backbone: Given an input feature map I, extract multi-scale features $C_i(i = 2, 3, 4, 5)$ using convolution network, i.e. ResNet. (b) Neck: This part consists of ICAF-FPN. The main role is to aggregate and distribute the multi-scale features C_i from the
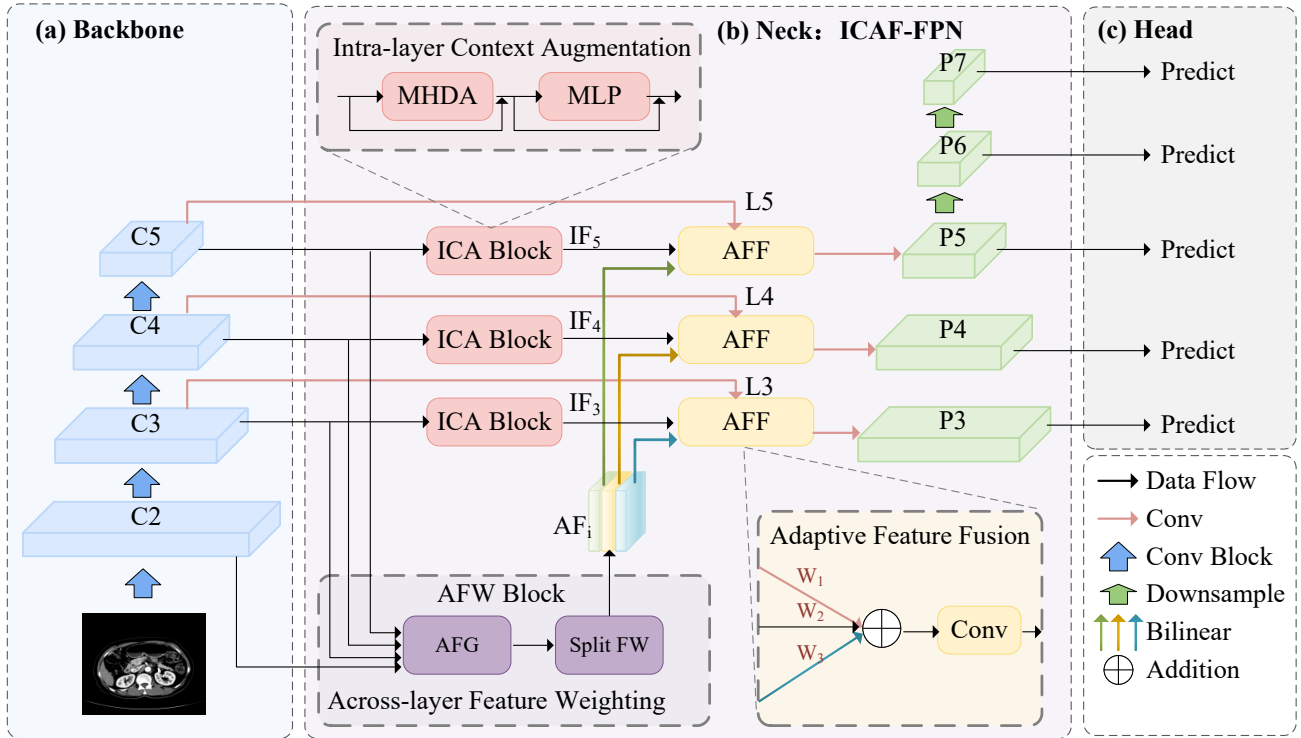
Fig. 2. Overview of the network architecture of the IAFI-FCOS detection framework, which mainly consists of three components: (a)a backbone network for feature extraction, (b)the ICAF-FPN neck and (c)the object detection head network. The C_i, P_i, IF_i, AF_i and L_i represent feature maps, the W_i indicate learnable weights.

backbone network. (c) Head: Outputs the final classification and localization prediction results.

A. ICAF-FPN

The accuracy of object detection relies on the processing of features at different scales by the NECK part. Low-level features tend to carry texture information and edge information, which helps in the localization of small targets. High-level features include semantic features and the location of larger objects. Effectively fuse features of different levels can improve the network's detection accuracy for objects of different sizes [28]. For the lesion detection task, we design a new multi-scale feature fusion mechanism, ICAF-FPN, which prevents information loss by aggregating features in both intra- and across-layer perspective.

Intra-layer. The recognition of a lesion depends not only on its inherent feature information, but also needs to be aided by the information provided by the background surroundings nearby the lesion. Traditional approaches employ convolution to process features at each layer, to capture the local features of the target. However, the nature of the convolutional operation, i.e., the limited receptive field, unavoidably has some drawbacks in establishing global dependencies. In our work, the ICA block is designed to deeply interact with each layer's inner features. We utilize multi-head dilated attention (MHDA) block to capture the global dependencies between the lesion area and the surrounding pixels, which significantly expands

the receptive fields. This approach preserves the model's sensitivity to local features while capturing global contextual information, enhancing the accuracy of lesion identification.

Across-layer. When inconspicuous lesion boundaries and small lesions appear (such as in Fig. 1 (b)), the amount of detail information determines the success of screening and detection. However, the traditional top-down and bottom-up transmission modes may lead to information loss, as each layer only receives complementary neighboring information while the cross-layer information (e.g. C_2 to C_4) is weakened and lost during transmission. Inspired by Glod-YOLO [28], we propose the AFW block, which first extract global effective features directly through across-layer feature gather (AFG) block to reduce the information loss. To retain more information about the positions of small targets, we specifically introduce the low-level feature map C_2 into the AFG block. After extracting the global features, they are assigned to different hierarchical levels. It's worth noting that directly fusing information of different densities may lead to semantic conflicts, thus limiting the expression of multi-scale features. Here, we design the split feature weighting (Split FW) block with different weights for each levels to fuse global features and filter conflicting information.

Finally, the layer features, the intra-layers interaction features and the across-layer interaction features are adaptively fused to obtain the output features of the neck by adaptive feature fusion (AFF) mechanism. Compared with existing

FPN, our proposed ICAF-FPN is more focused on the lesion detection task, which not only establishes global dependencies within each layer, but also realizes the depth of the information interaction between the different layers, further improves the expression ability of multi-scale features. This design alleviates the problems in lesion detection.

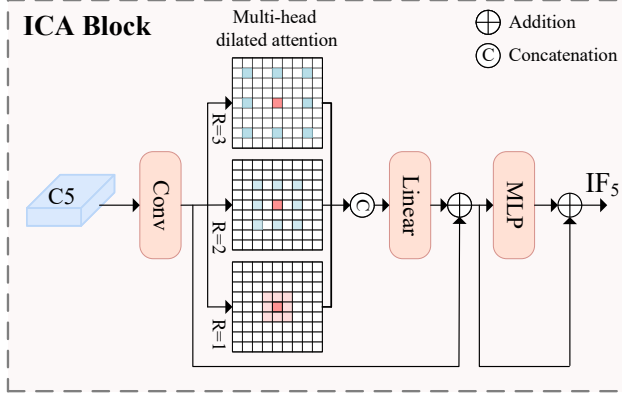## B. Intra-layer Context Augmentation Block

Fig. 3. The structure of ICA block.

Dilated convolution [29] and DilateFormer [30] expand the receptive field and capture rich contextual information by setting different dilation rates. To establish the relationship between the lesion region and the surrounding background, we design the ICA block. The structure is shown in Fig. 3, for the feature maps $C_i(i = 3, 4, 5)$, we first apply convolution to enrich local details. Subsequently, we utilize a multi-head dilated attention (MHDA) to establish dependencies between pixels. The feature map is divided into three heads by channel dimension and compute the dilated attention for each head. The dilated attention expands the receptive field by setting different the dilation rate R (e.g., R=1, 2, 3). Such as, an unflod operation utilizes a 3×3 kernel size with R = 3, and the size of receptive field is 7×7. Then, we concatenate the features from the different receptive field of three heads together and feed the concatenated features into a linear layer. The residual structure complement detailed information. Finally, the intra-layer interaction features IF_i is obtained through MLP and residual layer.

The above processes can be formulated as:

$$h_n = MHDA(C_i^{'}, R_n), \quad 1 \leq n \leq 3, \quad (1)$$

$$X_i = C_i^{'} + Linear(Concat[h_1, h_2, h_3]), \quad (2)$$

$$IF_i = X_i + MLP(X_i), \quad (3)$$

where $C_i^{'}$ is the feature map after convolution, R_n represent the dialation rate of the n-th head.

In contrast to conventional multi-head attention which calculate self-attention on the whole graph, we capture the spatial positional connections by building the different receptive fields sparsely. This approach successfully reduces computational complexity while captures global dependency relationships.

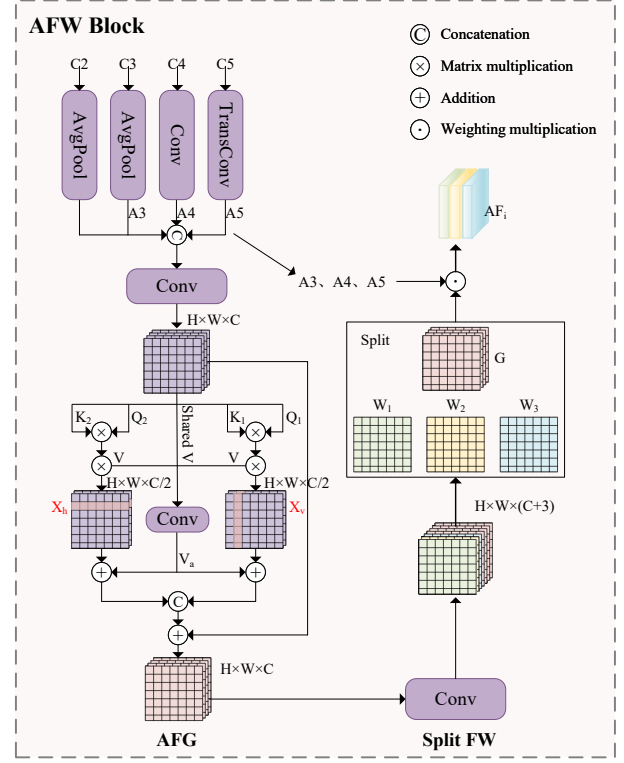## C. Across-layer Feature Weighting Block

Fig. 4. The structure of AFW block. The left and right sides of the figure represent the specific processes of AFG and Split FW, respectively.

The AFW block consists of the across-layer feature gather (AFG) block and the split feature weighting (Split FW) block, as illustrated in Fig. 4. Firstly, the AFG block aligns multi-layer features and extracts global features using a dual-axis attention operation [31]. Then, the cross-layer global features are weighted and fused with the aligned features, effectively facilitating the interaction of inter-layer information.

AFG. In the neck part of FCOS, a top-down mechanism is directly adopted to merge the $C3$, $C4$ and $C5$. To preserve more detailed information for smaller targets, we introduce features from the C2 layer and employ a different fusion mechanism. Initially, We align the four layers of different scale features to a unified size, avoiding computational overload by mapping images of different resolutions to the size of the C4. For high-resolution feature maps, the average pooling operation is used for downsampling, while the transpose convolution is taken for upsampling the low-resolution images. Due to the reason that the traditional up-sampling methods may lead to loss of information, we adopt the transposed convolution allows flexibility in retaining and reconstructing the information in the original input by learning the parameters. The formula is as follows:

$$A_i = F_{align}(C_i), \quad i = 2, 3, 4, 5 \quad (4)$$

where F_{aling} represent the alignment method, A_i represents the aligned features.

The aligned features are concatenated through the channel dimension and the global features are extracted by fusing the spliced features using the dual-axial attention mechanism. The dual-axial attention refer to the establishment of long-range dependencies in the vertical and horizontal directions, respectively. The feature map X is divided into two parts $X_v \in R^{H \times W \times C/2}$ and $X_h \in R^{H \times W \times C/2}$ by channel dimension. In the vertical direction, X_v is evenly split into W non-overlapping vertical axial stripes and projected as Q_1, K_1. In the horizontal direction, X_h is evenly split into H non-overlapping horizontal axial stripes and projected as Q_2, K_2. V is projected by the feature map X and shared in dual-axial. To compute self-attention separately, the formula is as follows:

$$X_v^{'} = Attention(Q_1, K_1, V) \qquad (5)$$

$$X_h^{'} = Attention(Q_2, K_2, V) \qquad (6)$$

where $X_v^{'}$ and $X_v^{'}$ denote the feature map after self-attention.

While the dual axial attention mechanism reduces the complexity of attention, the axis to axis interaction information is lost which is very critical for object detection task. Hence, convolutional are employed to spatially interact with the shared V, supplementing the connections between different axis. The final output of two parts are concatenated along the channel dimension. The process is formulated as:

$$X^{'} = X + Concat(X_v^{'} + V_a, X_h^{'} + V_a) \qquad (7)$$

where V_a denotes the value after interacting with the axial information, $X^{'}$ denotes the output of AFG.

Split FW. The AFG effectively aggregates across-layer information. In order to efficiently fuse global information into different layers, we use different weights to enhance features at each scale. We expand the feature map $X^{'}$ along the channel dimension by adding three dimensions, serving as learnable weights for each layer. After utilize the Split operation, these weights are divided into a globally effective feature weight map G and layer-specific feature weight maps $W_i (i = 1, 2, 3)$. W_i and G are multiplied by the aligned feature map A_i, respectively. It can complement the inter-layer correlated features while suppressing the conflicting information to enhance the representation of multi-scale features. The above processes can be formulated as:

$$AF_i = A_i \cdot G \cdot W_i, \quad i = 3, 4, 5 \qquad (8)$$

D. Adaptive Feature Fusion

The ICA and AFW blocks aggregate intra- and acoss-scale features, with the ultimate goal of fusing multiple semantic features: IF_i, AF_i and L_i. L_i is obtained by 1×1 convolution of C_i. We observe that a straightforward addition leads to confusion of distinct features, which reduce the ability to recognize the target. To address this problem, we employ an adaptive fusion mechanism, introducing weighting parameters to ensure that the contribution of each feature to the final fusion result is learned by the network. This process can be expressed as:

$$P_i = \alpha_1 \cdot IF_i + \alpha_2 \cdot AF_i + \alpha_3 \cdot Li, \ \alpha_1 + \alpha_2 + \alpha_3 = 1 \quad (9)$$

where α_i denotes different weights, P_i denotes the output of the neck part.

IV. MATERIALS AND EXPERIMENT

In this section, we introduce the datasets, the training schedule and the evaluation metrics used in our experiments. Then, we compare our proposed method with the other methods and analyze the effectiveness of our methodology by ablation study. Patient data were fully anonymized in this study to ensure confidentiality and privacy.

A. Dataset

We conducted experiments on two datasets. The first pancreatic lesion dataset is a contrast-enhanced CT images of pancreas provided by the First Affiliated Hospital, Zhejiang University School of Medicine. The dataset contains 1482 CT images, including 861 pancreatic serous cystic neoplasms (SCNs), 353 mucinous cystic neoplasms (MCNs) and 268 without tumors. CT slices of difficult-to-detect lesion (e.g., small lesions) comprise approximately 16% of the dataset. 1173 images from the dataset are used for training, and other additional 309 images are allocated for testing.

To validate the generalization performance of the model, we conducted experiments on the publicly available DeepLesion dataset [32]. This dataset contains 32,735 lesions on 32,120 axial slices from 4,427 patients. In this dataset, the lesion types are diverse and contain lesions from various sites (e.g., bone, abdomen, liver, etc.). The dataset is divided into training (70%), validation (15%), and test sets (15%) following official standards.

Data pre-process. Different HU window level and window width are set for each of the two datasets. Under different HU window, we can focus on lesions in certain specific organs. We set HU restriction in accordance with the window level and window width provided by expert radiologists, the pancreas dataset set 30 and 300, the Deeplesion dataset is set up as provided in the official documentation.

B. Experimental Setting

Implementation Details. All experiments are conducted on NVIDIA GeForce RTX 2080 11 GB GPUs. The input images training size is 640×640. The optimizer is set to stochastic gradient descent (SGD) with weight decay of 0.0001 and momentum of 0.9. The initial learning rate in our model is set to 0.02, which would automatically scaled according to batch size and GPU, and a total of 12 epochs are trained. Other comparative experiments retained the original design.

Evaluation Metrics. We mainly use two evaluation metrics: commonly used the Average Precision (AP) metric for object detection and the Free-Response Receiver Operating Characteristic (FROC). AP is defined as the area under precision-recall (PR) curve of a certain class, including AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L. Mean Average Precision (mAP) refers to the average of the summed APs for each class. In this experiment, AP is used to represent the mAP result. The alternative evaluation metric, FROC is generally used

| Method | AP | AP_{50} | AP_{75} | AP_S | AP_M | AP_L | Sensitivity | | | | |
							0.5	1	2	4	mFROC
Faster RCNN [8]	30.6	46.2	32.8	9.5	43.4	34.3	60.8	68.6	68.6	68.6	66.6
Cascade RCNN [12]	31.4	45.7	33.4	10.4	44.7	36.4	66.2	67.6	67.6	67.6	67.2
RetinaNet [14]	19.2	32.4	20.8	10.4	26.9	31.9	59.7	67.9	72.5	75.2	68.8
CenterNet [15]	32.0	48.5	35.6	10.0	44.4	35.6	67.1	72.9	80.7	81.8	75.6
TOOD [17]	32.7	46.6	35.3	16.8	44.1	38.8	66.7	79.5	80.6	82.2	77.2
Sparse RCNN [18]	24.3	42.5	25.0	4.8	33.4	32.9	59.9	62.8	65.4	66.5	63.6
YOLOx [19]	35.8	48.2	41.2	11.5	45.8	42.6	60.3	65.5	65.8	65.8	64.3
YOLOv8 [20]	37.0	49.1	41.7	19.0	42.4	38.8	61.2	63.5	66.7	66.7	64.5
Deformable DETR [21]	10.6	20.7	11.4	2.8	16.4	6.9	23.5	34. 5	36.3	36.3	32.6
DINO [22]	37.1	52.5	40.5	14.0	45.9	**45.1**	67.3	71.2	73.9	77.2	72.4
FCOS(Baseline) [16]	36.1	55.8	39.0	15.4	44.8	41.0	73.7	78.6	80.8	**82.9**	79.0
Our Method	**42.0**	**62.2**	**46.7**	**22.3**	**50.0**	44.1	**75.3**	**80.0**	**80.9**	82.3	**79.5**

in the medical field and allows the evaluation of arbitrary abnormalities on each image. Specifically, the detection of medical images requires extremely high sensitivity to ensure the detection of all abnormalities, allowing for some degree of false positives (FPs). FROC measures whether a detector can find more true positives (TPs) at the same false positive rate. In our experiments, we set the FPs per image to 0.5, 1, 2, and 4 to compare the sensitivity of different methods.

C. Results

In this section, we first evaluate the performance of our method and other different types of methods for detecting lesions on two datasets. Then ablation experiments are performed on the pancreas dataset.

Comparison study. For the pancreatic dataset, as shown in Table I, our method shows improvements in both AP and sensitivity compared to the baseline. The AP has increased by approximately 6%, and AP_S has shown an improvement of around 7%. In addition, under the more stringent mean FROC (mFROC) evaluation metric, our method improved by 0.5%. Notably, at FPs in per images of 0.5, our method obtained a higher sensitivity, achieving an improvement of 1.6%.

Comparing our method with other two-stage, one-stage and transformer detectors, our method outperforms the others in overall performance, as shown in Table I. The CenterNet and TOOD show relatively good sensitivity under the FROC metric, it still falls short of our method's performance. Additionally, the Dino achieved performance similar to the baseline, yet in comparison to our method, only has a slightly higher AP_L of 1%, while all other metrics are inferior.

For the Deeplesion dataset, which contains lesion types from different organ sites, it is more complex to realize the detection task compared to the pancreas dataset. As indicated in Table II, our method also achieves superior results on the Deeplesion dataset, with an AP_{50} reaching 60.0%, and it improves the AP_S from 0.1% to 0.5%, demonstrating the generalization of the model. In this dataset, the proportion of small lesions is extremely low. This scarcity of small lesions makes it more

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster RCNN	29.1	52.1	30.7	0.1	34.7	44.6
Cascade RCNN	29.8	51.7	32.6	0.1	35.7	45.4
RetinaNet	25.4	48.4	23.8	0.1	29.0	47.9
CenterNet	32.0	57.3	33.6	0.1	38.3	49.8
TOOD	33.5	59.3	35.6	0.3	**41.6**	50.8
Sparse RCNN	26.2	48.7	25.6	0.2	33.7	45.7
YOLOx	32.2	58.0	33.3	0.1	37.5	49.7
Deformable DETR	24.5	47.7	22.4	0.2	29.5	39.5
DINO	33.3	57.4	**36.1**	0.2	40.3	47.9
Liu [6]	-	57.5	-	-	-	-
Zhu [7]	-	60.4	-	-	-	-
FCOS(Baseline)	32.3	58.4	32.9	0.1	37.7	51.5
Our Method	**33.5**	**60.7**	34.7	**0.5**	38.8	**52.0**

challenging to learn features related to small targets during the training process, leading to difficulties in detecting these small lesions.
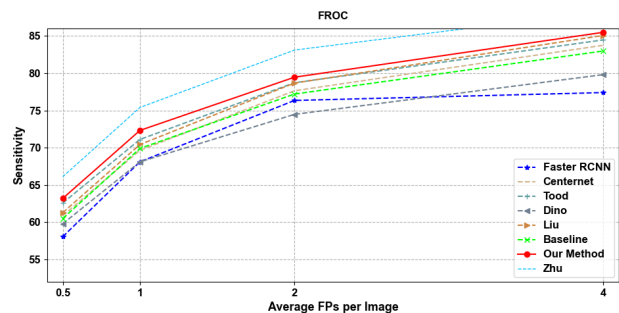

Fig. 5. The FROC curves of various methods for detection on the Deeplesion dataset.

TOOD does not differ much from the performance of our method, but the AP_{50}, AP_S and AP_L are 1.4%, 0.2% and

1.2% higher than TOOD, respectively. when comparing more stringent FROC metrics, our sensitivity is higher the TOOD, as shown in Fig. 5. FROC curves visually compare the sensitivity of methods with similar AP values, and our method were higher than the baseline at average FPs per image = 0.5, 1, 2, and 4 by 2.7%, 2.4%, 2.2%, and 2.5%, respectively. However, compared to Zhu's method [7], our FROC metric still fails to fully outperform it, despite performing better on the AP50. We will continue to make improvements in subsequent studies to optimize our approach.

As shown in Fig. 6, the first column visualizes the detection results of the slices in the pancreas dataset and the second column shows the results of the Deeplesion dataset. Our method (b) is better at recognizing categories compared to baseline (a), since there is no misdiagnosis of mucous cyst as serous cyst. Furthermore, the baseline results may exhibit some false positives, while our method avoids such occurrences and achieves higher confidence.
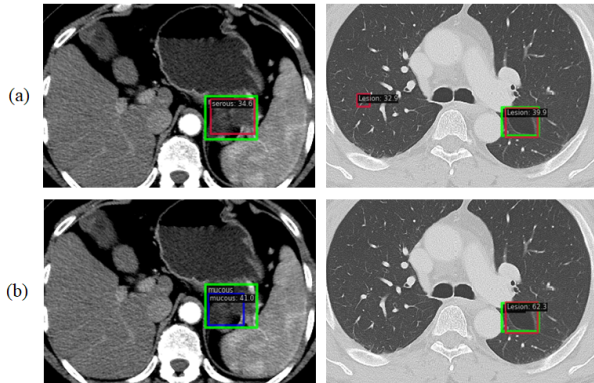

Fig. 6. Detection visualization of the comparison between baseline and our improved detector. (a): the results of baseline. (b): the results of our method. The green and red bounding boxes represent ground truths and predictions.

To validate the effectiveness of our designed ICAF-FPN structure, we replace the FPN structure in the neck part of baseline and compare the performance of detectors with different FPN mechanisms, as shown in Table III. Comparing with FPN, our method improves the AP by 6% and the AP_S by 7%, which is a significant improvement in the overall performance. Despite using the structure of PAFPN to improve the effectiveness, our method still outperformed PAFPN by 3%. This demonstrates that our proposed approach effectively improves lesion detection performance.

TABLE III
ABLATION STUDY ON FPN, DETECTION PERFORMANCE UNDER
DIFFERENT FPN STRUCTURES.

Neck	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FPN [10]	36.1	55.8	39.0	15.4	44.8	41.0
PAFPN [23]	39.4	59.7	42.1	17.9	47.4	42.7
NAS-FPN [24]	24.3	37.7	25.9	10.3	32.9	32.7
ICAF-FPN	**42.0**	**62.2**	**46.7**	**22.3**	**50.0**	**44.1**

Ablation study. We perform ablation experiments on the proposed method to evaluate the performance improvement in lesion detection. The same strategy and parameters are used during training and validation to ensure a fair comparison.

We evaluated the impact of different designs in the ICAF-FPN on the model. As shown in Table IV, when only the ICA block is introduced, resulted in an AP of 39.0%, an improvement of 2.9% compared to the baseline. Notably, the AP_S also achieves an improvement of 2.7%. This indicates that combining background information can help with lesion detection. When only introducing the AFW block for information interaction across different scales, we achieved better results for small target detection, with a 4.1% improvement over the baseline, while other metrics also showed improvements. This reflects the fact that our method retains more details that contribute to lesion identification. When combining ICA and AFW and not using AFF, we noticed that AP_S and AP_L are not as high as when using AFW blocks alone. We conjecture that feature redundancy and conflicts occur when using additional fusion of different semantic features, leading to confusion between small and large target information. To address this, we introduce adaptive features fusion (AFF) mechanism by assigning different weights to the features. As shown in the Table IV, after the introduction of AFF, AP_S reached 22.3%, and AP_L goes from 41.6% to 44.1%, which is an improvement in each metric, proving the effectiveness of the adaptive module.

TABLE IV
ABLATION STUDY ON ICAF-FPN, COMPARISON OF THE PERFORMANCE
FOR DIFFERENT BLOCKS.

ICA	AFW	AFF	AP	AP_{50}	AP_S	AP_M	AP_L
			36.1	55.8	15.4	44.8	41.0
✓			39.0	58.0	18.1	46.0	43.4
	✓		39.1	58.1	19.5	45.1	**44.6**
✓	✓		39.6	60.9	18.9	46.9	41.6
✓	✓	✓	**42.0**	**62.2**	**22.3**	**50.0**	44.1

In the AFW block, to retain more information about small targets, we additionally introduced the C2 layer compared to the baseline. To verify the necessity of the C2 layer, we conducted an ablation experiment, as shown in Table V.

TABLE V
ABLATION STUDY ON AFW BLOCK, W/O INDICATES THAT NO C2 LAYER
WAS INTRODUCED, W/ INDICATES THAT A C2 LAYER WAS INTRODUCED.

ICAF-FPN	AP	AP50	AP_{75}	AP_S	AP_M	AP_L
w/o C2	38.0	57.3	41.8	14.2	46.8	**44.1**
w/ C2	42.0	62.2	46.7	22.3	50.0	44.1

Our method fully utilizes the information extracted from both inter- and intra-layers to enhance the overall detection performance, which also effectively improving the detection rate of small targets without losing the accuracy of large targets.

V. CONCLUSION

In this paper, we propose a lesion detector based on one-stage method called IAFI-FCOS for assisting radiologists to achieve faster and more accurate early screening and detection of cancer lesions. We mainly focus on enhancing the neck part of the object detection framework, which extracts lesion-related information from both intra- and across-scale perspectives for features of different scales. Subsequently, we adaptively fuses the diverse semantic features of each scale and feed into the prediction head to obtain the final detection results. The novel multi-scale feature fusion mechanism ICAF-FPN alleviates the challenge of detecting ambiguous and small lesions. Extensive experiments have demonstrated a significant improvement in the detection of the pancreatic lesion dataset, as well as an enhanced detection performance of the Deeplesion dataset. Continuous efforts are still needed to achieve full generalizability across various lesion detection domains. In future studies, we will continue to improve and optimize our method to enhance domain generalizability and better meet the needs of the medical image analysis field.

REFERENCES

[1] D. Crosby, S. Bhatia, K. M. Brindle, L. M. Coussens, C. Dive, M.Esener, et al., "Early detection of cancer, " Science, vol. 375, pp.eaay9040, 2022.

[2] M. T. Zavalsiz, S. Alhajj, K. Sailunaz, T. Özyer, and R. Alhajj, "A comparative study of different pre-trained deep learning models and custom CNN for pancreatic tumor detection." Int. Arab J. Inf. Technol., vol. 20, pp. 515–526, 2023.

[3] J. Vosshenrich, P. Brantner, J. Cyriac, D. T. Boll, E. M. Merkle, and T. Heye, "Quantifying Radiology Resident Fatigue: Analysis of Preliminary Reports," Radiology, vol. 298, pp. 632–639, 2021.

[4] H. P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," Medical physics, vol. 47, pp. 218–227, 2020.

[5] H. Babiker, G. Hoilat, G. Menon, and S. K. Mukkamalla, "Mucinous cystic pancreatic neoplasms," StatPearls, 2023.

[6] Z. Liu, K. Han, K. Xue, Y. Song, L. Liu, Y. Tang, and Y. Zhu, "Improving CT-image universal lesion detection with comprehensive data and feature enhancements," Multimedia Systems, vol. 28, pp. 1741-1752, 2022.

[7] Y. Zhu, Z. Liu, Y. Song, K. Han, C. J. Qiu, Y. Y. Tang, et al. "CPSNet: a cyclic pyramid-based small lesion detection network," Multimed Tools Appl, 1-19, 2023.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, pp. 91–99, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 37, pp. 1904-1916, 2015.

[10] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117-2125, 2017.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969, 2017.

[12] Z. Cai, and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6154-6162, 2018.

[13] W. Liu et al., "SSD: Single shot multibox detector," In Computer Vision–ECCV, Switzerland: Springer, pp. 21–37, 2016.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," In Proceedings of the IEEE international conference on computer vision, pp. 2980-2988, 2017.

[15] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," arXiv preprint arXiv:1904.07850, 2019.

[16] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," In Proceedings of the IEEE/CVF international conference on computer vision, pp. 9627-9636, 2019.

[17] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3490-3499, 2021.

[18] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, et al., "Sparse r-cnn: End-to-end object detection with learnable proposals," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14454-14463, 2021.

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

[20] G. Jocher, A. Chaurasia, and J. Qiu, "Yolo by ultralytics," URL: https://github.com/ultralytics/ultralytics, 2023.

[21] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," arXiv preprint arXiv:2010.04159, 2020.

[22] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, et al., "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," arXiv preprint arXiv:2203.03605, 2022.

[23] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8759-8768, 2018.

[24] G. Ghiasi, T. Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7036-7045, 2019.

[25] J. Ding, A. Li, Z. Hu, and L. Wang, "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks," In Medical Image Computing and Computer Assisted Intervention, pp. 559-567, 2017.

[26] W. Zhu, C. Liu, W. Fan, and X. Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," In 2018 IEEE winter conference on applications of computer vision (WACV), pp. 673-681, March 2018.

[27] M. Fan, H. Zheng, S. Zheng, C. You, Y. Gu, X. Gao, et al., "Mass detection and segmentation in digital breast tomosynthesis using 3D-mask region-based convolutional neural network: a comparative analysis, Frontiers in molecular biosciences, vol. 7, pp. 599333, 2020.

[28] C. Wang, W. He, Y. Nie, J. Guo, C. Liu, K. Han, and Y. Wang, "Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism," arXiv preprint arXiv:2309.11331, 2023.

[29] F. Yu, and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015.

[30] J. Jiao, Y. M. Tang, K. V. Lin, Y. Gao, J. Ma, Y. Wang, and W. S. Zheng, "Dilateformer: Multi-scale dilated transformer for visual recognition," IEEE Transactions on Multimedia, 2023.

[31] H, Zhou, R, Yang, Y. Zhang, H. Duan, Y Huang, R. Hu, et al., "Uni-Head: Unifying Multi-Perception for Detection Heads," arXiv preprint arXiv:2309.13242, 2023.

[32] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," Journal of medical imaging, vol. 5, pp. 036501-036501, 2018.