MAFF: Multi-scale and self-adaptive attention feature fusion network for pancreatic lesion detection in PET / CT images

Heng Wang
Electronic and Information Engineering, Changchun University of Science and Technology, Changchun
wh1542335102@gmail.com

Zhongyi Wu
Department of Medical Imaging, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences
Suzhouwuzy@sibet.ac.cn

Fei Wang
Department of Medical Imaging, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences
Suzhouwongfei@shu.edu.cn

Wenting Wei
Electronic and Information Engineering, Changchun University of Science and Technology
Changchunweiwenting765@mails.cust.edu.cn

Kezhen Wei
Electronic and Information Engineering, Changchun University of Science and Technology
Changchunweikezhen19219@163.com

Zhaobang Liu†
Department of Medical Imaging, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou†
liuzb@sibet.ac.cn

ABSTRACT

Accurate, automated medical image detection is critical in clinical diagnosis and analysis. Since the pancreatic lesions in CT images are similar to the surrounding tissues, they are difficult to be detected. The lesions in PET images have the disadvantage of blurred edges that the precise localization is slightly insufficient. PET/CT integrates functional and anatomical imaging, combining the advantages of the high contrast of PET images and the high spatial resolution of CT images to assist doctors in detecting lesions. Therefore, it is significant for us to study the object detection of lesions based on PET/CT. At the same time, the context information extraction ability of the basic framework Faster R-CNN is insufficient. Therefore, we propose a multi-scale adaptive attention feature fusion network (MAFF) based on PET/CT to realize the automatic detection of pancreatic lesions. First, we fuse multi-scale features through a feature pyramid module to obtain richer contextual information and add an attention module to achieve preliminary screening of input features. Second, we design an adaptive attention feature fusion network to make feature semantic information selection more focused by recalibrating multimodal feature maps. Finally, we adopt a pooling module, which not only solves the problem of different sizes of region proposals but also avoids the localization error caused by quantization. Experimental results show that our proposed multimodal algorithm outperforms other algorithms on two challenging tasks, pancreatic cancer lesion detection, and head and neck cancer lesion detection.

EITCE 2022, October 21–23, 2022, Xiamen, China

ACM ISBN 978-1-4503-9714-8/22/10…$15.00
https://doi.org/10.1145/3573428.3573678

CCS CONCEPTS

• **Computer systems Organization** → Architectures; Other architectures; Neural networks.

KEYWORDS

Faster R-CNN, pancreatic lesions, PET/CT, object detection

ACM Reference Format:
Heng Wang, Zhongyi Wu, Fei Wang, Wenting Wei, Kezhen Wei, and Zhaobang Liu†. 2022. MAFF: Multi-scale and self-adaptive attention feature fusion network for pancreatic lesion detection in PET / CT images. In *2022 6th International Conference on Electronic Information Technology and Computer Engineering (EITCE 2022), October 21–23, 2022, Xiamen, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3573428.3573678

1 INTRODUCTION

Pancreatic cancer [1-2] is one of the most common malignant tumors of the digestive system, with rapid progression, early metastasis, high mortality, and poor prognosis. According to the American Cancer Society, about 45,000 people died in the United States in 2019, the third-highest number behind lung and colorectal cancers. Pancreatic cancer will become the second leading cause of cancer-related death in Western societies by 2030 [3-4].

Pancreatic cancer diagnosis is very challenging. Firstly, the pancreas is located deep in the abdominal cavity and is covered by abdominal organs such as the stomach, transverse colon, and greater omentum. Secondly, the pancreas has the characteristics of ambiguous positioning of internal organs. Thirdly, compared with ultrasound, magnetic resonance imaging (MRI), endoscopic ultrasonography, and positron emission tomography, computed tomography (CT) is the most commonly used imaging modality for the initial evaluation of suspected pancreatic cancer [5-6]. However, pancreatic lesions are difficult to observe on computed tomography (CT) because of their low contrast with surrounding tissues [7]. Some small lesions may easily lead to missed diagnosis and misdiagnosis. Even experienced radiologists are challenging to identify effectively in a short period, which is time-consuming [8]. At the same time, despite the high contrast of PET images, the edges of the

lesions are blurred, making precise localization difficult. PET/CT is a multimodal imaging technology that integrates functional imaging and anatomical imaging, which combines the advantages of the high contrast of PET images and the high spatial resolution of CT images. It has obvious benefits in diagnosis, staging, guiding treatment, and prognosis of patients. Therefore, to reduce the burden on radiologists while reducing image interpretation time, it is meaningful to seek an automatic and accurate method based on PET/CT to assist doctors in diagnosis.

With the development of deep learning technology, it is widely used in medical image analysis such as lesion detection [9-10], lesion segmentation [11-12], and disease classification [13-14]. Pancreatic cancer-based automated detection is also gaining more and more attention. Because Faster R-CNN [15] is mature, expansible, and widely used in the medical field, we choose Faster R-CNN as the basic framework to detect pancreatic lesions based on PET/CT images. However, the detection of pancreatic lesions by Faster R-CNN is not ideal, and there are mainly the following problems: (1) The ability to obtain contextual information is poor. (2) There is quantization, which leads to errors in positioning.

To improve the problem of poor context acquisition ability, Zhang et al. [16] used the multi-resolution features extracted in the feature pyramid structure in gastric polyp detection, reused the information discarded by max pooling, and stitched these data as additional features with the output features to enhance classification and detection. Chen et al. [17] added an attention module to focus on helpful feature channels and weaken helpless feature channels in colorectal cancer detection, further improving the quality of feature maps generated by the feature extraction network and the feature extraction ability. To solve the influence of quantization on the localization error, He et al. [20] proposed ROI Align in Mask R-CNN, which uses bilinear interpolation to obtain accurate values of input features, thus avoiding the use of quantization.

In this paper, based on the discussion of the first question, we first add a pyramid module to obtain rich multi-scale contextual information. Secondly, an attention module is added to realize the preliminary screening of input features. Finally, we design a multi-scale adaptive attention feature fusion (MAFF) module. Specifically, the two multi-scale input feature maps are spliced and fused. Then, by recalibrating the multimodal feature maps, the feature semantic information selection is more focused so that context-rich and prominent multimodal feature maps are obtained. Based on the discussion of the second question, we add ROI Align to Faster R-CNN to solve the above problem. Based on the above improvements, we propose a multi-scale adaptive attention feature fusion network based on pancreatic lesion detection in PET/CT images. The main contributions of this paper include:

(1). A multi-scale adaptive attention feature fusion network is proposed to obtain feature maps with wealthy and prominent contextual information through multi-scale information fusion and adaptive attention feature screening.

(2). According to the experimental results of the pancreatic and head and neck cancer datasets, our algorithm has a certain generalization ability.

2 RELATED WORK

Object detection is a hot topic that has received much attention and has been widely used in various fields. Next, I will introduce related work from three aspects:

1) General object detection algorithms
2) Algorithm application based on pancreatic cancer
3) Algorithm application based on multimodal PET/CT

2.1 General object detection algorithms

Object detection aims to determine the classes of instances in an image and indicate their locations with bounding boxes. Existing deep learning-based object detection networks can be divided into two- and one-stage networks. The main difference is that the two-stage network extracts a set of regions of interest (ROIs) before making detections. In contrast, the one-stage network directly performs dense sampling-based detection.

The two-stage network is based on R-CNN [18], which combines region proposal and convolutional neural network for the first time. On this basis, Girshick et al. [19] utilized softmax in Fast R-CNN for classification and bounding box regression using a multi-task loss function, speeding up training and testing time. Ren et al. [15] proposed the RPN network to replace the SS algorithm in Fast R-CNN to extract region proposals, further reducing the running time and improving the detection accuracy. He et al. [20] enhanced and extended Faster R-CNN by adding segmentation branches and designing ROI Align to replace ROI Pooling in Mask R-CNN. The one-stage network is dominated by the YOLO series, including YOLO [21], YOLO-V2 [22], YOLOV3 [23], SSD [24], and so on. They are fast compared to two-stage networks and can be real-time.

2.2 Algorithm application based on pancreatic cancer

In recent years, with the increasing proportion of pancreatic cancer patients and the clinical emphasis on diagnosis and treatment, deep learning research on pancreatic cancer has been the focus of many researchers. Research on it has also emerged in an endless stream. Liu et al. [25] used the Faster R-CNN algorithm to establish an artificial intelligence diagnosis system for pancreatic cancer based on sequence-enhanced CT images and realized the automatic detection of pancreatic lesions. On the framework of the Faster R-CNN algorithm, Zhang et al. [26] achieved effective detection of pancreatic cancer by extracting multi-scale context information and capturing the interaction information between candidate regions and surrounding tissues. Wang et al. [27] proposed an Induced Attention Guided Network (IAG-Net) for regular/PDAC classification and semi-supervised PDAC segmentation, effectively eliminating the background interference on lesion segmentation. Si et al. [28] proposed a fully end-to-end deep learning (FEE-DL) model for automatically diagnosing pancreatic tumors from raw abdominal CT images. Natáliaet al. [29] developed an automated framework for PDAC detection using the nnUnet model, enabling efficient diagnosis of small lesions. Ma et al. [30] designed a CNN model to classify pancreatic CT images to aid in diagnosing pancreatic cancer.

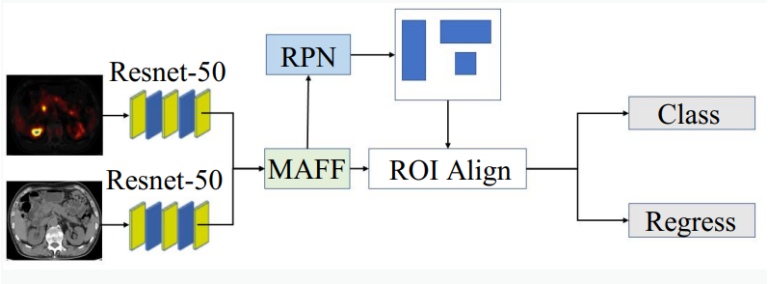

Figure 1: Frame diagram of the improved Faster R-CNN based on PET/CT

2.3 Algorithm application based on multimodal PET/CT

Compared with single-modality CT and PET, multi-modality PET/CT has become a popular research topic due to its complementary advantages of using the high contrast of PET images and the high spatial resolution of CT images. In object detection, Jonathan Wehrend et al. [31] used a deep, fully convolutional neural network based on 68Ga DOTATATE PET/CT in the automatic detection of liver lesions and realized the effective detection of liver lesions. Avi et al. [32] combined FCN and GAN to detect malignant liver lesions, used CT images to generate simulated PET images, and used the original PET images for supervision, effectively reducing false positives.

In addition, in areas such as segmentation and classification, Li et al. [33] proposed a computer-aided diagnosis model of pancreatic cancer based on PET/CT images, which realized the effective diagnosis of pancreatic cancer. Li et al. [34] A 3D fully convolutional network (FCN) was designed to achieve accurate lung tumor segmentation by generating CT probability maps combined with PET images. Kumar et al. [35] first encoded PET/CT images to extract semantic features, then spliced and multiplied the encoded features in the decoding stage to obtain complementary multimodal information at accurate locations, thereby achieving precise segmentation of lung cancer. At HECKTOR 2020, Yuan et al. [36-39] proposed a dynamic scale attention mechanism on a fully automatic segmentation network based on encoder-decoder architecture, which combines low-level details from different scale feature maps with high-level semantics, and achieved fourth-place result in the Head and Neck Cancer Challenge.

3 METHODS

Our network architecture is shown in Fig. 1, which uses 2D Faster R-CNN as the base model, which is divided into the following four parts:

(1) The first part is the Dual Feature Extraction Network, which consists of ResNet-50, Feature Pyramid Network (FPN) [40], and the Attention Module (AM) [41]. This part mainly obtains the feature map of the input image.

(2) The second part is the Multimodal Self-adaptive Attention Feature Fusion Module (MAFF), which realizes multi-modal feature fusion.

(3) The third part is the RPN network, which divides the candidate regions.

(4) The fourth part comprises two sub-networks, mainly responsible for object classification and bounding box regression of effectual feature map output to predict the category and location of pancreatic lesions.

3.1 Dual Feature Extraction Network

We designed a dual feature extraction network to extract feature maps from PET and CT images to achieve multimodal lesion detection. ResNet designs a residual module that allows us to train deeper networks and solve the problem of vanishing gradients. At the same time, after comparing with ResNet-34 and ResNet-101, ResNet-50 has a better effect, so we use ResNet-50 as the essential feature extraction network. In the residual network of ResNet-50, we added an attention module (AM) (Fig. 3), as shown in the Fig. 2. We exploit the channel-to-channel relationships of features to generate channel attention maps. Feature maps are recalibrated to enhance feature representation by explicitly emphasizing informative feature channels, weakening the saliency of less important feature channels, and generating high-quality region proposals.

To get rich contextual information, we add the pyramid module. Bottom-up (the process of upsampling the Resnet-50 feature map after convolution) feature maps have lower-level semantics, but they are sampled less often, making localization more accurate. Each lateral connection merges feature maps of the same spatial size from the bottom-up and top-down paths. And the prediction is made independently at each pyramid level, enabling the model to fully utilize contextual multi-scale information to detect lesions of various sizes and improve speed.

3.2 Multimodal Self-adaptive Attention Feature Fusion (MAFF)

The traditional multimodal fusion uses splicing, which is prone to feature redundancy and affects the detection effect. Therefore, we design MAFF (Fig. 4) to achieve multimodal fusion. We first realize the preliminary fusion of the PET feature map u1 and the CT feature map u2 by splicing.

$$u_c = \varsigma(u_1, u_2) \tag{1}$$

Then, the fused feature maps are input into a global average pooling layer (GAP) and converted into channel descriptors Z_c, where H and W represent spatial dimensions. In this way, the channel's statistical properties can be well obtained from a global

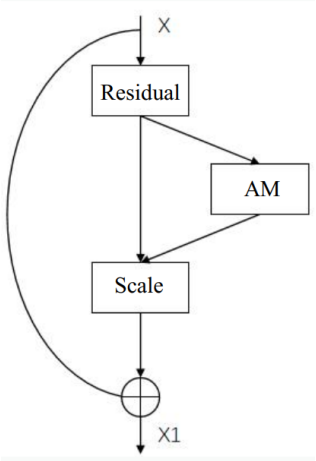

Figure 2: Improved residual network

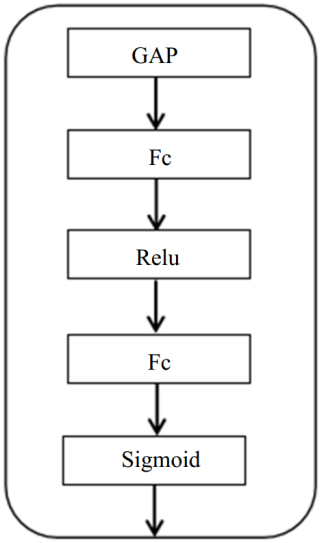

Figure 3: Structure of AM

perspective.

$$Z_c = F_{sq}(u_c) = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{2}$$

Next, two fully connected layers, W1 and W2, are connected to the formed channel descriptors to exploit the channel-wise interdependencies further. To fully obtain channel dependencies, the module needs to learn non-mutually exclusive relations and nonlinear interactions between channels. Therefore, the first fully connected layer is modulated by the rectified linear unit ReLU, and the sigmoid function modulates the second. After passing through two fully connected layers, a probabilistic form of attention descriptor S_c is generated, each element of which reflects the saliency and importance of the relevant channels of the input feature map. This attention descriptor is used as a weight to modify the input feature map.

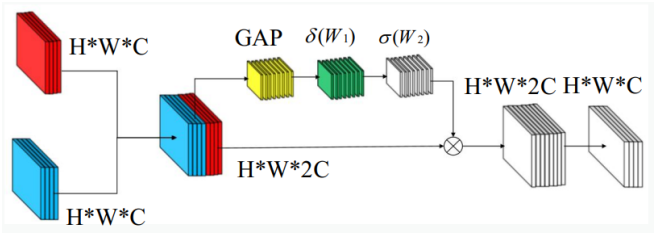

Figure 4: Multimodal Self-adaptive Attention Feature Fusion structure diagram

$$S_c = F_{ex}(z, W) = \sigma(W_2 \delta(W_{1Z})) \tag{3}$$

Finally, an informative feature-emphasized feature map X is generated by multiplying the attention descriptor channel-wise by the input feature map.

$$X_c = s_c * u_c \tag{4}$$

3.3 Region Proposal Network

The role of RPN (Region Proposal Network) is mainly used to extract preselected boxes. We know that a performance bottleneck in R-CNN and Fast R-CNN is extracting preselected boxes, and RPN optimizes this part very well. The reason is that it introduces the convolutional neural network and generates the position of the preselection box in the form of feature extraction, thus reducing the computational time overhead caused by the selective search algorithm.

RPN learns to generate candidate boxes from the feature maps generated by the feature extraction network. Specifically, it operates on the feature map through a sliding window of size m*n. For each sliding window, a specific set of anchor boxes is generated with three different aspect ratios (1:1, 1:2, 2:1 in the original text) and three different aspect ratios (128, 256, and 512 in the original text)) box, so there are nine schemes for each pixel in the feature map. In [42], based on the different scales of the lesions, the author selected three different length-width ratios (1:2, 2:1, 1:3) and four different length-width boxes (64, 128, 256, 512) through multiple experiments. In this paper, based on the different scales of the pancreatic lesions, we have selected five boxes with various lengths and widths (16, 32, 64, 128, 256) and five different aspect ratios (0.4:1, 0.8:1, 1:1, 1.5:1, 2:1) after many experiments.

3.4 Classification and Regression Network

The classification and regression network uses the region proposals provided by the RPN and the feature maps output by the feature extraction network as input to detect lesions. However, the sizes of region proposals are different, leading to the loss of lesion detection accuracy. So we take a pooling operation to convert them to the same size. Among them, the bilinear interpolation method is used to avoid the influence of ROI Pooling quantization on the positioning. Specifically, the mapped coordinates are floating-point numbers, which will be rounded by quantization. The rounding will cause the target feature block to be reduced or enlarged to introduce noise, resulting in positioning errors.

4 EXPERIMENTS AND DISCUSSIONS

4.1 Data

We use two datasets to test the performance of our algorithm. One is the pancreatic cancer dataset, and the other is the head and neck cancer dataset. The pancreatic cancer dataset is provided by Changhai Hospital. which contains 880 PET/CT images of 93 cases, and the size of each image is 512*512 pixels, of which 792 are used for training, and 88 are used for testing. There is no overlap between the training and testing sets, and experienced doctors label all images. The head and neck cancer dataset is a segmentation dataset publicly available from HECKTOR 2020. The pixel values of the segmentation labels are detected to obtain accurate bounding box labels. The dataset contains 1613 PET/CT images of 224 cases with a resolution of 512*512, and we use it to verify the generalization ability of our model. We normalize the dataset before feeding it into the feature extraction network. To obtain stable performance, we adopt cross-validation in the experiment. Among them, to ensure the amount of training data, ten-fold cross-validation is selected for pancreatic cancer experiments, and five-fold cross-validation is chosen for head and neck cancer-related experiments due to relatively abundant data and to balance the amount of calculation verified.

4.2 Experimental setup

The method proposed in this paper is implemented in Python 3.7 using PyTorch. We used the Faster R-CNN framework in this pancreatic cancer experiment and adopted the stochastic gradient descent method with a momentum optimizer. The momentum value is 0.9. The initial learning rate is 0.01. The learning rate decays every five epochs. The decay coefficient is 0.95. And a total of 100 epochs are trained. According to the tumor diameter distribution, we selected anchor box areas of 16^2, 32^2, 64^2, 128^2, 256^2, 512^2, a total of 6 scales, 0.4:1, 0.8:1, 1:1, 1.5 :1,2:1 total 5 anchor ratios. In the head and neck cancer experiment, only the anchor ratio was changed to 0.5:1, 1:1, 1.5:1, 2:1, and the rest remained unchanged. The hardware configuration is Intel (R) Core i9-10900K CPU, Nvidia GeForce RTX3080 Ti GPU, 12GB memory, and support Windows 64-bit.

4.3 Evaluation Criteria

This paper uses several evaluation indexes commonly used in target detection tasks: Joint intersection (IOU), precision, recall, AR (average recall, only 100 predicted AR values are allowed in each image) and mAP (mean average precision). The intersection between the predicted box Bp calculated for each result, and the corresponding ground-truth box Bgt is defined as IOU. The detection results with IOU greater than 0.5 are regarded as valid results. This paper uses mAP and mAP_small ($<20^2$) indicators in COCO indicators to verify the model's pros and cons. The mAP is usually used as object detection's final performance evaluation index. The specific calculation is shown in the formula:

$$\mathrm{IOU} = \frac{\mathrm{Bp} \cap \mathrm{Bgt}}{\mathrm{Bp} \cup \mathrm{Bgt}} \tag{5}$$

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$AP = \sum_{i=1}^{n-1} (R_i + 1 - R_i) P(R_i + 1) \tag{8}$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{9}$$

where P represents the precision, R represents the recall. TP represents the number of positive samples predicted to be positive samples, and FP represents the number of negative samples expected to be positive. FN represents the number of negative samples predicted to be negative samples.

4.4 Comparison of Ablation Experiments

In this experiment, the multi-modal detection network obtained by adding dual Resnet-50 feature extraction networks and splicing based on Faster R-CNN will be used as the multi-modal baseline. After adding the feature pyramid network, as shown in Table 1, the detection results are improved from mAP: 46.9% to mAP: 80.43%. This aspect verifies the defect of the primary network's poor ability to extract contextual information. On the other hand, it also shows that the feature pyramid network uses a multi-scale fusion method to obtain rich contextual information to improve the detection effect of lesions. Among them, we can see that the improvement of the detection results of small lesions is more significant, from mAP: 25.9% to mAP: 66.25%. This shows that the progress of the detection effect of small lesions is key to improving the overall lesion detection effect.

However, we can also see from it that the detection effect of only adding the feature pyramid network is still insufficient. There is still a gap of about 5% mAP in the overall detection of our final improved algorithm and a 9%mAP difference in detecting small lesions. There are two main reasons for this. The first reason is the feature redundancy and low utilization rate caused by the simple splicing of multimodal features. The second reason is that the ROI pooling in Faster R-CNN has its quantization operation, and the bounding box will have two errors in the regression process. We first designed Multimodal Self-adaptive Attention Feature Fusion (MAFF) based on these two points. The essence is that after the multimodal feature map is spliced, the feature semantic information selection is made more focused so that context-rich and prominent feature maps are obtained by recalibrating the multimodal feature maps. Secondly, ROI Align is introduced, using the bilinear interpolation method, avoiding the influence of quantization on positioning so that the detection effect has been improved to a certain extent. At the same time, because the algorithm parameters added to the feature pyramid network are too large, an attention mechanism is introduced in the Resnet-50 to focus on the more critical features of the current task, reduce the attention to other parts, and improve the efficiency and accuracy of task processing. Each module has improved the lesion detection results to varying degrees from the ablation experiment results.

4.5 Comparison with different algorithms and modals

To further evaluate the performance of our proposed algorithm, we conduct comparative experiments with some different algorithms, as shown in Table 2. Devaguptapu [43] selected Faster RCNN as the

Table 1: Performance comparison of ablation experiments (pancreatic cancer)

MAFF	AM	ROI Align	FPN	mAP	$mAP^{small}(<20^2)$
				46.90%	25.90%
			√	80.43%	66.52%
		√	√	82.47%	72.86%
	√	√	√	83.53%	73.97%
√	√	√	√	85.01%	75.98%

Table 2: Comparison of detection performance of different algorithms (pancreatic cancer)

Methods	mAP	mAP^{small} $(<20^2)$
Devaguptapu [43]	47.12%	25.90%
Pei [44]	82.26%	61.71%
Ming [45]	62.98%	47.97%
OUR	85.01%	75.98%

base frame. It has a poor ability to extract contextual feature information in medical images containing many small lesions. Secondly, in terms of multimodal feature fusion, only using splicing will cause feature redundancy, and the effective feature utilization is limited. Finally, the problem of quantization error in the ROI Pooling layer remains unsolved. Therefore, it can be seen that the overall detection effect is inferior. Pei [44] is based on RetinaNet and uses focal loss to divide positive and negative samples. The author uses post fusion after FPN. The two subnets have the same Conv+BN+ReLU, ResNet and FPN settings. The difference is that the weights and deviations of these networks are automatically updated according to the characteristics of the input images. The author sums the five scales of FPN output to feature maps of different scales to achieve feature fusion. Ming [45] selected YOLO-V5 as the base frame. Instead of using the traditional dual feature extraction network, the PET and CT datasets are fused into one dataset before inputting the data and then sent to the network for detection. As for the algorithm itself, the backbone network adopts a pyramid structure, but the C2 layer is also abandoned. From the experimental results, this method is not ideal for the detection of pancreatic cancer and head and neck cancer.

At the same time, we also conducted comparative experiments on CT, PET, and PET/CT, as shown in Table 3. Based on our multimodal algorithm, we changed the dual feature extraction network to a single feature extraction network and removed the adaptive fusion module. Because the texture structure of the lesion is very similar to the surrounding tissue, the detection effect of single-modal CT is much worse than that of single-modal PET. And there is still some gap between PET images and PET/CT due to blurred edges(Table 3). Compared with different algorithms and modes, Our multimodal algorithm has achieved the best results. It shows that Multimodal Self-adaptive Attention Feature Fusion combines the advantages of the high contrast of PET image and sufficient structural information of high spatial resolution of CT image, improves the problems of fuzzy tumor edge of single-mode PET image and low contrast of CT image lesions and surrounding tissues, and can effectively mine the spatial and texture features of lesions.And as shown in Table 4 and Table 5, this result is also verified on the head and neck cancer dataset, indicating that our algorithm has certain applicability.

5 CONCLUSION

In this paper, we propose a Multi-scale and self-adaptive attention feature fusion network for pancreatic lesion detection in PET/CT images. Compared with the basic Faster R-CNN, our improvements are mainly reflected in the multimodal and MAFF modules. By adding a dual feature extraction network, the combined use of PET features and CT features are realized. On this basis, we propose multi-scale adaptive feature fusion. First, through feature pyramid networks, propagating low-level precise localization information and capturing richer contextual information at multiple scales. Then, an attention module is introduced to perform preliminary feature screening on the PET and CT feature maps, respectively. Finally, it is input to the adaptive feature fusion module to recalibrate the multimodal feature maps so that the feature semantic information selection is more focused and contextual information-rich and prominent feature maps are obtained. At the same time, we introduce ROI Align and use bilinear interpolation to avoid the problem of ROI Pooling quantifying the impact on the positioning. Comprehensive evaluation and comparison show that the method has good detection performance. In the future, we will continue to research pancreatic cancer diagnosis and assist doctors in clinical diagnosis.

ACKNOWLEDGMENTS

This study was funded by the National Natural Science Foundation of China (No. 62101551, 62001417) and completed in the Suzhou Institute of Biomedical Engineering Technology, Chinese Academy of Sciences.

REFERENCES

[1] Kamisawa T, Wood L D, Itoi T, *et al.* Pancreatic cancer [J]. The Lancet, 2016, 388(10039): 73-85.
[2] Mizrahi J D, Surana R, Valle J W, *et al.* Pancreatic cancer [J]. The Lancet, 2020, 395(10242): 2008-2020.
[3] National Cancer Institute. Cancer stat facts: pancreatic cancer [J]. 2019.

Table 3: Comparison of detection performance of different modals (pancreatic cancer)

Methods	mAP	mAP^{small} ($<20^2$)
CT	66.11%	48.47%
PET	83.06%	73.57%
OUR	85.01%	75.98%

Table 4: Comparison of detection performance of different algorithms (head and neck cancer)

Methods	mAP	mAP^{small} ($<20^2$)
Devaguptapu [43]	56.89%	27.37%
Pei [44]	62.68%	40.26%
Ming [45]	70.40%	49.36%
OUR	79.90%	67.61%

Table 5: Comparison of detection performance of different modals (head and neck cancer)

Methods	mAP	mAP^{small} ($<20^2$)
CT	57.00%	33.23%
PET	78.99%	56.74%
OUR	79.90%	67.61%

[4] Ferlay J, Colombet M, Soerjomataram I, *et al.* Cancer statistics for the year 2020: An overview [J]. International Journal of Cancer, 2021, 149(4): 778-789.

[5] ChariST. Detecting early pancreatic cancer: problems and prospects [C]// Seminars in oncology. WB Saunders, 2007, 34(4): 284-294.

[6] Al-Hawary M M, Francis I R, Chari S T, *et al.* Pancreatic ductal adenocarcinoma radiology reporting template: consensus statement of the Society of Abdominal Radiology and the American Pancreatic Association [J]. Radiology, 2014, 270(1): 248-260.

[7] Kenner B, Chari S T, Kelsen D, *et al.* Artificial intelligence and early detection of pancreatic cancer: 2020 summative review [J]. Pancreas, 2021, 50(3): 251.

[8] Yoon S H, Lee J M, Cho J Y, *et al.* Small (≤ 20 mm) pancreatic adenocarcinomas: analysis of enhancement patterns and secondary signs with multiphasic multidetector CT [J]. Radiology, 2011, 259(2): 442-452.

[9] Dargan S, Kumar M, Ayyagari M R, *et al.* A survey of deep learning and its applications: a new paradigm to machine learning [J]. Archives of Computational Methods in Engineering, 2020, 27(4): 1071-1092.

[10] Su Y, Li D, Chen X. Lung nodule detection based on faster R-CNN framework [J]. Computer Methods and Programs in Biomedicine, 2021, 200: 105866.

[11] Cao W, Zheng J, Xiang D, *et al.* Edge and neighborhood guidance network for 2D medical image segmentation [J]. Biomedical Signal Processing and Control, 2021, 69: 102856.

[12] Mou L, Zhao Y, Fu H, *et al.* CS2-Net: Deep learning segmentation of curvilinear structures in medical imaging [J]. Medical image analysis, 2021, 67: 101874.

[13] Chen K, Xuan Y, Lin A, *et al.* Esophageal cancer detection based on classification of gastrointestinal CT images using improved Faster RCNN [J]. Computer Methods and Programs in Biomedicine, 2021, 207: 106172.

[14] Srinivasu P N, SivaSai J G, Ijaz M F, *et al.* Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM [J]. Sensors, 2021, 21(8): 2852.

[15] Ren S, He K, Girshick R, *et al.* Faster r-cnn: Towards real-time object detection with region proposal networks [J]. Advances in neural information processing systems, 2015, 28.

[16] X.u. Zhang, F. Chen, T. Yu, J. An, Z. Huang, J. Liu, W. Hu, L. Wang, H. Duan, J. Si, S. Diciotti, Real-time gastric polyp detection using convolutional neural networks [J], PLoSONE 14, (3), 2019, e0214133, https://doi.org/10.1371/journal.

[17] Chen B L, Wan J J, Chen T Y, *et al.* A self-attention based faster R-CNN for polyp detection from colonoscopy images [J]. Biomedical Signal Processing and Control, 2021, 70: 103019.

[18] Girshick R, Donahue J, Darrell T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[19] Girshick R. Fast r-cnn [C]// Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[20] He K, Gkioxari G, Dollár P, *et al.* Mask r-cnn [C]// Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[21] Redmon J, Divvala S, Girshick R, *et al.* You only look once: Unified, real-time object detection [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[22] Redmon J, Farhadi A. YOLO9000: better, faster, stronger [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

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

[24] Liu W, Anguelov D, Erhan D, *et al.* Ssd: Single shot multibox detector [C]// European conference on computer vision. Springer, Cham, 2016: 21-37.

[25] Liu S L, Li S, Guo Y T, *et al.* Establishment and application of an artificial intelligence diagnosis system for pancreatic cancer with a faster region-based convolutional neural network [J]. Chinese medical journal, 2019, 132(23): 2795.

[26] Zhang Z, Li S, Wang Z, *et al.* A novel and efficient tumor detection framework for pancreatic cancer via CT images [C]// 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020: 1160-1164.

[27] Wang Y, Tang P, Zhou Y, *et al.* Learning inductive attention guidance for partially supervised pancreatic ductal adenocarcinoma prediction [J]. IEEE transactions on medical imaging, 2021, 40(10): 2723-2735.

[28] Si K, Xue Y, Yu X, *et al.* Fully end-to-end deep-learning-based diagnosis of pancreatic tumors [J]. Theranostics, 2021, 11(4): 1982.

[29] Alves N, Schuurmans M, Litjens G, *et al.* Fully Automatic Deep Learning Framework for Pancreatic Ductal Adenocarcinoma Detection on Computed Tomography [J]. Cancers, 2022, 14(2): 376.

[30] Ma H, Liu Z X, Zhang J J, *et al.* Construction of a convolutional neural network classifier developed by computed tomography images for pancreatic cancer diagnosis [J]. World Journal of Gastroenterology, 2020, 26(34): 5156.

[31] Wehrend J, Silosky M, Xing F, *et al.* Automated liver lesion detection in 68Ga DOTATATE PET/CT using a deep fully convolutional neural network [J]. EJNMMI research, 2021, 11(1): 1-11.

[32] Hervella Á S, Rouco J, Novo J, *et al.* Retinal microaneurysms detection using adversarial pre-training with unlabeled multimodal images [J]. Information Fusion, 2022, 79: 146-161.

[33] Li S, Jiang H, Wang Z, *et al.* An effective computer aided diagnosis model for pancreas cancer on PET/CT images [J]. Computer methods and programs in biomedicine, 2018, 165: 205-214.

[34] Li L, Zhao X, Lu W, *et al.* Deep learning for variational multimodality tumor segmentation in PET/CT [J]. Neurocomputing, 2020, 392: 277-295.

[35] Kumar, A., Fulham, M., Feng, D. and Kim, J., Co-learning feature fusion maps from PET-CT images of lung cancer. IEEE Trans. Med. Imaging, 39(1), 2019, pp.204-217.

[36] Andrearczyk V, Oreiller V, Jreige M, *et al.* Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT[C]//3D Head and Neck Tumor Segmentation in PET/CT Challenge. Springer, Cham, 2020: 1-21.
[37] Yuan Y. Automatic head and neck tumor segmentation in PET/CT with scale attention network [C]//3D Head and Neck Tumor Segmentation in PET/CT Challenge. Springer, Cham, 2020: 44-52.
[38] Andrearczyk V, Oreiller V, Boughdad S, *et al.* Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images [C]// 3D Head and Neck Tumor Segmentation in PET/CT Challenge. Springer, Cham, 2021: 1-37.
[39] Oreiller V, Andrearczyk V, Jreige M, *et al.* Head and neck tumor segmentation in PET/CT: the HECKTOR challenge [J]. Medical image analysis, 2022, 77: 102336.
[40] Woo S, Park J, Lee J Y, *et al.* Cbam: Convolutional block attention module [C]// Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
[41] Lin T Y, Dollár P, Girshick R, *et al.* Feature pyramid networks for object detection [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
[42] Sha G, Wu J, Yu B. Detection of Spinal Fracture Lesions Based on Improved Faster-RCNN [C]// 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS). IEEE, 2020: 29-32.
[43] Devaguptapu C, Akolekar N, M Sharma M, *et al.* Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019: 0-0.
[44] Pei D, Jing M, Liu H, *et al.* A fast RetinaNet fusion framework for multi-spectral pedestrian detection [J]. Infrared Physics & Technology, 2020, 105: 103178.
[45] Ming Y, Dong X, Zhao J, *et al.* Deep learning-based multimodal image analysis for cervical cancer detection [J]. Methods, 2022.