*Computers, Materials & Continua*
DOI: 10.32604/cmc.2023.027627
***Article***

Tech Science Press

# DSAFF-Net: A Backbone Network Based on Mask R-CNN for Small Object Detection

**Jian Peng[1,2], Yifang Zhao[1,2], Dengyong Zhang[1,2,*], Feng Li[1,2] and Arun Kumar Sangaiah[3]**

[1]Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha, 410114, China
[2]School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, 410114, China
[3]School of Computing Science and Engineering, Vellore Institute of Technology (VIT), Vellore, 632014, India
*Corresponding Author: Dengyong Zhang. Email: zhdy@csust.edu.cn
Received: 22 January 2022; Accepted: 10 May 2022

**Abstract:** Recently, object detection based on convolutional neural networks (CNNs) has developed rapidly. The backbone networks for basic feature extraction are an important component of the whole detection task. Therefore, we present a new feature extraction strategy in this paper, which name is DSAFF-Net. In this strategy, we design: 1) a sandwich attention feature fusion module (SAFF module). Its purpose is to enhance the semantic information of shallow features and resolution of deep features, which is beneficial to small object detection after feature fusion. 2) to add a new stage called D-block to alleviate the disadvantages of decreasing spatial resolution when the pooling layer increases the receptive field. The method proposed in the new stage replaces the original method of obtaining the P6 feature map and uses the result as the input of the regional proposal network (RPN). In the experimental phase, we use the new strategy to extract features. The experiment takes the public dataset of Microsoft Common Objects in Context (MS COCO) object detection and the dataset of Corona Virus Disease 2019 (COVID-19) image classification as the experimental object respectively. The results show that the average recognition accuracy of COVID-19 in the classification dataset is improved to 98.163%, and small object detection in object detection tasks is improved by 4.0%.

**Keywords:** Small object detection; classification; RPN; MS COCO; COVID-19

## 1 Introduction

Deep learning has become the most efficient technology in computer vision recently. It shows great advantages in image recognition, object tracking, et al. Object detection is the basic task of it, and also it is one of its core tasks. The task of object detection is to identify the objects of interest in an image and determine their categories and positions. In other words, it is to answer the question of “where? “And” What is it?” Traditional object detection methods were generally divided into three steps. First of all, this method generally uses the selective search algorithm [1] to obtain candidate areas using sliding window frames of different sizes. Then, the method adopts diverse methods to extract the relevant visual features of the candidate area, such as harr features [2] frequently used in face detection, histogram of oriented gradient (HOG) features [3] popularly used in pedestrian detection, and public object detection, scale-invariant feature transform (SIFT) algorithm [4] for detecting local features et al. Finally, use a trained classifier, such as a support vector machine (SVM) classifier [5], to classify. However, the traditional object detection methods have many defects, such as slow detection speed, low accuracy, poor real-time performance, et al.

With the development of technology, object detection method has veered from the traditional algorithm to the deep neural network technology. In 1998, LeNet-5 [6] proposed by LeCun first successfully applied convolutional neural networks (CNNs) to image recognition and achieved good performance in letter recognition. Deep learning has been greatly promoted because of its emergence. In 2012, Krizhevsky et al. from the University of Toronto proposed the structure of AlexNet [7], which not only attracted widespread attention of people to convolutional neural networks but also had milestone significance for image processing research based on convolutional neural networks. CNNs have greatly contributed to the development of computer vision, such as image retrieval [8], object detection, et al. The object detectors based on CNNs are separated into two categories: 1)one-stage detectors like you only look once (YOLO) [9], the single-shot multi-box detector (SSD) [10], and RetinaNet [11], which do not need to specially design a network to find candidate regions, but can directly extract features to predict the category and regression of objects. 2) two-stage detectors like region-based CNN (R-CNN) [12], Fast R-CNN [13], et al. This approach is implemented in two steps. The network first obtains proposal boxes (boxes that may contain the object to be detected) and then recognizes the category and location regression information of the object in proposal boxes. Comparing the two algorithms, the former has more advantage in speed, and the latter has higher accuracy. However, with the continuous optimization of object detection methods, accuracy and speed have been greatly improved. Meanwhile, object detection has also been used to do specific types of detection, such as intelligent transportation, face detection [14], and text detection [15], et al. Excellent object detection provides reliable information for more elaborate computer vision tasks studies.

Small object detection is a significant part of object detection. It widely exists in a large range, long-distance, and other imaging pictures. There are two official definitions of a small object: 1) Relative size. A small object means that the object area in a $256 \times 256$ image is less than 80 pixels, that is, less than 12% of $256 \times 256$ is a small object. 2) Absolute size. Microsoft Common Objects in Context (MS COCO) [16] defines that an object with a size smaller than $32 \times 32$ pixels can be regarded as a small object. Most object detectors based on CNNs use public datasets for detection. However, the edge features of small objects in pictures are easily blurred or even missing, which makes the small object detection algorithms inefficient on public datasets. Therefore, a greatly increased number of experts and scholars propose methods to optimize small object detection. In 2014, Goodfellow et al. proposed the generative adversarial networks (GANs) [17], which have brought some major technological breakthroughs to deep learning and are widely used in image generation, information steganography [18], object detection, and other fields. GANs improve the detection performance of small objects by expanding the characteristics of large objects and reducing the representation difference between small objects and large objects. In addition, the current ideas to optimize small object detection also include data enhancement [19], feature fusion [20,21], using context information [22,23], appropriate training methods [24], more denser anchor sampling and matching strategy [25,26].

This paper mainly researches two aspects. First, we present a three-way feature attention fusion module–sandwich attention feature fusion module (SAFF module). Its purpose is to improve the resolution of deep features and strengthen the semantic information of shallow features, and effectively combine with the feature pyramid network (FPN) [27] to optimize detection and regression, especially for small object feature processing. In addition, we create a new stage in the backbone network by dilated convolution [28] to alleviate the disadvantage of loss of resolution when pooling expands the reception field.

The structure of this paper is as follows: Related work is shown in Section 2 about object detection techniques. In Section 3, the proposed methodology, including feature extraction, receptive filed expansion, and the overall detection are introduced. The experimental setup, validation, and results, as well as comparative analysis with other techniques, are in Section 4. Finally, Section 5 summarizes the conclusion of the paper.

## 2 Related Work

### *2.1 Backbone Network*

Simonyan et al. proposed the visual geometry group (VGGNet) [29] in 2014. The structure alternately uses 3 × 3 convolutional kernels and 2 × 2 maximum pooling layers to deepen the network to 19 layers, thus the performance of CNNs is dramatically improved. In the same year, the inception module was presented by Szegedy et al. and built the GoogleNet [30] on this basis. The number of network layers has reached an unprecedented layer of 22. Although the network becomes deeper, it does not mean that the experimental effect is better. The network is prone to overfit, computational resource consumption, gradients disappearing, and other problems. The network performance degrades as the number of network layers increases. To address these problems, many experts and scholars have studied and explored them in many ways. In 2015, He et al. presented a residual network structure (ResNet) [31], that is, adding shortcut connections to the forward neural network. A shortcut connection can be regarded as a sample equivalent mapping. The input signal can be directly propagated from any low layer to the high layer without generating additional parameters or increasing computational complexity. There is no doubt that it can, to a large degree, improve the problem of network degradation. At the same time, the training network can still pass the end-to-end backpropagation algorithm to alleviate the troublesome of gradient disappearance (even if the weight of the intermediate layer matrix is small, the gradient will not disappear). Therefore, the residual network not only enables us to train deeper layers, but also guarantees good network performance, and further makes the network layer depth to a new height.

### *2.2 Object Detectors Based on Image Classification*

Currently, the feature extraction network is mainly coming from image classification, so the accuracy of classification will exert a considerable influence on object detection. But object detection is not classification, which includes two tasks: classification and positioning, there are two weaknesses with using classified networks as the backbone network of object detection: 1) Information about the small object is easy to lose. In feature extraction, the shallow feature maps have high resolution and can return object location relatively accurately, but the semantic information is too weak to adequately identify the object. Instead, the deep feature maps have strong semantic information and low resolution, which is not conducive to object regression. In response to this point, the FPN [27] structure was designed by Lin et al. in 2017 to improve the detector's shortcomings in dealing with multi-scale changes. The FPN structure effectively combines the shallow and deep features to promote the semantic expression ability of the shallow features and the resolution of the deep features. After that, He et al. proposed Mask R-CNN [32], which uses the FPN structure to further advance Faster R-CNN [33], and the processing layers of the proposal boxes are changed from single to multi-layer. The accuracy of bounding box regression and the performance of small object detection are greatly improved. 2) In the Mask R-CNN network, the feature extraction network based on residual network and feature pyramid network (ResNet-FPN) results (P2, P3, P4, P5, P6) are taken as input to the regional proposal network (RPN). P6 is only used to process anchors of 512 × 512, which is obtained by P5 through maximum pooling down-sampling with a step size of 2. Although the method of obtaining P6 expands the receptive field of corresponding pixels on the feature maps, it will cause some parameters that cannot be learned and lose part of the spatial resolution, which is not conducive to accurately locating large objects and identifying small objects.

### *2.3 Attention Mechanism*

In recent years, the attention model has been widely used in various types of deep learning tasks such as natural language processing, image recognition, and speech recognition, and is one of the most noteworthy core technologies in deep learning technology. The attention mechanism is to imitate the way humans observe things, deepen and highlight local information, and select more critical information to the current mission goal from much information. For example, when people observe a picture because each person's attention or focus on the object of observation is different, certain local information that different people pay attention to is also different. The attention mechanism needs to decide which part of the input information needs more attention throughout the whole paragraph and then extracts features from the key parts to get more important information. The attention mechanism has achieved good results in computer vision tasks such as image segmentation and object detection.

## 3 The Proposed Method

The main purpose of this paper is to advance the shortcomings of object detectors based on the image classification backbone and increase the detection accuracy of small object detection by fusing better features. In this paper, the backbone network of feature extraction is improved on Mask R-CNN of the two-stage detector.

### *3.1 SAFF Module*

It is well known that the detection performance of small objects based on the MS COCO dataset is far inferior to that of large objects. There are several reasons for this: 1) Features of network feature extraction. Object detection networks usually use CNNs for feature extraction. The deeper the network layer is, the larger receptive field of pixel points on the feature map and the stronger the semantic information of the feature, but the size of the feature map also decreases. There is no doubt that the information of a small area, the feature information of the small object, is hard to be transmitted to the later stage of the object detector because the feature maps become smaller. As a result, small object features are difficult to extract or even disappear, and their detection performance is naturally poor. 2) Unbalanced distribution of objects with various sizes in datasets. The proportion of large and small objects in the MS COCO dataset is unbalanced, and the number of large objects is far more than small objects, which makes detection networks based on deep learning not very friendly to small object detectors, which also brings some difficulties for the network to adapt to different size objects. 3) Network loss function. When positive and negative samples are selected, the network loss function is not friendly to small objects.

For object detection algorithms, the underlying features of an image generally refer to features such as contours, edges, colors, textures, and shapes that reflect the general condition of the object, and they are mostly found in the shallow feature map of the detection network, so the features in such feature layers facilitate location regression for object detection. Semantic information can be simply understood as what we can see, for example, detecting a face in the shallow feature map, we can extract information such as the outline of the face, nose, eyes, etc. The deeper the feature layer, the richer the semantic information, the stronger the network's ability to identify the object, but the deeper features have low resolution, which is not conducive to the location regression of the object. In other words, the features used to detect small objects should have both high resolution and strong semantic information, which is very difficult for the network, so to ease the relationship between the two, we designed a three-way feature attention fusion module–sandwich attention feature fusion module (SAFF module), which is formed by the alternating superposition of two channel attention mechanisms and a spatial attention mechanism, see Fig. 1 below. Its purpose is to enhance the semantic information of features in shallow feature maps and improve the resolution of features in deep feature maps.

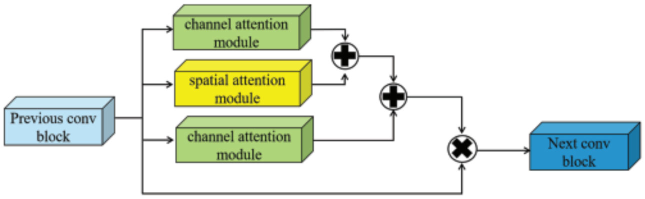

**Figure 1:** SAFF module architecture

### *3.1.1 Channel Attention Mechanism*

The channel attention mechanism, such as squeeze-and-excitation networks (SENet) [34], can be simply understood as what the neural network wants to see. The mechanism focuses on the correlation between feature channels, and automatically obtains the feature information of each channel through learning, Fig. 2 shows its structure. Then, according to the importance of features, this mechanism will enhance useful features and discard useless information, to advance the performance of the model. If channel attention is added to the shallow feature map, the expression ability of features can be improved and their semantic information can be enhanced. However, after adding channel attention, the global average pooling (GAP) [35] may lead to a part of the spatial information loss, the lack of interdependencies between the channel dimension and the spatial dimension. To reduce these losses, the SAFF module superimposes a layer of spatial attention mechanism to avoid adverse effects on the image position information. The spatial attention mechanism can be understood as where a neural network is looking. It transforms the original image's spatial information into another space and retains its key information. For deep features, they do not lack rich semantic information for object classification, they lack information that can ensure the accurate regression of the detection position. Introducing spatial attention will, to a large extent, can help the feature maps of different network layers to retain the location information of the features. Therefore, this information will not be lost too much in the deep feature map, and alleviate the drawbacks of object regression of the final detection network.

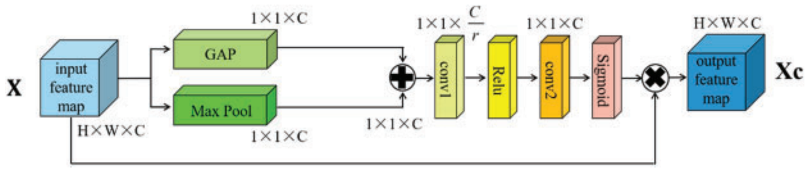

**Figure 2:** Channel attention details. “r” represents channel compression ratio, “Xc” means output

First, the GAP and max pooling operations are performed on the input feature map X to obtain two channel features with a size of 1 × 1 × C respectively. GAP compresses the global information into a real number. The real number, to some extent, has a receptive field of global information, which directly endows each channel with actual category meaning. GAP also greatly reduces the network parameters. Maximum pooling, that is, taking the maximum value of each block, means extracting the relatively strongest features and discarding other weak feature information to enter the next layer. In the second step, superimpose the features obtained by GAP and maximum pooling, and input them to the next convolutional layer. The first convolutional kernel is 1 × 1 × C/r, which compresses the channel to C/r of its original size and reduces the dimension of the feature map. Then, the method adopts the rectified linear unit (Relu) [36] to activate the resulting feature map, which increases the nonlinearity of the extracted features and improves their feature expression ability. After that, the feature map obtained in the last step is performed by using a filter with sizes 1 × 1 × C to reduce the number of channels to C to increase the dimension of the feature map. Finally, a Sigmoid activation function captures the important information in the channel, enhances the effective feature information, suppresses the irrelevant features, and obtains a new feature layer after scaling.

The output features after processing can be expressed as follows. Where $\sigma$ represents Relu nonlinear activation function and δ represents the Sigmoid nonlinear activation function

$$C(x) = \delta\left(Conv\left(\sigma\left(Conv\left(GAP(x) + MaxPool(x)\right)\right)\right)\right) \tag{1}$$

Through shortcut, the enhanced C(x) and original input feature map are multiplied by phase, and then get a new fused feature F1(x).

$$F1(x) = C(x) * X \tag{2}$$

Similarly, the feature map F3(x) after the attention of the next channel is strengthened is obtained.

### *3.1.2 Spatial Attention Mechanism*

The spatial attention mechanism is different from the channel attention mechanism in that it focuses on enhancing the position information of features. Firstly, two feature maps with the same dimension are obtained by using two different approaches, GAP and global maximum pooling (GMP). Then, these two feature maps are merged to get a special feature map. After that, the feature map undergoes a dimensionality reduction operation through a convolutional layer, and a spatial matrix with spatial attention weights is obtained. Finally, the matrix with spatial weight is multiplied by the original feature map, as shown in Fig. 3.

The new feature layer F2(x) after obtaining space reinforcement can be expressed in the following formula:

$$S(x) = \delta\left(Conv\left(AvePool(x);\ MaxPool(x)\right)\right) \tag{3}$$

$$F2(x) = S(x) * X \tag{4}$$

The input feature map X, after passing the SAFF Module, will get a feature map $X'$ with enhanced channel information and spatial information at the same time.

$$X' = X * (F1(x) + F2(x) + F3(x)) \tag{5}$$

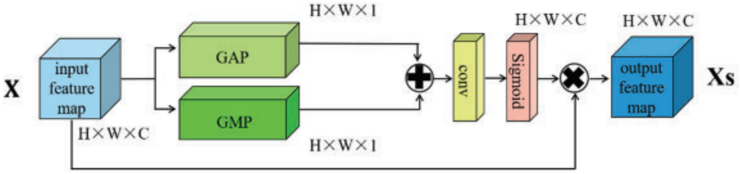

**Figure 3:** Spatial attention details, "Xs" represent the feature map after processing

### *3.2 D-Block Module*

The receptive field size of pixels on the feature map indicates the size of the area they correspond to in the original map. A large receptive field shows that the area in the original image is large. A small receptive field shows that the area in the original image is small. In an object detection task, if the receptive field is smaller than the feature area to be extracted, too much local information about the object will be obtained, resulting in a loss of global information and affecting the recognition of objects, such as pixel points on a shallow feature map. If the receptive field is larger than the area where the features are to be extracted, then this results in the object becoming background and being simply ignored, and no information about the object is extracted. Since the size of the object to be detected is different, it is important to select the right receptive field to obtain information of different sizes.

DetNet [37], proposed by Megvii Technology in 2018, is a backbone network specifically designed for object detection tasks from which we were inspired to design a D-block module. In ResNet-FPN, the backbone extraction network of Mask R-CNN, the P6 layer is specifically designed for the RPN network and is obtained from P5 by down-sampling. P6 is only used for the 512 × 512 proposal box and is not involved in the subsequent processing of the whole network, so the network does not pre-train the P6 layer and the parameters are not learned by the network. Using down-sampling to reduce the dimension, the network will only leave the information it considers important, resulting in some feature information loss. Therefore, the method of extracting P6 from the original network affects the effectiveness of object detection to some extent. Our proposed network retains stages 1–5 in the original feature extraction network and adds a new D-block module to obtain P6. The D-block module consists of two dilated residual blocks, a dilated convolutional block and a dilated identity block, as shown in Fig. 4 below. Using the D-block module to obtain P6, P6 can be trained by the network and its parameters can be learned. The dilated convolution also alleviates the disadvantages of down-sampling leading to a loss of part of the feature information and resolution to some extent. There is no doubt that the D-block module can optimize the effectiveness of object detection.

### *3.3 DSAFF-Net*

The DSAFF-Net presented in this paper uses the SAFF module and D-block module to modify the backbone structure of the object detection network to extract features, as shown in Tab. 1. Resnet-50 in the chart represents a 50-layer residual net and ResNet-101 represents a 101-layer residual net, Conv shows convolution with different sizes.

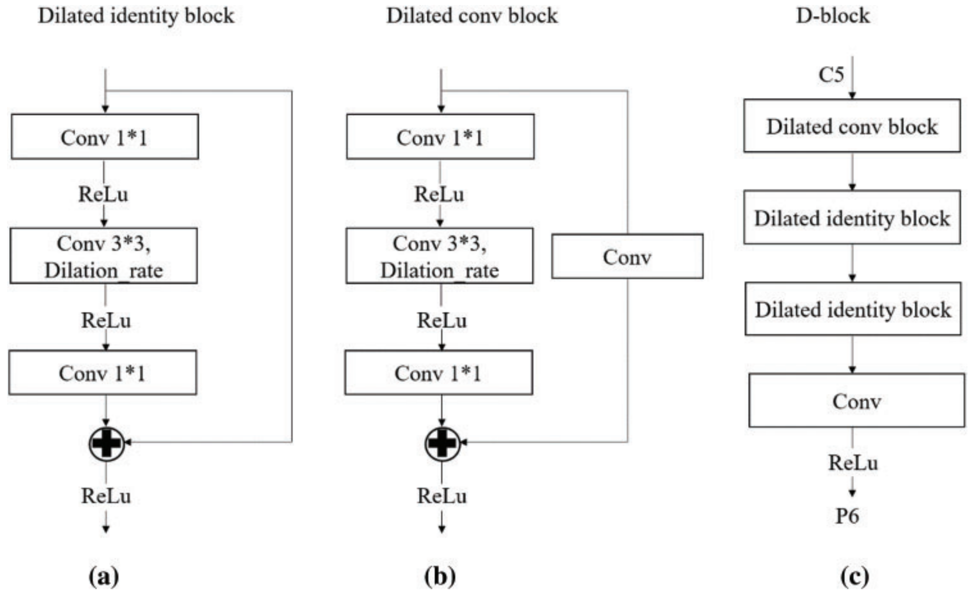

**Figure 4:** (a) Identity block with dilated convolution is for connecting the network. (b) Convolutional block with dilated convolution is for changing network dimension. (c) D-block module details

**Table 1:** DSAFF-Net feature extraction backbone details

| | Resnet-50 + SAFF module + D-block | | Resnet-101 + SAFF module + D-block | |
|---|---|---|---|---|
| Stage 1 | Conv 7 × 7, 64, stride = 2<br>Max Pool 3 × 3, stride = 2 | | | |
| Stage 2 | Conv 1 × 1, 64<br>Conv 3 × 3, 64<br>Conv 1 × 1, 256 | ×3 | Conv 1 × 1, 64<br>Conv 3 × 3, 64<br>Conv 1 × 1, 256 | ×3 |
| | SAFF module | | SAFF module | |
| Stage 3 | Conv 1 × 1, 128<br>Conv 3 × 3, 128<br>Conv 1 × 1, 512 | ×4 | Conv 1 × 1, 128<br>Conv 3 × 3, 128<br>Conv 1 × 1, 512 | ×4 |
| | SAFF module | | SAFF module | |
| Stage 4 | Conv 1 × 1, 256<br>Conv 3 × 3, 256<br>Conv 1 × 1, 1024 | ×6 | Conv 1 × 1, 256<br>Conv 3 × 3, 256<br>Conv 1 × 1, 1024 | ×23 |
| | SAFF module | | SAFF module | |
| Stage 5 | Conv 1 × 1, 512<br>Conv 3 × 3, 512<br>Conv 1 × 1, 2048 | ×3 | Conv 1 × 1, 512<br>Conv 3 × 3, 512<br>Conv 1 × 1, 2048 | ×3 |

(Continued)

**Table 1:** Continued

| | | | | |
|---|---|---|---|---|
| | SAFF module | | SAFF module | |
| D-block | D-conv1 × 1, 512<br>D-conv3 × 3, 512<br>D-conv1 × 1, 2048 | ×3 | D-conv1 × 1, 512<br>D-conv3 × 3, 512<br>D-conv1 × 1, 2048 | ×3 |
| | Conv 7 × 7, 256 | | Conv 7 × 7, 256 | |

## 4 Experimental Results and Analysis

The experiment is divided into two parts. One is to add the SAFF module to the 50-layer residual net (ResNet-50) and the 101-layer residual net (ResNet-101) to form new network structures which are ResNet-58 and ResNet-109, and carry out three classification experiment on the Corona Virus Disease 2019 (COVID-19) dataset. The other is based on Mask R-CNN, a new feature extraction network, DSAFF-Net, formed by fusing the SAFF module and the D-block module with ResNet-FPN for object detection on the MS COCO dataset, focusing on the results of small object detection.

### *4.1 Classification*

On the eve of the Spring Festival in 2020, COVID-19 [38,39], an acute respiratory infectious disease caused by novel coronavirus infection, is a global outbreak and is extremely contagious. With the normalization of the epidemic, most countries are facing huge pressures on public resources and medical resources. The pneumonia diagnostic kit (RT-PCR), the most widely used COVID-19 detection technology, has the limitations of high cost, time consumption, and low sensitivity. To help healthcare workers identify and classify COVID-19 quickly and correctly from countless pictures, the COVID-19 dataset was used as the study object. We download Chest X-Ray (CXR) images from the public image database [40] and construct a dataset, including a training set and test set. There are 5526 images in the training set, including 310 for COVID-19, 3875 for ordinary pneumonia, and 1341 for the normal image. The test set has 726 images, including 102 for COVID-19, 390 for ordinary pneumonia, and 234 for normal images. This paper shows the results of the new network classification through a confusion matrix, see Figs. 5 and 6.

Experimental effects use Accuracy, Precision, Recall, F1-score, and Cohen's Kappa coefficients as metrics for classification model evaluation. Accuracy shows how many positive samples are correctly predicted in the whole dataset. Precision represents how many of the predicted positive samples are correctly classified. Recall indicates how many of the true positive examples are correctly predicted. The F1-score evaluation metric is a harmonic average of precision and recall used to reconcile the extremes of the two, with a larger F1-score indicating a more effective model. The Kappa coefficient is based on the calculation of the classification confusion matrix and is used to measure the accuracy of the classification.

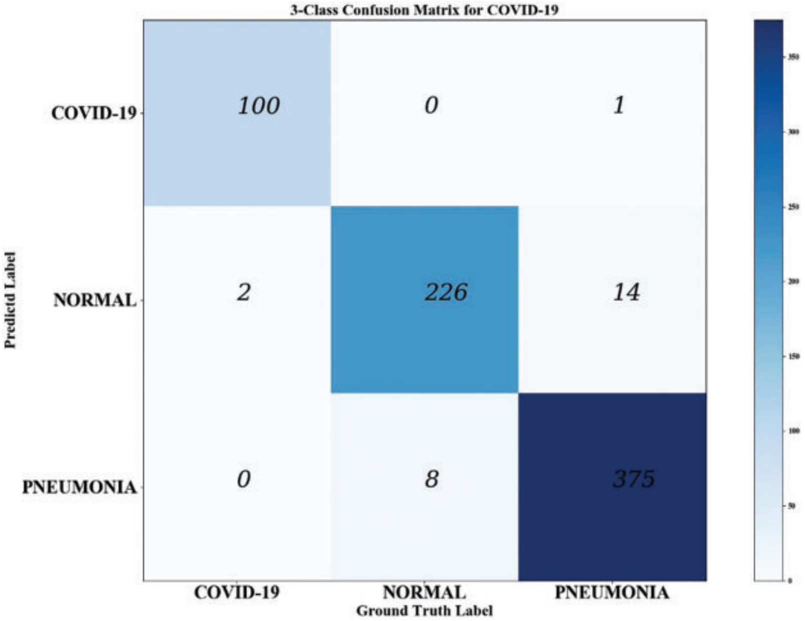

**Figure 5:** DSAFF-Net-58's triple classification confusion matrix for COVID-19

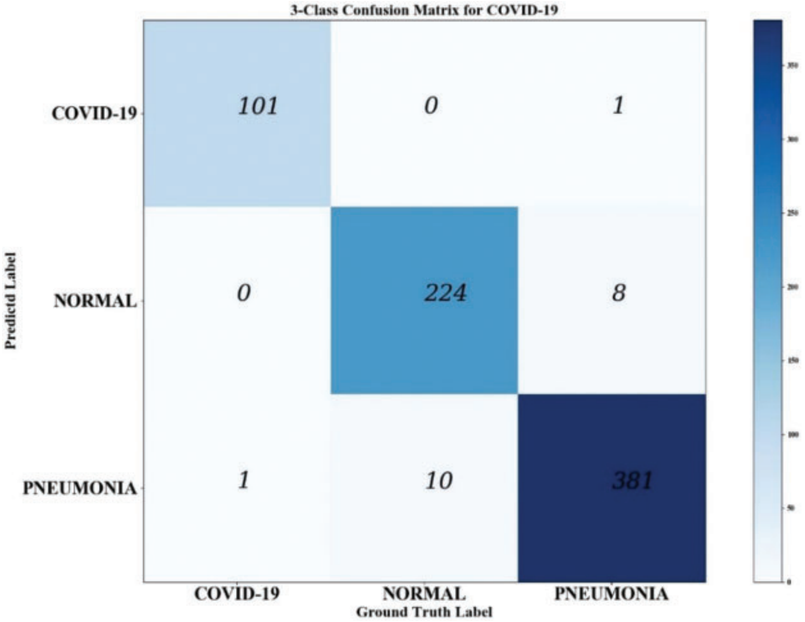

**Figure 6:** DSAFF-Net-101's triple classification confusion matrix for COVID-19

The calculation of each indicator is shown below

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

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

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

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{9}$$

$$kappa = \frac{p0 - pe}{1 - pe} \tag{10}$$

Where TP means the number of positive samples correctly predicted, FP shows the number of positive samples incorrectly predicted, TN denotes the number of negative samples correctly predicted and FN presents the number of negative samples incorrectly predicted as negative. p0 is the overall classification accuracy, and pe is the product of the number of correctly predicted and the actual number of categories in each category as a proportion of the square of the total number of samples. Tabs. 2–4 show specific data on the results of the classification experiments after feature extraction using the new method.

**Table 2:** Classification results for the three categories on the SAFF-Net-58

| Class | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Covid-19 | 99.586% | 99.010% | 98.039% | 98.522% |
| Normal | 96.694% | 93.388% | 96.581% | 94.958% |
| Pneumonia | 96.832% | 97.911% | 96.154% | 97.025% |
| Average | **97.704%** | **96.770%** | **96.925%** | **96.835%** |

**Table 3:** Classification results for the three categories on the SAFF-Net-109

| Class | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Covid-19 | 99.725% | 99.020% | 99.020% | 99.196% |
| Normal | 97.521% | 96.552% | 95.726% | 96.137% |
| Pneumonia | 97.245% | 97.194% | 97692% | 97.442% |
| Average | **98.163%** | **97.588%** | **97.588%** | **97.533%** |

**Table 4:** Average results of three classifications in different feature extraction networks for the COVID-19 dataset

| Model | Accuracy | Precision | Recall | F1-score | kappa |
|---|---|---|---|---|---|
| ResNet-50 | 91.427% | 85.338% | 90.731% | 86.833% | 77.183% |
| **SAFF-Net-58** | **97.704%** | **96.770%** | **96.925%** | **96.835%** | **94.160%** |
| ResNet-101 | 92.745% | 93.030% | 86.705% | 89.057% | 80.748% |
| **SAFF-Net-109** | **98.163%** | **97.588%** | **97.588%** | **97.533%** | **95.309%** |

### *4.2 Object Detection*

We assess the performance of DSAFF-Net in small object detection on the MS COCO 2014 datasets. The dataset contains 80 object categories, 80 k images in the training set, and 40 k images in the validation set. We divided the validation set with 40 k images into 35 k trainval datasets and 5 k minival datasets, and then tested on the minival datasets. Standards for evaluating network performance include average precision (AP) values and AP of different sizes ($AP_S$, $AP_M$, and $AP_L$ represent detector AP measurements for small, medium, and large objects, respectively).

101-layer residual net and feature pyramid network (ResNet-101-FPN) is the original feature extraction network in Mask R-CNN. Integration of the SAFF module and ResNet-101-FPN (DSAFF-Net-109) indicates that adding the SAFF module to the ResNet-101-FPN feature extraction plate, its feature extraction network is ResNet-101-FPN-SAFF. Integration of the D-block module and ResNet-101-FPN (DSAFF-Net-111) means adding the D-block module to replace the method of obtaining P6 from the RPN network in the original detection network, its feature extraction network is ResNet-101-FPN-D-block. Integration of D-block module and SAFF module with ResNet-101-FPN (DSAFF-Net-119) shows that the network adds both D-block and SAFF modules, its feature extraction network is ResNet-101-FPN-SAFF-D-block. To verify the advantages of DSAFF-Net in small object detection, we compared DSAFF-Net-109, DSAFF-Net-111 and DSAFF-Net-119 with Mask R-CNN. The result is represented in Tab. 5.

**Table 5:** DSAFF-Net experimental result. The detector uses the MS COCO dataset for training and detection. Comparison of Average Precision (AP) and AP with different bounding box scales

| Model | Feature extraction backbone network | mAP | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|
| Mask R-CNN | ResNet-101-FPN | 35.7% | 15.5% | 38.1% | 52.4% |
| DSAFF-Net-109 | ResNet-101-FPN-SAFF | 36.0% | **17.4%** | 41.5% | 53.9% |
| DSAFF-Net-111 | ResNet-101-FPN-D-block | 36.3% | 17.8% | 42.6% | **54.1%** |
| DSAFF-Net-119 | ResNet-101-FPN-SAFF-D-block | **36.7%** | **19.5%** | **42.5%** | **54.4%** |

For small objects, the information of shallow feature maps is the most favorable. SAFF module added in these layers is helpful to strengthen and preserve the small object features that are easily lost in a deep feature map and the detection of small objects by DSAFF-Net-109 is significantly improved. There is no doubt that the resolution of high-level feature maps has a great impact on the regression of large-scale objects. The higher resolution of high-level features, the more position information of objects is saved, and the better the detection effect of large objects. Tab. 5 is obvious that the detection performance of large objects has been significantly improved. Some of the results are shown in Fig. 7.

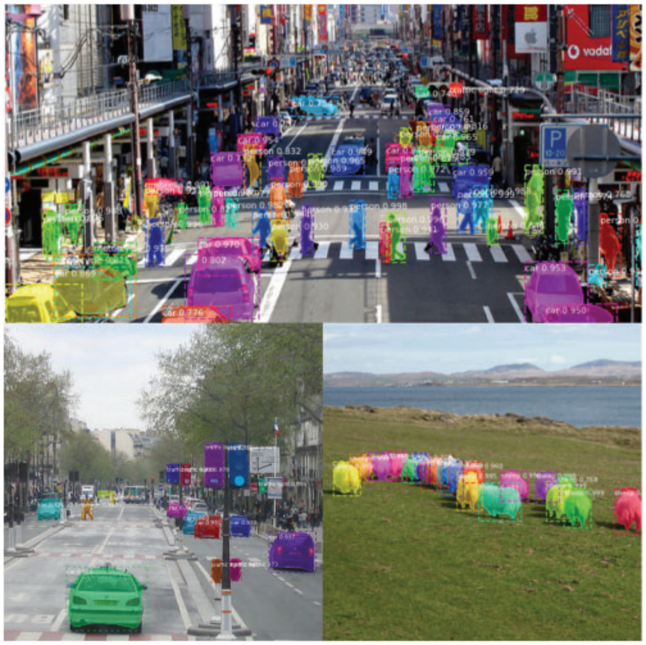

**Figure 7:** Visual results for DSAFF-Net

## 5 Conclusion

This study uses the advantages of dilated convolution and attention mechanism to obtain better features in different feature layers and expands the receptive field without losing resolution, thereby improving the object detection network based on image classification feature extraction. Effectively improve the accuracy of object detection results, especially small object detection.

We used classification experiments and object detection experiments to test the effectiveness of DSAFF-Net, on the one hand, verifying whether the new feature extraction method can extract better features will benefit the final classification results. On the other hand to test whether this method improves the extraction of small object features to some extent and optimizes the detection effect of the network. In terms of classification experiments, there are many limitations to the recognition methods of COVID-19 images just after the outbreak of the epidemic. In order to help health care workers quickly and correctly identity and classify COVID-19 images from countless pictures, improve speed and accuracy, which can greatly save costs and help control the spread of viruses, this paper uses the COVID-19 dataset as the research object of classification experiments. For the object detection experiments, we use the original Mask R-CNN running environment, language, settings of various parameters, and the same dataset, the public dataset MS COCO 2014, as the experimental object. This gives a more direct view of the improvement in the effectiveness of our proposed new strategy on Mask R-CNN for small objects.

Although DSAFF-Net has some effect on extracting better features for classification and target detection, further research and experiments are needed. For example, the small number of dataset images used in classification experiments may lead to some bias in model recognition. More pictures are needed to construct a larger dataset, which can be optimized according to practical applications. A new P6 method is obtained by using dilatation convolution construction so that its parameters can be trained in the network, but its effect is not as good as expected, possibly due to the void rate, which needs further study.

**Funding Statement:** This project is supported by the National Natural Science Foundation of China under grant 62172059 and 62072055, Hunan Provincial Natural Science Foundations of China under Grant 2020JJ4626, Scientific Research Fund of Hunan Provincial Education Department of China under Grant 19B004, "Double First-class" International Cooperation and Development Scientific Research Project of Changsha University of Science and Technology under Grant 2018IC25, and the Young Teacher Growth Plan Project of Changsha University of Science and Technology under Grant 2019QJCZ076.

**Conflicts of Interest:** The authors declare no conflicts of interest regarding the present study.

## References

[1] K. E. Van de Sande, J. R. Uijlings, T. Gevers and A. W. Smeulders, "Segmentation as selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2011.

[2] P. Viola and M. Jones, "Rapid objection detection using a boosted cascade of simple features," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, HI, USA, pp. 511–518, 2001.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, no. 1, pp. 886–893, 2005.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] S. R. Gunn, "Support vector machines for classification and regression," *ISIS Technical Report*, vol. 14, no. 1, pp. 5–16, 1998.

[6] Y. LeCun, L. Bootou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[7] A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 37, no. 9, pp. 1097–1105, 2012.

[8] X. B. Shen, G. H. Dong, Y. H. Zheng, L. Lan, Q. S. Sun *et al.,* "Deep co-image-label hashing for multi-label image retrieval," *IEEE Transactions on Multimedia*, vol. 24, pp. 1116–1126, 2022.

[9] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 779–788, 2016.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "SSD: Single shot multibox detector," in *Proc. of the European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 21–37, 2016.

[11] T. Y. Lin, P. Goyal, R. Girshick and K. He, "Focal loss for dense object detection," in *IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 2980–2988, 2017.

[12] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 580–587, 2014.

[13] R. Girshick, "Fast R-CNN," in *IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 1440–1448, 2015.

[14] A. Verma, M. Baljon, S. Mishra, I. Kaur, R. Saini *et al.,* "Secure rotation invariant face detection system for authentication," *Computers, Materials & Continua*, vol. 70, no. 1, pp. 1955–1974, 2022.

[15] H. P. Wu, Y. L. Liu and J. W. Wang, "Review of text classification methods on deep learning," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1309–1321, 2020.

[16] T. Y. Lin, M. Maire, S. Belongie, J. Hays and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. of the European Conf. on Computer Vision*, Zurich, Switzerland, pp. 740–755, 2014.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley *et al.,* "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2014.

[18] C. L. Wang, Y. L. Liu, Y. J. Tong and J. W. Wang, “GAN-GLS: Generative lyric steganography based on generative adversarial networks,” *Computers, Materials & Continua*, vol. 69, no. 1, pp. 1375–1390, 2021.

[19] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec and K. Cho, Augmentation for small object detection. In: *arXiv:1902.07296v1*, pp. 1–15, 2019.

[20] M. Najibi, P. Samangouei, R. Chellappa and L. S. Davis, “SSH: Single stage headless face detector,” in *IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 4875–4884, 2017.

[21] J. Deng, J. Guo, Y. Zhou, J. Yu and S. Zafeiriou, Retinaface: Single-stage dense face localisation in the wild. In: *arXiv,1905.00641v2*, pp. 1–10, 2019.

[22] X. Tang, D. K. Du, Z. Q. He and J. T. Liu, “Pyramidbox: A context-assisted single shot face detector,” in *Proc. of the European Conference on Computer Vision*, Munich, Germany, pp. 797–813, 2018.

[23] L. Zhao and M. Zhao, “Feature-enhanced refinedet: fast detection of small objects,” *Journal of Information Hiding and Privacy Protection*, vol. 3, no. 1, pp. 1–8, 2021.

[24] B. Singh and L. S. Davis, “An analysis of scale invariance in object detection snip,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 3578–3587, 2018.

[25] S. F. Zhang, X. Y. Zhu, Z. Lei, H. L. Shi, X. B. Wang *et al.,* “Faceboxes: A CPU real-time face detector with high accuracy,” in *2017 IEEE Int. Joint Conf. on Biometrics*, Denver, Colorado, pp. 1–9, 2017.

[26] D. Y. Zhang, J. W. Hu, F. Li, X. L. Ding, A. K. Sangaiah *et al.,* “Small object detection via precise region-based fully convolutional networks,” *Computers, Materials & Continua*, vol. 69, no. 2, pp. 1503–1517, 2021.

[27] T. Y. Lin, P. Dollár, R. Girshick and K. M. He, “Feature pyramid networks for object detection,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2117–2125, 2017.

[28] F. Yu, V. Koltun and T. Funkhouser, “Dilated residual networks,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 472–480, 2017.

[29] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Int. Conf. on Learning Representations*, San Diego, CA, USA, pp. 1–14, 2015.

[30] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Reed *et al.,* “Going deeper with convolutions,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 1–9, 2015.

[31] K. M. He, X. Y. Zhang, S. Q. Ren and S. Jian, “Deep residual learning for image recognition,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770–778, 2016.

[32] K. M. He, G. Gkioxari, P. Dollár and R. Girshick, “Mask R-CNN,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2961–2969, 2017.

[33] S. Ren, K. M. He, R. Girshick and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[34] J. Hu, L. Shen and G. Sun, “Squeeze-and-excitation networks,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7132–7141, 2018.

[35] M. Lin, Q. Chen and S. Yan, Network in network. In: *arXiv:1312.4400*, pp. 1–10, 2013.

[36] V. Nair and G. E. Hinton, “Rectifified linear units improve restricted boltzmann machines,” in *Int. Conf. on Machine Learning*, Haifa, Israel, 2010.

[37] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng *et al.,* Detnet: A backbone network for object detection. In: *arXiv:1804.06215*, pp. 1–17, 2018.

[38] A. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim *et al.,* “Automated detection of COVID-19 cases using deep neural networks with X-ray images,” *Computers in Biology and Medicine*, vol. 121, no. 10, pp. 103792–103803, 2020.

[39] E. Hussain, M. Hasan, M. A. Rahman, I. Lee, T. Tamanna *et al.,* “CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images,” *Chaos, Solitons & Fractals*, vol. 142, no. 5, pp. 110495–110507, 2020.

[40] J. P. Cohen, P. Morrison and L. Dao, Covid Chest x-ray Dataset. In: *arXiv 2003.11597*, 2020. [Online]. Available at: https://github.com/ieee8023/covid-chestxray-dataset.