Pattern Recognition 143 (2023) 109801

Contents lists available at ScienceDirect

Pattern Recognition

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

Construction of a feature enhancement network for small object detection☆

Hongyun Zhang[a,*], Miao Li[a], Duoqian Miao[a], Witold Pedrycz[b], Zhaoguo Wang[a], Minghui Jiang[a]

[a] Department of Computer Science and Technology, Tongji University, Shanghai, 201804, PR China
[b] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2G7, Canada

ARTICLE INFO

Article history:
Received 25 November 2022
Revised 27 May 2023
Accepted 2 July 2023
Available online 4 July 2023

Keywords:
Collision detection
Granular computing
High-Resolution block
FENet
HR-FPN
Small object detection,

ABSTRACT

Limited by the size, location, number of samples and other factors of the small object itself, the small object is usually insufficient, which degrades the performance of the small object detection algorithms. To address this issue, we construct a novel Feature Enhancement Network (FENet) to improve the performance of small object detection. Firstly, an improved data augmentation method based on collision detection and spatial context extension (CDCI) is proposed to effectively expand the possibility of small object detection. Then, based on the idea of Granular Computing, a multi-granular deformable convolution network is constructed to acquire the offset feature representation at the different granularity levels. Finally, we design a high-resolution block (HR block) and build High-Resolution Block-based Feature Pyramid by parallel embedding HR block in FPN (HR-FPN) to make full use different granularity and resolution features. By above strategies, FENet can acquire sufficient feature information of small objects. In this paper, we firstly applied the multi-granularity deformable convolution to feature extraction of small objects. Meanwhile, a new feature fusion module is constructed by optimizing feature pyramid to maintain the detailed features and enrich the semantic information of small objects. Experiments show that FENet achieves excellent performance compared with performance of other methods when applied to the publicly available COCO dataset, VisDrone dataset and TinyPerson dataset. The code is available at https://github.com/cowarder/FENet.

1. Inroduction

Object detection technology based on deep learning has been developed rapidly and comes with improved accuracy. Although the accuracy of object detection has considerably improved on various large-scale datasets, there is still a significant discrepancy between the detection of small and large objects. Existing models are relatively inefficient to detect small objects, especially both the localization and classification accuracy of small objects are lower compared to the one when dealing with large ones. However, small object detection is a common problem in numerous application scenarios, such as UAV aerial photography [1], face detection [2], video surveillance [3], and action recognition [4,5] etc. Currently, there are two main definitions of the small object [6]: (i) expressed in terms of pixel size. An object with a resolution lower than 32 × 32 pixels is considered small; (ii) in terms of relative size. An object with both width and height less than 0.1 of the size of the whole image is called small.

In recent years, there have been proposed different solutions to the challenges in detecting small objects. On the one hand, some efforts attempt to extend small object dataset by data augmentation, which is the simplest and effective way to enhance the representability of small objects by the techniques such as oversampling and image processing. One common practice of these methods [7] rotate small objects to improve the diversity of dataset and others [8] copied and pasted existing small objects in original dataset to expand the data. On the other hand, feature representation in the algorithms plays an important role in improving the detection performance. A straight forward idea is to design a different convolution network to learn feature representation of small objects such as the methods of adjusting receptive fields [9] and the dilated or deformable convolution [10]. Besides, the feature representation with different resolution is also the focus of small

☆ This work was supported by the National Natural Science Foundation of China under Grants 62076182, 61976158, 61976160, 62076184 and the Natural Science Foundation of Shanghai 22ZR1466700.

* Corresponding author at No.4800 Cao'an Highway, Jiading District Shanghai, Room 476, Telecom Building
E-mail address: zhanghongyun@tongji.edu.cn (H. Zhang).

https://doi.org/10.1016/j.patcog.2023.109801
0031-3203/

object detection. There are some studies to obtain enhanced feature representations from existing neural networks models (such as VGG-Net [11], Recurrent Neural Networks [12]) by using high-resolution features fused with high-dimensional features of low-resolution images. In addition, Feature pyramid network [13] is the most popular method which can make full use of different resolution features. However, there are still some shortcomings in the exiting methods. Data augmentation methods that have been proposed based on random copy-paste algorithm can easy lead to the occlusions between objects and incorrect context information. The feature extraction methods combined FPN [13] with single granularity CNN have good performance in normal-scale object detection, but the results for small object are not satisfactory.

In our study, a new small object detection method called FENet is proposed, whose main objective is to study, implement and improve the small object detection task by performing data augmentation and feature representation enhancement to construct a more robust small object detector. First of all, collision detection and spatial context extension are introduced to improve current data augmentation methods based on copy-paste. Collision detection and spatial context extension solve the problems of object collision and incorrect context information caused by random copy-paste. Then, aiming at the problem that the small objects are vulnerable to scale variation, combined with the ideas of Granular Computing, a multi-granular deformable convolution network is constructed to acquire offset feature representation in different granularity levels by granulating and fusing the offset features, which allows the model not only to learn the changes in the shape of the object, but also to capture the changes in the scale of the object. Finally, to deal with the low resolution of small objects, we design a high-resolution block (HR block) that can bring more semantics and detailed information by maintaining high resolution features through the whole process and fusing the feature of different resolution. To fully utilize different levels of granularity and resolution features, we build High-Resolution Block-based Feature Pyramid by parallel embedding HR block in FPN. The application of above strategies can reduce the loss of small object information and acquire more sufficient feature information. The main contributions of this paper are as follows:

- To more effectively expand the possibility of small objects appearing, we improve current copy-paste based data augmentation method by introducing collision detection and spatial context position extension to avoid object collision and incorrect context information caused by random copy-paste.
- To solve the problem that the small objects are vulnerable to scale variation, we construct a multi-granular deformable convolution network to learn and capture the changes in the shape and scale of the object, and offset feature representations in different granularity are acquire by granulating and fusing the offset features.
- A high-resolution block (HR-block) is designed to bring more semantics while maintaining high-resolution features, and high-resolution block-based Feature Pyramid is built by parallel embedding HR block in FPN to further enhancing the feature representation.
- A large number of experiments are reported to demonstrate the effectiveness of the proposed method. At the same time, we set up ablation experiments to analyze the rationality of proposed different strategies.

In summary, a novel small object detector is proposed, which uses improved data augmentation, multi-granularity deformable convolution and optimized feature pyramid with designed high-resolution blocks to obtain richer semantic information and more precise detailed features of small objects.

Table 1
MS COCO 2017 Labeling object information.

Object	Proportion of objects	Area	Proportion
Small objects	31.13%	0.58%	43.54%
Medium objects	34.90%	5.99%	64.72%
Large objects	33.97%	93.44%	91.22 %

The study is organized as follows. Section 2 introduces the fundamentals of related research. Proposed method is described in Section 3. Section 4 reports on the experiments compared with other methods and the ablation experiments. Parameter analysis are given in Section 5.

2. Related works

2.1. Data augmentation methods

There are many data augmentation methods in object detection, such as random cropping, random flipping, Gaussian noise, etc. These augmentation methods are inherently more general and well suited for vision tasks by encoding the invariance of data transformations.

Since small objects are intrinsically small in size, the corresponding area they cover is also relatively small, which means that the locations of small objects lack diversity. By summarizing various information in the MS COCO2017 dataset in Table 1, we can see that although there is small difference in the number of large and small objects, the area occupied by small objects is only 0.58% of the total area for all images, which is almost negligible compared to 93.44% of are occupied by large objects (resolution higher than 96×96).

From the aspect of the number of included samples, the number of images containing small objects is only 43.54%, much lower than that of medium-sized (64.72%) and large objects (91.22%). Table 1 points at significant differences in the number and area proportion of small objects compared to medium and large objects.

It is crucial to enhance the diversity of small objects in the dataset through data augmentation. A simple approach of copying instances of objects from one image and pasting them onto another is called the copy-paste algorithm. Deng *et al.* [14] consider the spatial context information of the objective image and the current instance when pasting, making the final generated image more reasonable. Fang et al. [15] extracted the instances in the images and blends them into different contextual images to train on the enhanced images outside the original dataset. Kisantal et al. [16] firstly oversample the small object images and then paste them in the image multiple times at any position. Lee and Bae [17] build a GAN-based object generator to improve feature interpolation of feature pyramid networks.

2.2. Feature extraction methods for small object

Feature extraction is an import part in small object detection. There are many implementation methods. One of them is to use the transformable convolution kernels to obtain different receptive filed. Dai et al. [10] first used deformable convolution for object detection to adjust receptive fields by learning the offset information of the convolution kernel. Yang et al. [18] introduced the form of point sets to represent the objects by using the deformable convolution network.

On the other hand, the multi-scale learning based on different resolutions can also produce the feature information. Liu *et al.* built a single shot multibox detector (SSD) [9] that is able to learn features of different resolutions obtained by normal convolution neural network to detect small objects. Different from

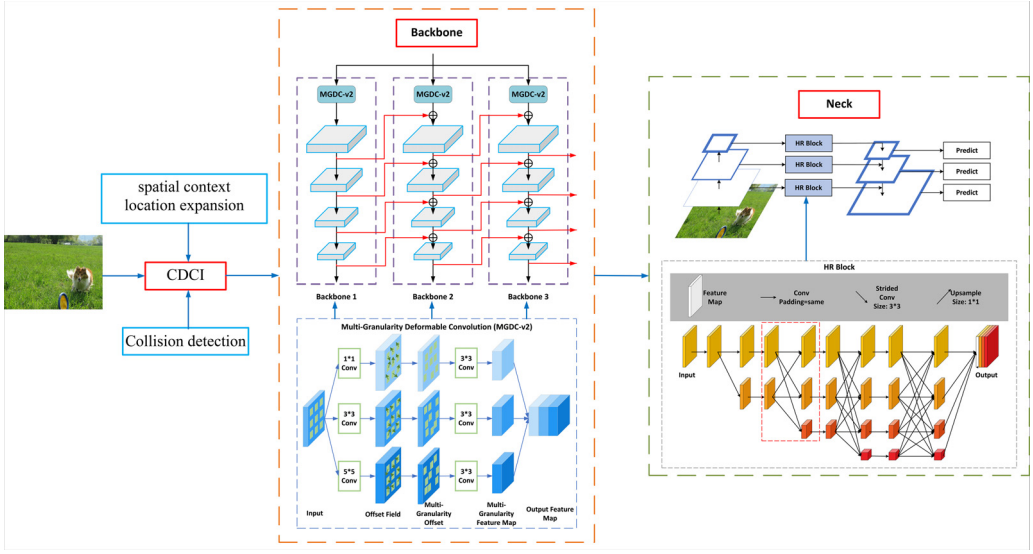


Fig. 1. The structure of proposed model.

SSD, feature pyramid network (FPN) by Lin et al. [13] established a top-down architecture with lateral connection to extract high-level feature maps at different scales. The feature representations of different resolution are obtained by FPN to detect objects. Xia et al. [19] *et al.* proposed a TRAN-based self-supervised model and designed a deformable attention converter to achieve target recognition. In recent years, Sun et al. [20] provides a new structure (HRNetV1), which maintains the high-resolution represent of the feature in the whole model. Based on HRNetV1, Wang et al. [21] extended the model into four-resolution representations and connected the high-to-low resolution convolution stream in parallel to exchange the information of the features in different resolutions.

However, most data augmentation methods based on copy-paste are random, which causes objects to collide with each other. In addition, unreasonable pasting position can result in incorrect context information. The receptive field of deformable convolution is not sufficient due to the single level of granularity. Meanwhile, the feature pyramid network cannot make full use of the feature information under high resolution. To solve these problems, A Feature Enhancement Network (FENet) is proposed in our study.

3. Proposed method

In this section, we construct the Feature Enhancement Network (FENet) and propose a small object detection algorithm to improve its performance. FENet mainly consists of three parts: data processing part, backbone part and neck part. Fig. 1 shows the overall structure of the proposed algorithm.

In Fig. 1, the first part on the left side of the figure is data augmentation method named CDCI, aiming to expand small objects while preserving them context. The second part composed of triple multi-granularity deformable convolution networks is a backbone, whose purpose is to extract the features of small object better. The last part is a neck combining FPN with high-resolution block (HR Block), which is used to fuse features obtained in different granularities and resolutions.In what follows, Section 3.1 introduces the CDCI method, Section 3.2 introduces the middle part of Fig. 1, which called backbone, and Section 3.3 introduces the neck module based on FPN.

3.1. Data augmentation method based on collision detection and spatial context location expansion (CDCI)

3.1.1. Problems with the random copy-paste method

To solve the problems of the limitation of the location of small objects, most of the previous methods use the strategy of random copy-paste objects. However, there are usually the following two problems.

Object collision. The positions of the pasted small objects are entirely random, and this may lead to overlapping objects. This means that the covered objects cannot correspond to the annotations, which reduces the performance of the detector.

Object location irrationality. The increase of position diversity means that the unreasonableness of position also increases. For a pasted position, the pasted object should probably not be present at that position at all, which destroys the semantic information in the image and causes the detector to get the wrong information about location of small objects.

3.1.2. CDCI data augmentation algorithm

Because the algorithm [8] was the state-of-the-art and strong copy-paste method for data augmentation, we choose it as the baseline data augmentation method. The CDCI consists of two main parts.

Object collision detection. The collision detection means determining whether the pasted object will overlap with original objects in the image. The proposed method utilizes Intersection over Union (IOU) as a measure of the degree of collision.

To speed up processing, the objects are directly judged to be in non-collision state when the value of IOU is zero. Assuming that $box_{small} = \left(x_1^s, y_1^s, x_2^s, y_2^s\right)$ represents the bounding box of the small object to be pasted, where $\left(x_1^s, y_1^s\right)$ represents the coordinates of the top-left vertex of the bounding box and $\left(x_2^s, y_2^s\right)$ shows coordinates of the bottom-right vertex of the bounding box. $box_{other} = \left(x_1^o, y_1^o, x_2^o, y_2^o\right)$ represents the bounding box of the object to be compared for the existence of collision. According to Eq. 1, the top-left vertex $\left(x_1^i, y_1^i\right)$ and the bottom-right vertex $\left(x_2^i, y_2^i\right)$ of the intersecting region can be determined.

$$\begin{cases} x_1^i = max\left(x_1^s, x_1^o\right), \\ y_1^i = max\left(y_1^s, y_1^o\right), \\ x_2^i = min\left(x_2^s, x_2^o\right), \\ y_2^i = min\left(y_2^s, y_2^o\right) \end{cases} \tag{1}$$

Fig. 2. A small copied object with context information.

The formula to determine if there is a collision between two bounding boxes is expressed as:

$$have\_collision = \left(x_1^i < x_2^i \;\&\&\; y_1^i < y_2^i\right), \tag{2}$$

if *have collision* is True, a collision occurs and one needs to find a new paste position, otherwise, paste the object directly.

Location extension based on spatial context. Unlike the previous method of only pasting small objects, we paste the information around the small objects to the new location together to preserve the scene information to a certain extent. This method not only reduces the possibility of unreasonable positions of objects after pasting, but also preserves the spatial context information of small objects as much as possible. When performing the copy-paste operation, we add equally spaced padding around the bounding box to expand the pixel value of the scene around object. Suppose that $box_{small} = \left(x_1^s, y_1^s, x_2^s, y_2^s\right)$ is the original bounding box of an object, and after adding padding, the copied area becomes $box_{copied} = \left(x_1^s - pd, y_1^s - pd, x_2^s + pd, y_2^s + pd\right)$, where pd is the pixel value expanded in all directions. As shown in Fig. 2, the red box is the genuine bounding box. We extend the bounding box of the object to a larger scene, and then the extended box area is pasted to the new area.

Fig. 3 shows the results of adding contextual information to small objects in different images, where the red boxes represent the real bounding boxes and the brown boxes indicate the extended areas after adding padding. In Fig. 3, we can see that pasted targets contain more effective contextual information by adding padding, such as the grass in Fig. 3(a), the human face in Fig. 3(b), the sky and the ocean in Fig. 3(c), and the light line in Fig. 3(d). The introduction of these contextual information enhanced the information of the objects in the original bounding box and was more beneficial to the understanding of the objects.

3.1.3. CDCI data augmentation algorithm flow

The CDCI is implemented based on the current mainstream copy-paste algorithm. Assuming that it needs to generate N enhanced images from the original dataset SI, two images IA and IB are first randomly selected from this dataset. The datasets of instances in images IA and IB are represented by SIA and SIB, respectively, and the two images are randomly scaled and dithered. Compared with the traditional random copy-paste algorithm, the algorithm [8] uses a larger dithering scale (the range extends from [0.8,1.24] to [0.1, 2.0]) while performing a random horizontal flipping operation on each of these two images, and then randomly selecting a subset ST from the set of instances SIA or SIB of one image. To retain more spatial context information of the objects, CDCI adds a padding of the corresponding size when extracting the instances and pastes this subset of instances ST into another image, and collision detection is performed at the same time.

Table 2
Algorithm 1. CDCI data augmentation (pseudo code).

Input:	original dataset SI=I1, I2, I3,...,IM, the number of images in the target dataset N, the empty set used to store the enhanced images SAUG=
Output:	enhanced dataset SAUG=IA1, IA2, IA3,..., IAN
1.	For i :=1:N
2.	Randomly select images IA and IB from SI
3.	Do random scaling dithering on IA and IB
4.	Do a random horizontal flip on IA and IB
5.	Extract instances from IA and IB and extend the spatial context location of the instances to form the set SIA, SIB
6.	Extract a random subset of ST from SIA or SIB
7.	Paste ST into another image, then do collision detection when pasting, update annotation to generate enhanced image IAi
8.	SAUG←AUG+IAi
9.	END FOR
10.	RETURN SAUG

Finally, the annotation of the pasted instance is updated to generate the enhanced image IAUG. then we repeat the above process N times, and form the SAUG which contains a collection of N data-enhanced images. The specific improved algorithm is shown in Table 2.

3.2. Deformable convolution network based on multi-granularity

In traditional convolutional neural networks (CNN), the convolution kernels are often the same size (e.g., 3×3). For small objects, due to their inherently small scale, a slight scale change can have drastic changes in feature extraction. Therefore, we introduce the deformable convolution instead of traditional CNN to obtain the representation of offset under different granularity.

3.2.1. Offset feature learning of the sampling locations

As shown in Fig. 4 and Fig. 5, the traditional convolution can only focus on a fixed square region, and deformable convolution [10] can make the network more adaptable to the changes of the sampling locations by learning the offset of each location of the convolution kernel.

The single deformable convolution is shown in Fig. 6. The input feature is denoted as $f_{in} = [C, H, W]$, where C is the number of channels of the input feature map, H and W are the height and width of the input feature map, respectively.

In Fig. 6, the proposed method can obtain the offset feature map $f_{offset\_field} = [2C, H, W]$ of the input feature map, and the number of channels is changed to twice as much as the original, it represents the offset of each pixel in two directions (horizontal

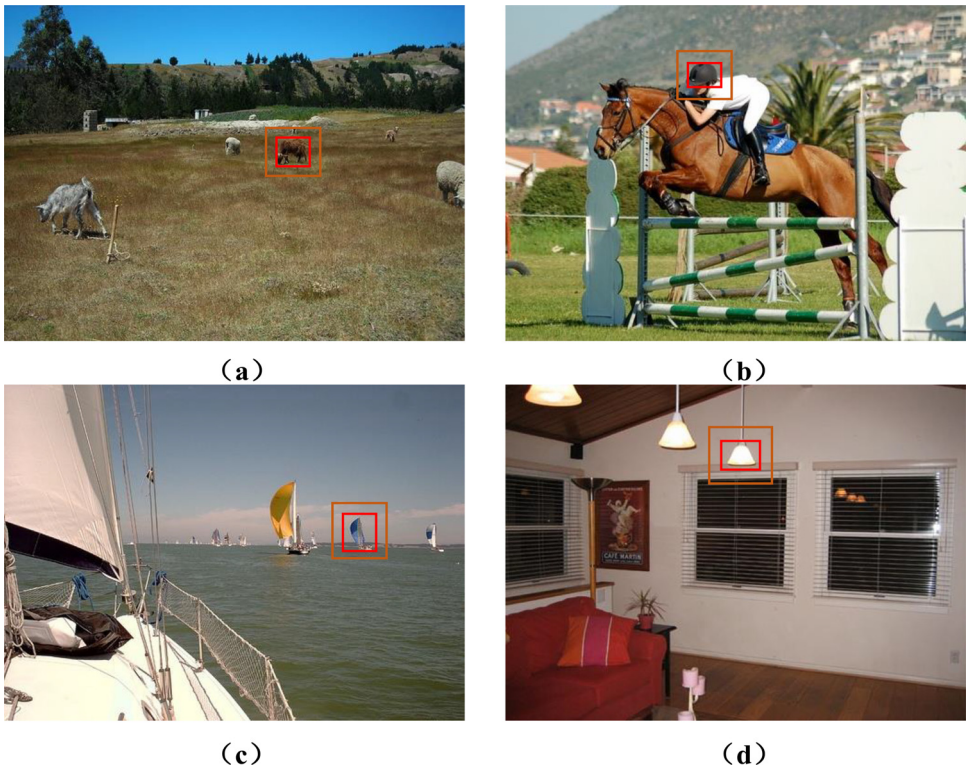


Fig. 3. Adding context information to small objects in different scenarios.

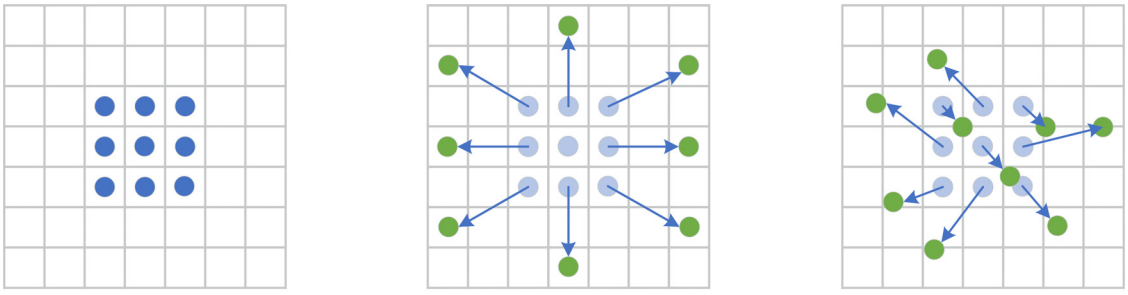


Fig. 4. Various deformable convolution kernels.

direction X and vertical direction Y). By adding the index of the position (in f_{in}) and the corresponding offset in the x and y directions (in f_{offset_field}), we obtain the new index for each pixel (it is necessary to ensure that the index is within the range of feature map). However, the predicted offset is a real number, but the value of new index is not necessarily an integer. For the index value $P = (x, y)$ of the new index, we get the four integers nearest to P: $P_{11} = (x_1, y_1)$, $P_{12} = (x_1, y_2)$, $P_{21} = (x_2, y_1)$, and $P_{22} = (x_2, y_2)$, where $x_1 = floor(x)$, $x_2 = ceil(x)$, $y_1 = floor(x)$, and $y_2 = ceil(x)$.

Fig. 7 shows the process of bilinear interpolation. We obtain the eigenvalues of R_1 and R_2 by interpolating among the x-coordinate.

$$f(R_1) = \frac{x_2 - x}{x_2 - x_1} f(P_{11}) + \frac{x - x_1}{x_2 - x_1} f(P_{21}), \tag{3}$$

$$f(R_2) = \frac{x_2 - x}{x_2 - x_1} f(P_{12}) + \frac{x - x_1}{x_2 - x_1} f(P_{22}). \tag{4}$$

Then we obtain the eigenvalues of f(P) by interpolating among the y-coordinate.

$$f(P) = \frac{y_2 - y}{y_2 - y_1} f(R_1) + \frac{y - y_1}{y_2 - y_1} f(R_2). \tag{5}$$

After calculating the eigenvalues of all new index points, we get a new feature map $f_{offset} = [C, H, W]$, each eigenvalue in f_{offset} is the value after the offset. Finally, we get the final output feature map $f_{out} = [C_{out}, H_{out}, W_{out}]$, where C_{out}, H_{out} and W_{out} are the number of channels, height and width of the output feature respectively.

3.2.2. Multi-granularity deformable convolution design

Granular Computing is a new computing paradigm in the field of artificial intelligence to simulate human way of thinking and solve complex problems by reducing the problem to some smaller problems occurring at different levels of abstraction (granularity). The most important step in the multi-granularity theory is to determine the granulation object of the problem.

In our research, the granulation object is the offset feature map. Usually, a 3×3 convolutional kernel can be employed to learn the offsets of each point on the feature map in the Deformable convolution [10]. We granulate the offsets learning into three different levels of abstraction to gain the representation of offsets feature in multi-granularity. Fig. 8 shows the multi-granularity offsets.

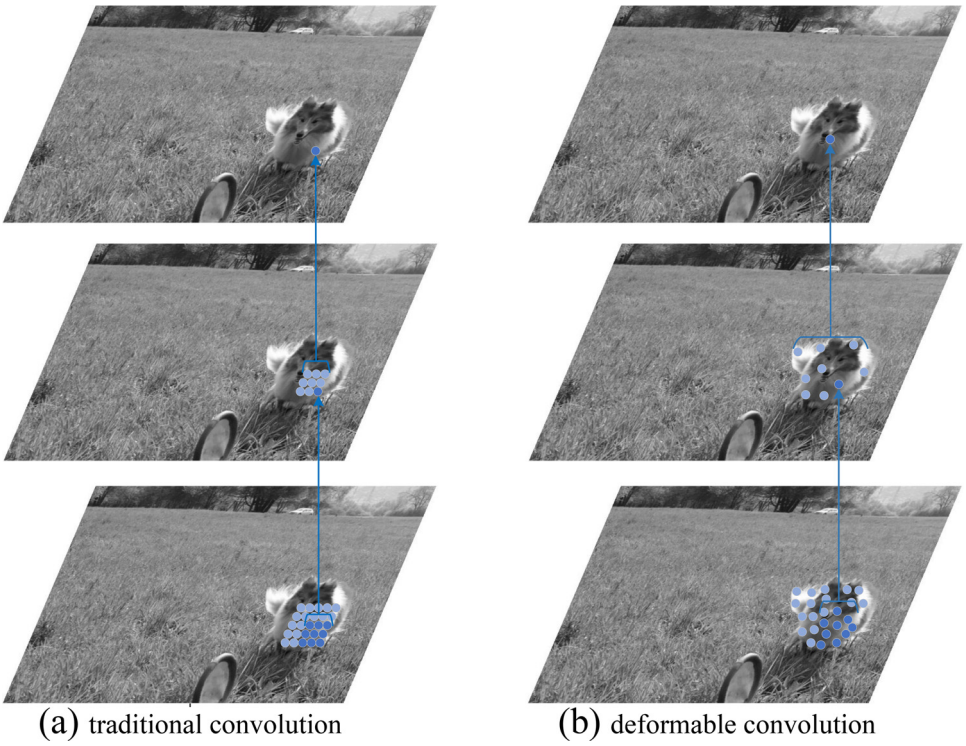


Fig. 5. Comparison of traditional and deformable convolution.

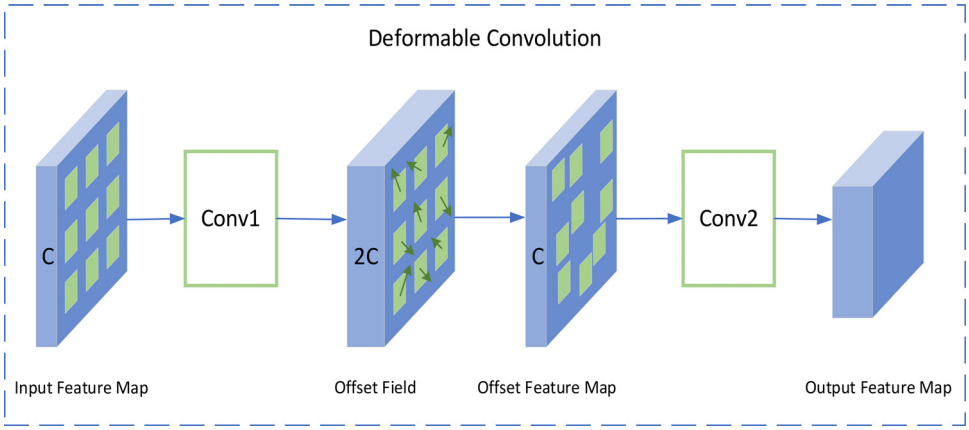


Fig. 6. Single deformable convolution.

Low-rank granule learning: The offset field can be learned from the input feature map by 1×1 convolution. Under this granularity, we learn the offset information of each sampling point separately, without their surrounding information.

Middle-rank granule learning: 3×3 convolution is utilized for offset field learning. Here, offsets of each point are learned using a corresponding 3×3 grid area in the input feature map, which has a higher spatial perception range than the low-rank learning.

High-rank granule learning: 5×5 convolution is utilized for offset field learning. Offsets of each point are learned from a corresponding 5×5 grid, which has the largest receptive field compared with low-rank and middle-rank counterpart.

After obtaining multi-granularity representation of location offsets, these feature maps coming at different level of granularity need to be fused. In this way, we can get an integrated representation from input feature map.

Generally, there are two ways of fusing feature maps. One is element-wise, which adds or multiplies the corresponding locations of different feature maps in different levels, and the number of channels are kept the same after fusing. The second is channel-wise, meaning that the obtained multi-level feature map is concatenated along the channel dimension to form a new feature map with two times the number of channels. However, it is difficult for element-wise to determine the offsets in fusion under different granularities. Correspondingly, the channel-wise method can completely preserve the offsets of all granularity levels.

Fig. 9 shows the detailed process of multi-granularity deformable convolution (MGDC). Input feature map can be represented as $f_{in} = [C, H, W]$, and then the maps are fed into three branches to get the offset information from three different granularity, and the value of the offset is expressed as $f^{i}_{offset\_field} = [2C, H, W]$ $(i = 1, 2, 3)$. Subsequently, the multi-granularity offset can be generated by using bilinear interpolation, and then we use 3×3 convolution to extract features to obtain multi-granularity feature maps. Lastly, the fused feature map is obtained by concatenating these three levels of feature maps.

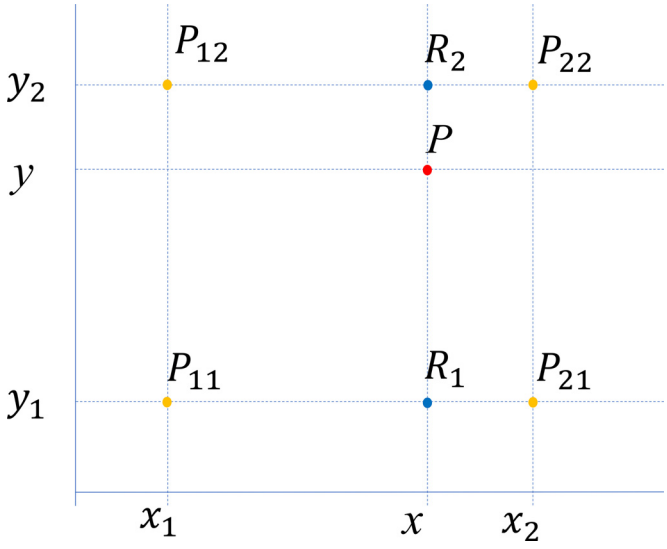


Fig. 7. Bilinear interpolation method.

3.3. High resolution block-based feature pyramid

Now, we build High Resolution Block (HR Block) and combine it with FPN to propose the HR-FPN with richer information representation.

3.3.1. High resolution block design

To maintain high resolution, inspired by HRNet [21], the proposed HR-Block contains multiple parallel features maps of different resolutions, whose combination makes up the final output feature maps. The HR Block in the Fig. 1 shows the structure of a single HR Block.

High-resolution features contain a lot of detail, and low-resolution features have richer semantic information. Based on these advantages, every layer of HR-Block uses high-resolution to maintain the feature information of small object, at the same time, HR-Block also combine low-resolution features to better extract the information of small objects.

In the HR Block of Fig. 1, the cube whose size indicates the difference of resolution represents the feature map, and the horizontal arrow represents the convolution. The down arrow means that 3×3 convolution with stride of 2, 3, or 4 (gradually increasing according to the number of vertically parallel feature maps). After these convolutions, the resolution of feature map is reduced, but semantic information will be richer. Considering the loss of information in the dimensionality reduction, we use the convolution with 3×3 stride instead of pooling layers to minimize the loss of information. The arrow pointing up represents bilinear interpolation. We use the 1×1 convolution to change the number of channels for the convenience of subsequent feature fusion.

Fig. 10 shows the part framed by a dashed line in HR Block, which is located at the lower-right corner of Fig. 1. For feature maps A and B, different colors indicate convolution with different parameters in Fig. 10. Next, element-wise addition is utilized to fuse the feature maps with same resolution in the same vertical level.

3.3.2. Embedding high resolution block in FPN

The reasons why the FPN structure can promote small object detection are two-fold. First, FPN increases the resolution of the feature of the small object, and can retain more effective information of the object. Second, FPN captures more contextual information to small objects, due to the existence of top-down path, the

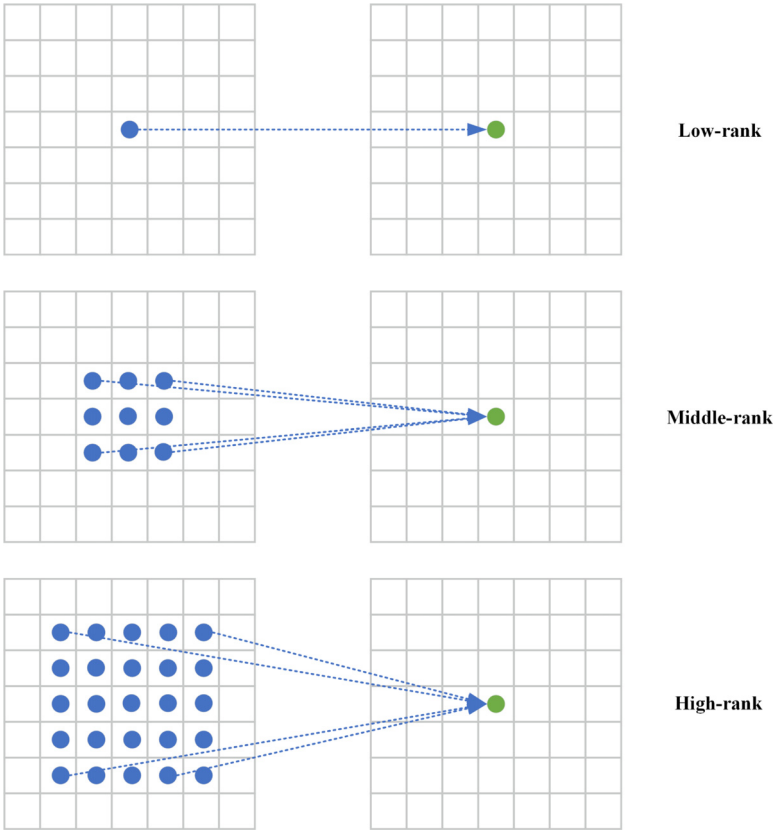


Fig. 8. A multi-granularity representation of offset.

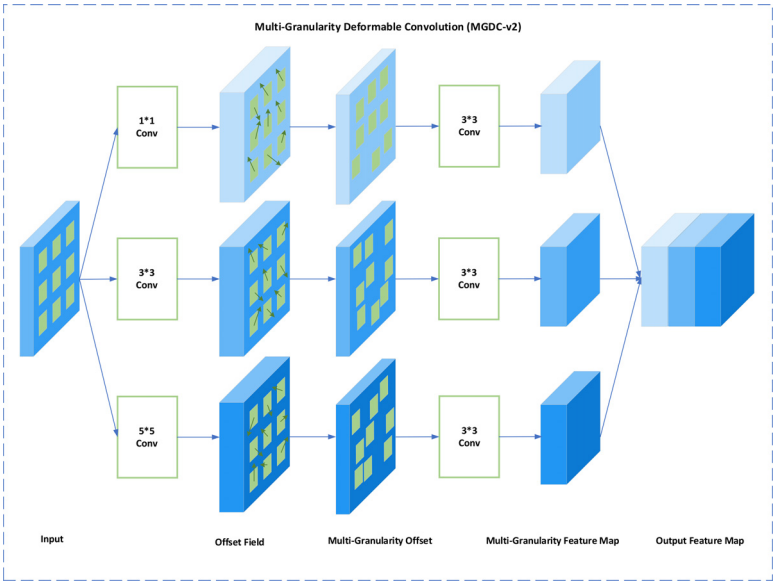


Fig. 9. Multi-granularity deformable convolution.

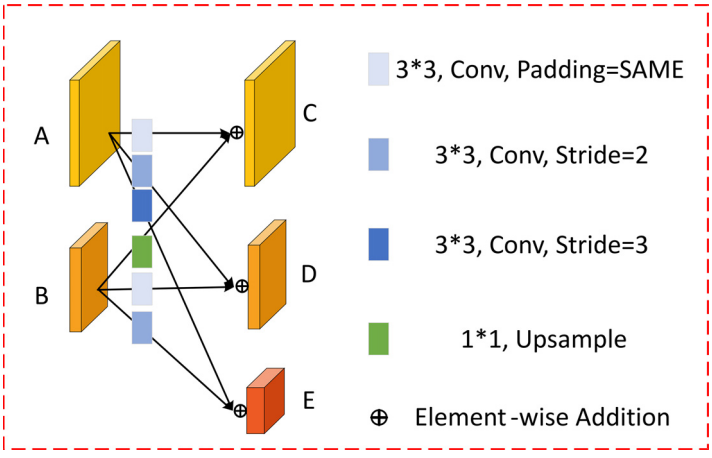


Fig. 10. The feature fusion of HR-Block.

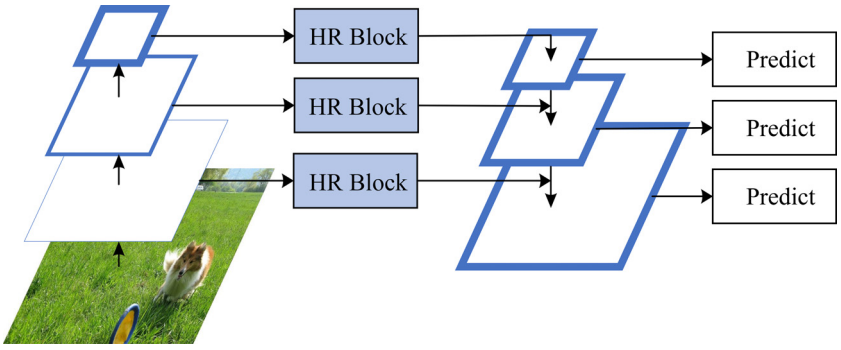


Fig. 11. The feature fusion of HR-Block.

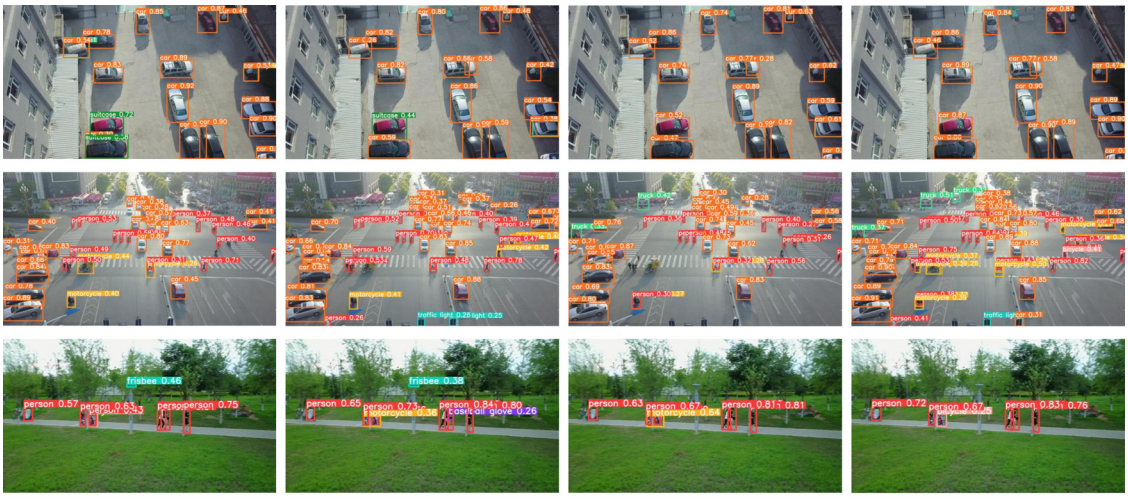
semantic information at the top level can be transmitted to the bottom level.

We introduced the HR block in FPN to improve the ability of small object detection. As shown in Fig. 11, we use multi-layer downsampling to obtain feature maps at different resolutions, these feature maps are sent to HR blocks and the output of each HR block is fused along the top-down path. Specifically, the top-level feature map is upsampled to obtain a new high-resolution feature map. Finally, the subsequent classification and regression tasks are performed on the fused feature maps.

4. Experimental studies

4.1. Experiment setup

We adopted Cascade R-CNN [22] as the baseline with Triple-ResNext152 for backbone. The experimental results are all obtained



(a) RetinaNet (b) ATSS (c) YOLOv5 (d) FENet

Fig. 12. Comparison of subjective performance.

on the MS COCO 2017 dataset [6]. We used the Adam optimizer with an initial learning rate of 1e-3. We used the computer equipment with 128G RAM, Tesla V100GPU × 8, and Intel(R) Xeon(R) Silver 4114 CPU.

4.2. Comparison of subjective performance

To fully represent the detection capability of the proposed model, we compare the FENet proposed in this paper with RetinaNet [23], ATSS, and YOLOv5. We specifically select the samples with relatively small objects present in the images. Fig. 12 shows the detection results: FENet has better performance for small object detection in the practical detection results, and both the missed and false objects are less than the other three comparison models.

For the number of boxes selected in the detection results, there is little difference between YOLOv5 and FENet, but the degree of confidence of the objects boxed by YOLOv5 is generally lower than FENet. This indicates that it has insufficient discriminatively ability for the attributes of the objects. Meanwhile, the second image in Fig. 12(d) also shows that there are missed objects in the FENet in the case of more pedestrians and covered persons. Currently, the problem of missed and false detection is also one of the difficulties in object detection, and the method proposed in this paper still has some limitations in solving the occlusion problem.

4.3. Comparison of objective performance

4.3.1. Algorithm performance comparison on small object detection datasets

The VisDrone detection dataset [24] consists of UAV vision images designed for small object detection, in which the objects are small in size and large in number, VisDrone can effectively verify the performance of the network model for small object detection. Although the TinyPerson dataset [25] has only human annotation information, the object size is relatively smaller and the resolution is lower, and it is more difficult to detect. We tested the FENet with other models on the VisDrone and the TinyPerson dataset. Table 3 shows the detection results on the VisDrone dataset, and Table 4 shows the results on the TinyPerson dataset. In Table 3, the detection accuracy of FENet is only slightly lower than that of DroneEye2020 and TAUN, while Table 4 shows that FENet is only marginally lower than the method in the literature [25] on the TinyPerson dataset. The experimental results further demonstrate the effectiveness of FENet on the small object detection.

4.3.2. Comparison with mainstream methods

For a fair comparison with other methods, MS COCO test-dev 2017 is also selected as the test dataset, which is the most authoritative and challenging dataset in object detection. Table 5 shows the experimental results. We divided the various algorithms into two categories: one-stage and two-stage methods, the method with * indicate a multi-scale approach is used. This paper directly adopts the source code and pre-trained models provided by the authors, and we only use the multi-scale testing method in the testing phase to ensure the fairness of the experiments.

In Table 5, FENet is relatively advanced among the mainstream object detection models both in overall detection accuracy and small object detection accuracy. By adopting a multi-granularity deformable convolution and high-resolution feature pyramid network, the overall detection accuracy and small-object detection accuracy of baseline improved to 54.6 and 37.2. With the addition of the pre-processed data augmentation method CDCI, the overall detection accuracy of FENet improves from 54.6 to 55.5, and the small object detection accuracy improves from 37.2 to 38.0. It illustrates that the CDCI algorithm proposed in this paper has improved both the overall detection accuracy and the effectiveness for small object detection. In addition, FENet reaches the optimal accuracy in the series of Cascade Mask RCNN-based detection algorithms. Although the CenterNet2 methods in Table 5 is superior than FENet in accuracy, CenterNet2 uses higher resolution input images in the testing phase.

4.3.3. Model efficiency

Here we use Cascade Mask R-CNN [22] as a baseline model. We propose the FENet (Feature Enhancement Networks) by applying CDCI (Collision Detection and Contextual Information based augmentation), MGDC (Multi-Granular Deformable Convolution) and HR-FPN (High Resolution block-based Feature Pyramid Networks) to the baseline. The FPS (Frame Per Second) metric is utilized to evaluate model efficiency. Under the proposed experimental conditions, the baseline model runs at 23 FPS, and the proposed FENet runs at 15 FPS. The inference efficiency of FENet is 30% lower than the baseline.

Table 3
Performance comparison of the proposed method on dataset VisDrone.

Methods	Model	AP	AP50	AP75	AR1	AR10	AR100	AR500
DroneEye2020	Cascade RCNN	**34.57**	58.21	**35.74**	0.28	1.92	6.93	52.37
TAUN	ATSS [26]	34.54	59.42	34.97	0.14	0.72	12.81	49.80
CDNet	Cascade RCNN	34.19	57.52	35.13	0.80	8.12	39.39	52.62
CascadeAdapt	Cascade RCNN	34.16	58.42	34.50	0.84	8.17	39.96	47.86
HR-Cascade+	Cascade RCNN	32.47	55.06	33.34	0.94	7.81	37.93	50.65
MSC-CenterNet	CenterNet [27]	31.13	54.13	31.41	0.27	1.85	6.12	50.48
Proposed	Cascade RCNN	34.50	**59.73**	32.31	0.57	1.20	35.70	51.55

Table 4
Performance comparison of the proposed method on dataset TinyPerson.

Detector	AP_{50}^{tiny}	AP_{50}^{tiny1}	AP_{50}^{tiny2}	AP_{50}^{tiny3}	AP_{50}^{small}	AP_{25}^{tiny}	AP_{75}^{tiny}
FCOS [28]	17.90	2.88	12.95	31.15	40.54	41.95	1.50
RetinaNet [23]	33.53	12.24	38.79	47.38	48.26	61.51	2.28
FreeAnchor [29]	44.26	25.99	49.37	55.34	60.28	67.06	4.35
Libra RCNN [30]	44.68	27.08	49.27	55.21	62.65	64.77	6.26
Grid RCNN	47.14	30.65	52.21	57.21	62.48	68.89	6.38
Faster RCNN-FPN	47.35	30.25	51.58	58.95	63.18	68.43	5.83
RCNN-FPNMSM	50.89	33.79	55.55	61.29	65.76	71.28	6.66
RCNN-FPNMSM+	**52.61**	34.20	**57.60**	**63.61**	**67.37**	72.54	**6.72**
Proposed	51.33	**37.02**	55.03	62.44	66.92	**72.81**	6.20

Table 5
Quantitative assessment comparison of different methods applied to COCO test-dev2017 dataset.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
One-stage methods:							
ATSS [26]	ResNetXt10+DCN*	50.7	68.9	56.3	33.2	52.9	62.4
PAA [31]	ResNext152+DCN*	53.5	71.6	59.1	36.0	56.3	66.9
EfficientDet+DI [32]	Efficient-B7	53.6	71.8	57.0	32.2	51.3	56.8
EFPN [33]	ResNeXt-101	44.6	64.7	49.4	28.0	47.5	54.2
M2YOLOF [34]	ResNet-101	42.6	62.3	46.2	24.3	47.3	57.8
OTA [35]	ResNext-101-DCN	51.5	68.6	57.1	34.1	53.7	64.1
AFI-GAN [17]	ResNext-50	43.8	61.7	47.6	26.9	46.6	53.4
AugFCOS [36]	ResNext-152-DCN	53.5	71.6	59.1	36.0	56.3	66.9
Object [37]	ResNet-50-FPN	42.3	60.3	46.0	24.9	46.0	55.9
UniverseNet [38]	ResNext-101	54.1	71.6	59.9	35.8	57.2	67.4
DAT [19]	Transformer	47.9	69.6	51.2	32.3	51.8	63.4
Two-stage methods:							
Faster RCNN [13]	ResNet101+FPN	36.2	59.1	39.0	18.2	39.0	48.2
Sparse RCNN [39]	ResNeXt101-DCN	43.5	62.1	47.2	26.1	46.3	59.7
TridentNet [40]	ResNet101-Deformable	48.4	69.7	53.5	31.8	51.3	60.3
Cascade Mask RCNN [22]	Triple-ResNeXt152*	53.3	71.9	58.5	35.5	55.8	66.7
CenterNet2 [41]	Res2Net-101-DCN	56.4	74.0	61.6	38.7	59.7	68.6
Proposed method:							
FENet (without pre-processing)	ResNet101 +MGDC-v2*	52.0	71.0	58.1	35.6	56.3	66.2
FENet (with pre-processing)	ResNet101 +MGDC-v2*	53.4	72.0	58.8	36.8	57.0	65.8
FENet (without pre-processing)	Triple-ResNext152 +MGDC-v2*	54.6	73.3	60.2	37.2	57.1	67.3
FENet (with pre-processing)	Triple-ResNext152 +MGDC-v2*	55.5	73.4	60.7	38.0	58.6	67.9

5. Ablation studies and analysis

5.1. Evaluation on data augmentation strategies

The experiments explore the effect of different padding sizes in the expansion of spatial contextual information and compare the proposed data augmentation algorithm CDCI with two conventional copy-paste algorithms to verify the effectiveness of the improved method in this paper.

Table 6
Impact of different padding size on performance.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
no padding	53.4	71.2	58.0	35.3	55.7	66.2
padding=1	53.4	71.7	58.2	35.6	55.5	66.6
padding=2	53.6	72.1	58.1	35.8	**55.9**	66.5
padding=3	**53.8**	**72.4**	**58.4**	**36.0**	**55.9**	66.8
padding=4	53.7	72.3	58.1	35.7	55.7	**66.9**

5.1.1. Effect of different padding size

Table 6 shows the impact of adding padding to small objects on the model performance, and the collision detection is used in all models. When the padding is too small, the performance degrades because small edges affect the characteristics of the object and add some redundant information to the feature of small objects. When the padding becomes larger, the surrounding context information can improve the performance. When the padding size continues to increase, the detection performance for small objects also decreases slightly.

Table 7
Performance comparison of different copy-paste methods.

Methods	base-method	**AP**	**AP$_{50}$**	**AP$_{75}$**	**AP$_S$**	**AP$_M$**	**AP$_L$**
base*	-	52.4	71.2	57.4	34.6	55.3	65.7
w/o Col	base*+cp2	52.8	71.6	57.7	34.9	55.8	66.0
w Col	base*+cp2	52.6	71.8	57.6	35.0	**55.9**	65.9
w Col	base*+cp2+padding	**53.1**	**71.9**	**58.0**	**35.2**	55.6	**66.2**
w/o Col	base*+cp1	53.5	72.0	58.1	35.5	55.4	**67.0**
w Col	base*+cp1	53.4	71.2	58.0	35.3	55.7	66.2
w Col	base*+cp1+padding	**53.8**	**72.4**	**58.4**	**36.0**	**55.9**	66.8

Table 8
Effect of granulation mode on the performance of multi-granularity deformable convolution..

Methods	**AP**	**AP$_{50}$**	**AP$_{75}$**	**AP$_S$**	**AP$_M$**	**AP$_L$**
baseline [22]	53.3	71.9	58.5	35.5	55.8	66.7
baseline*	52.4	71.2	57.4	34.6	55.3	65.7
DCN	53.2	72.0	58.2	35.5	56.0	66.6
MGDC-v1	53.7	72.4	58.7	36.3	56.5	67.2
MGDC-v2	54.0	72.8	**59.0**	**36.6**	**56.7**	67.7
MGDC-v3	54.2	73.0	**59.0**	36.4	56.6	67.9
MGDC-v4	**54.3**	**73.1**	58.8	36.1	56.5	**68.1**
MGDC-v5	54.0	72.7	58.5	35.8	56.3	**68.1**

5.1.2. Performance comparison of data augmentation algorithm

This paper proposes two improved methods for copy-paste. To verify the effectiveness of these methods, we apply them to two mainstream copy-paste algorithms and compare them. Table 7 shows the experimental results, where cp1 is proposed in ref. [8], cp2 is proposed in ref. [16], *Col* represents collision detection, padding represents Spatial context extension. For small objects, it is not easy for them to collide with each other during data augmentation, so the effect improvement brought by single collision detection is small or even performance degradation. Compared to cp2, cp1 randomly resize and rotate images to increase the number of targets (especially small object because of the reduction of the size). Therefore, in Table 6, using cp2 (without padding) can bring some performance improvement (APm: 55.9 vs 55.8 and APs: 35.0 vs 34.9), while cp1 does not significantly depend on a single collision detection. According to Table 6 and Table 7, it is concluded that using Spatial context extension and collision detection simultaneously can effectively improve the performance of the model.

5.2. Evaluation on feature augmentation methods

5.2.1. Effect of different granulation methods

The size of the granularity is decisive to the multi-granularity deformable convolution. Different numbers of granularities are used in experiments for comparison. Table 8 shows the experimental results. DCN is the deformable convolution network chosen as the base in this paper. MGDC-v1 represents the proposed multi-granularity deformable convolution, which use two kinds of convolution kernels 1×1 and 3×3 to obtain the offsets of different levels. MGDC-v2 adds additional 5×5 granularity compared to MGDC-v1. MGDC-v3 adds 7×7 to v2. MGDC-v4 adds 9×9 granularity to v3, while MGDC-v5 adds 11×11 granularity to v4.

According to the results (baseline* represents the implementation of mmdetection), from DCN to MDGC-v1, we can see that double granularities significantly improve the detection performance (35.5 to 36.3) in comparison to single granularity, and the overall detection performance has also been improved (53.2 to 53.7). From MDGC-v1 to MDGC-v2, the model obtains a performance increase of 0.3 for small objects and overall performance with three levels of granularity, respectively. When the granularity continues increasing to 4, MGDC-v3 could improve the overall detection performance by 0.2, but its performance for small objects is worse than MGDC-v2. In addition, the training process of MGDC-v3 is more time-consuming due to its complicated model structure, so the improvement of overall detection performance in MGDC-v3 is not enough to compensate for the extra training time. When MGDC-v4 continues to add 9×9 granularity, the overall detection accuracy slightly increases, but the small object detection accuracy continues to decline. When MGDC-v5 adds 11×11 granularity, both the overall detection accuracy and the small object detection accuracy show decreasing trends, and the computational overhead will also increase significantly.

5.2.2. Visualization of multi-granularity deformable convolution

According to the conclusion of previous subsection, we compare the difference between deformable convolution with granularity of 1×1, 3×3 and 5×5. Taking ResNet50 as an example, we extract a point (the blue point shown in Fig. 13) on the output feature map of the 4th convolutional block in the network. This point can theoretically get the corresponding $(1 \times 1)^2 = 1, (3 \times 3)^2 = 81, (5 \times 5)^2 = 625$ feature points (the red point shown in Fig. 13) under low-rank, middle-rank, and high-rank granules, these feature points are used to learn the offset information of the blue point. The number may be lower than the calculated value because some points are out of the bounds.

In the overall view, when the blue feature point belongs to an object, this means that the object is needed to be detected (the red ones almost cover the entire object and can adaptively be adjusted according to different sizes of objects). When the blue feature point falls on the background, the red points will expand outwards constantly to find and determine whether the next point is still a background point. For the different levels of granularity, the higher the rank of offset granule, the wider the coverage area of corresponding red points, and the more the information used for learning offsets. Low-rank offset granule only needs one feature point to learn offsets, middle-rank utilizes a small local area around the feature point, and high-rank granule uses a large area around the feature point for learning.

5.3. Evaluation on high resolution block FPN

The baseline uses the conventional FPN, we add HR Block on its basis directly. Fig. 14 shows the experimental results, which show that the detection performance of the HR-FPN proposed in this paper is significantly better than that of the ordinary FPN for small and medium-sized objects, while the detection performance for large objects is close to that of the ordinary FPN. On the one hand, high-resolution feature maps from the HR-FPN can provide extra rich information about the details of image for small and medium size objects, it is important for the model to distinguish these objects from the background. On the other hand, since FPN has already represented the multi-layer features of large objects, HR-FPN use the high-resolution module again may lead to the generation of redundant information. Hence, the detection accuracy of HR-FPN is improved by about 0.7, it shows that HR-FPN can bring

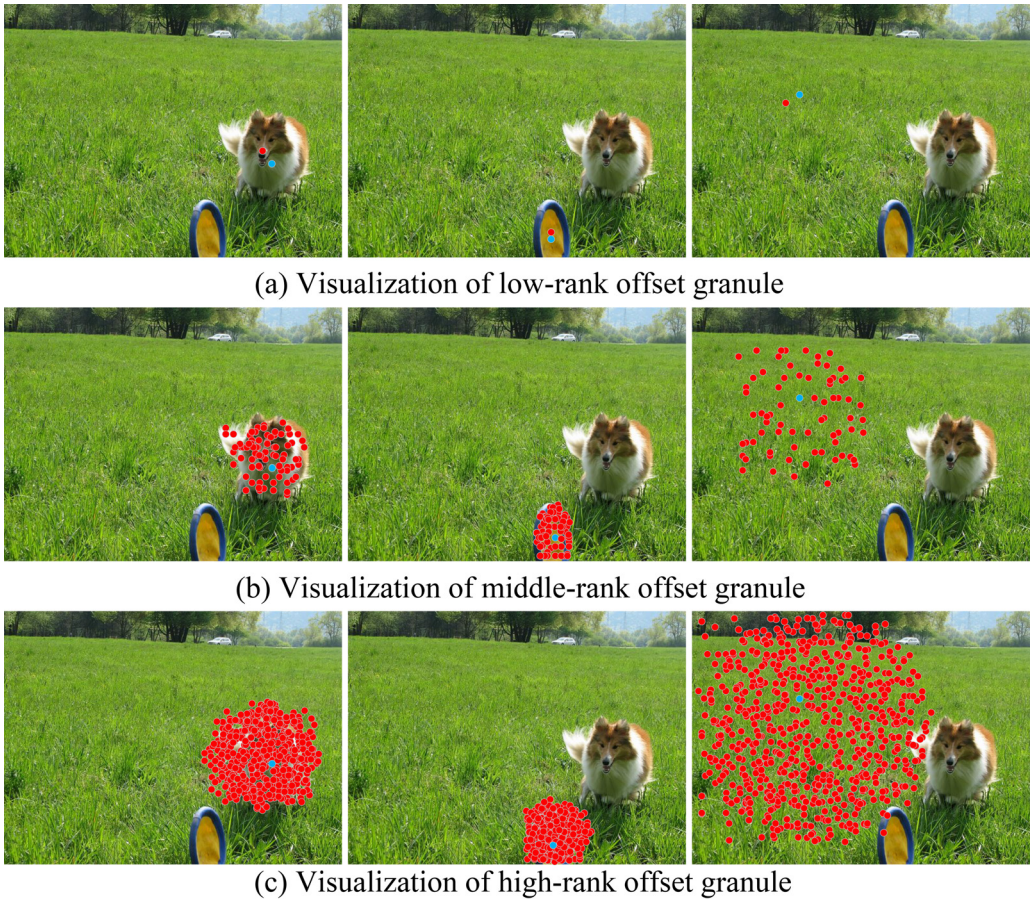


Fig. 13. The feature fusion of HR-Block.

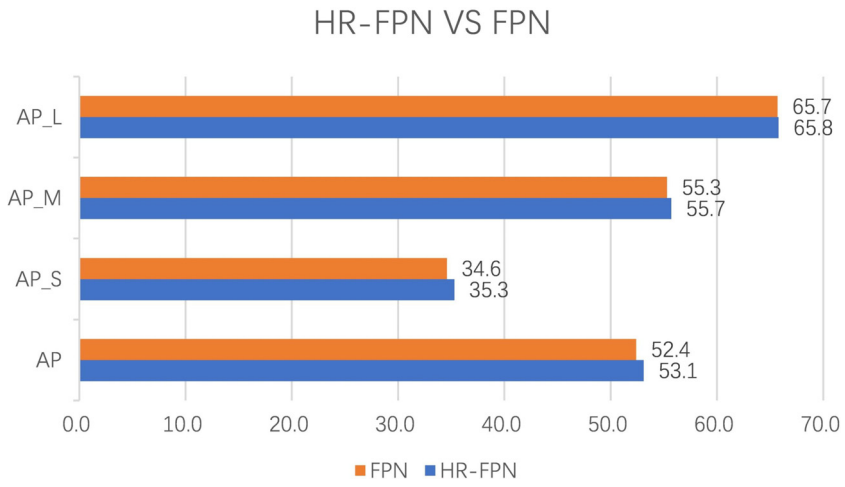


Fig. 14. The feature fusion of HR-Block.

a certain degree of improvement to the overall detection effect of the model.

5.4. FENet Feature-enhanced ablation experiment

We performed ablation experiments on the MGDC-v2 and HR-FPN. As displayed in Table 9, the model leads to a certain degree of improvement by using MGDC-v2 and HR-FPN alone, respectively. MGDC-v2 can improve the performance of the baseline model by 2.0 for small object detection and 1.6 on the overall performance, this indicates that both small and large object detection benefit from multi-granularity deformable convolution. HR-FPN brings a relatively small improvement of 0.7 on small objects and 0.7 over-

Table 9
The results of feature enhancement ablation experiments.

MGDC	HR	**AP**	**AP_{50}**	**AP_{75}**	**AP_S**	**AP_M**	**AP_L**
colrule ×	×	52.4	71.2	57.4	34.6	55.3	65.7
✓	×	54.0	72.8	59.0	36.6	56.7	**67.7**
×	✓	53.1	71.7	57.8	35.3	55.7	65.8
✓	✓	**54.6**	**73.3**	**60.2**	**37.2**	**57.1**	67.3

all, which also shows the effectiveness of HR-FPN. The ablation experiments in Table 9 show that both the proposed multi-grain deformable convolution and the high-resolution feature pyramid network can improve the overall detection accuracy of the model as well as the small object detection accuracy. The multi-granularity deformable convolution improves the small object detection accuracy of baseline from 34.6 to 36.6, while the high-resolution feature pyramid improves the small object detection accuracy of baseline from 34.6 to 35.3. The combined application makes the accuracy for small object detection finally reaching 37.2.

6. Conclusion

This paper presents the design of the detection network FENet (Feature Enhancement Network) based on collision detection, spatial context information, multi-granularity deformable convolution, and high-resolution feature pyramids, which can generate multi-pose, multi-scale and robust feature representations for small objects. In this study, a new data augmentation strategy, a novel module combining idea of multi-Granularity and deformable convolution network and a optimized HR-FPN are proposed to improve the performance of small object detection. Ablation experiments show that proposed modules can improve the ability of the small object detector, and the subjective and objective experiments demonstrate that our FENet can perform better in the overall detection capability compared with the current mainstream detection methods.

In future, FENet can be switched to different applications by changing head part, such as Swin, SSD and DETR. FENet has strong robustness and can be easily applied to multipli CNN and transformer module. However, there are also some weaknesses in this paper such as lightness of the model, performance of the object detection in scenes of occlusion (the objects are obscured by the obstacle) or crowd (the objects cover each other). Future studies are focused on the object detection in a crowd or occlusion, and by improving the proposed FENet, we will construct a new model which can detect the objects by using only a part of the object.

At the same time, small object detection plays an important role in the field of intelligent medical, aerial detection, mechanical defect detection and other tasks. The proposed method could offer some inspiration in the future studies of small object tracking, aerial object detection, which is also one of our next major works.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

References

[1] Z. Wu, K. Suresh, P. Narayanan, H. Xu, H. Kwon, Z. Wang, Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1201–1210.

[2] P. Hu, D. Ramanan, Finding tiny faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 951–959.

[3] Z. Yang, X. Wang, J. Wu, Y. Zhao, Q. Ma, X. Miao, L. Zhang, Z. Zhou, Edgeduet: tiling small object detection for edge assisted autonomous mobile vision, IEEE/ACM Trans. Networking (2022).

[4] Z. Tu, H. Li, D. Zhang, J. Dauwels, B. Li, J. Yuan, Action-stage emphasized spatiotemporal VLAD for video action recognition, IEEE Trans. Image Process. 28 (6) (2019) 2799–2812.

[5] Z. Tu, W. Xie, Q. Qin, R. Poppe, R.C. Veltkamp, B. Li, J. Yuan, Multi-stream CNN: learning representations based on human-related regions for action recognition, Pattern Recognit 79 (2018) 32–43.

[6] T.Y. Lin, M. Maire, S.J. Belongie, J. Hays, et al., Microsoft COCO: common objects in context, in: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V, in: Lecture Notes in Computer Science, volume 8693, Springer, 2014, pp. 740–755.

[7] J. Xu, W. Wang, H. Wang, J. Guo, Multi-model ensemble with rich spatial information for object detection, Pattern Recognit 99 (2020) 107098.

[8] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.Y. Lin, E.D. Cubuk, Q.V. Le, B. Zoph, Simple copy-paste is a strong data augmentation method for instance segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, Computer Vision Foundation / IEEE, 2021, pp. 2918–2928.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, C. B. Alexander, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, 2016.

[10] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: IEEE International Conference on Computer Vision, IEEE Computer Society, 2017, pp. 764–773.

[11] W.i. Ma, Y. Wu, F. Cen, G. Wang, MDFN: multi-scale deep feature learning network for object detection, Pattern Recognit 100 (2020) 107149.

[12] S. He, L. Schomaker, GR-RNN: global-context residual recurrent neural networks for writer identification, Pattern Recognit 117 (2021) 107975.

[13] T. Lin, P. Doll, R.B. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2017, pp. 936–944.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[15] H. Fang, J. Sun, R. Wang, M. Gou, Y. Li, C. Lu, InstaBoost: boosting instance segmentation via probability map guided copy-pasting, in: IEEE International Conference on Computer Vision, IEEE, 2019, pp. 682–691.

[16] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho, Augmentation for small object detection, IEEE Conference on Computer Vision and Pattern Recognition abs/1902.07296 (2019).

[17] S.-H. Lee, S.-H. Bae, AFI-GAN: improving feature interpolation of feature pyramid networks via adversarial training for object detection, Pattern Recognit 138 (2023) 109365.

[18] Z. Yang, S. Liu, H. Hu, L. Wang, S. Lin, Reppoints: Point set representation for object detection, in: IEEE International Conference on Computer Vision, IEEE, 2019, pp. 9656–9665.

[19] Z. Xia, X. Pan, S. Song, L.E. Li, G. Huang, Vision transformer with deformable attention, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 4794–4803.

[20] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, Computer Vision Foundation / IEEE, 2019, pp. 5693–5703.

[21] J. Wang, K. Sun, T. Cheng, B. Jiang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3349–3364.

[22] Z. Cai, N. Vasconcelos, Cascade R-CNN: delving into high quality object detection, in: CVPR, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6154–6162.

[23] T. Lin, P. Goyal, R.B. Girshick, K. He, P. Doll, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision, IEEE Computer Society, 2017, pp. 2999–3007.

[24] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, H. Ling, Detection and tracking meet drones challenge, IEEE Trans Pattern Anal Mach Intell (2021). 1–1

[25] N. Jiang, X. Yu, X. Peng, Y. Gong, Z. Han, SM+: refined scale match for tiny person detection, in: ICASSP, IEEE, 2021, pp. 1815–1819.

[26] S. Zhang, C. Chi, Y. Yao, Z. Lei, S.Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in: IEEE Conference on Computer Vision and Pattern Recognition, Computer Vision Foundation / IEEE, 2020, pp. 9756–9765.

[27] T. Yin, X. Zhou, P. Krähenbühl, Center-based 3D object detection and tracking, CVPR (2021).

[28] T. Wang, X. Zhu, J. Pang, D. Lin, FCOS3D: fully convolutional one-stage monocular 3d object detection, in: ICCVW, IEEE, 2021, pp. 913–922.

[29] X. Zhang, F. Wan, C. Liu, R. Ji, Q. Ye, FreeAnchor: learning to match anchors for visual object detection, Neural Information Processing Systems, 2019.

[30] J. Pang, K. Chen, Q. Li, Z. Xu, H. Feng, J. Shi, W. Ouyang, D. Lin, Towards balanced learning for instance recognition, Int J Comput Vis 129 (5) (2021) 1376–1393.

[31] K. Kim, H.S. Lee, Probabilistic anchor assignment with iou prediction for object detection, in: European Conference on Computer Vision, in: Lecture Notes in Computer Science, volume 12370, Springer, 2020, pp. 355–371.

[32] G. Tian, J. Liu, H. Zhao, W. Yang, Small object detection via dual inspection mechanism for UAV visual images, Appl. Intell. 52 (4) (2022) 4244–4257.

[33] C. Deng, M. Wang, L. Liu, Y. Liu, Y. Jiang, Extended feature pyramid network for small object detection, IEEE Trans. Multim. 24 (2022) 1968–1979.

[34] Q. Wang, Y. Qian, Y. Hu, C. Wang, X. Ye, H. Wang, M2YOLOF: Based on effective receptive fields and multiple-in-single-out encoder for object detection, Expert Syst Appl 213 (2023) 118928.

[35] none

[36] X. Zhang, W. Guo, Y. Xing, W. Wang, H. Yin, Y. Zhang, AugFCOS: augmented fully convolutional one-stage object detection network, Pattern Recognit 134 (2023) 109098.

[37] Y. Song, P. Zhang, W. Huang, Y. Zha, T. You, Y. Zhang, Object detection based on cortex hierarchical activation in border sensitive mechanism and classification-Glou joint representation, Pattern Recognit 137 (2023) 109278.
[38] Y. Shinya, Usb: Universal-scale object detection benchmark, in: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21–24, 2022, BMVA Press, 2022.
[39] P. Sun, R. Zhang, Y. Jiang, et al., Sparse R-CNN: end-to-end object detection with learnable proposals, in: IEEE Conference on Computer Vision and Pattern Recognition, Computer Vision Foundation / IEEE, 2021, pp. 14454–14463.
[40] Y. Li, Y. Chen, N. Wang, Z. Zhang, Scale-aware trident networks for object detection, in: IEEE International Conference on Computer Vision, IEEE, 2019, pp. 6053–6062.
[41] X. Zhou, V. Koltun, P. Krähenbühl, Probabilistic two-stage detection, arXiv preprint arXiv:2103.07461, 2021.

Hongyun Zhang received the Ph.D. degree in pattern recognition and intelligence system from Tongji University, Shanghai, China, in 2005. She is doctoral supervisor and currently an Associate Professor at Tongji University. She is the author or co-author of nearly 70 journal papers and conference proceedings in principal curves, pattern recognition, machine learning granular computing, and rough set Her current research interests include computer vision and pattern recognition, principal curves, data mining, rough set theory, and granular computing. E-mail: zhanghongyun@tongji.edu.cn.

Miao Li is currently pursuing his Ph.D. in Computer Sciences at Tongji University, Shanghai, China. He received his master's degree from School of Information science and Engineering, Yunnan University in 2022. His research interests include computer vision and pattern recognition, image enhancement, object detection and image segmentation. E-mail: lmiao@tongji.edu.cn.

Duoqian Miao is professor of College of Electronics and Information Engineering of Tongji University, Fellow of International Rough Set Society (IRSS), Fellow of Chinese Association for Artificial Intelligence (CAAI). Prof. Miao works in Department of Computer Science and Technology of Tongji University. Prof. Miao's research interests include Artificial Intelligence, Machine Learning, Big Data Analysis, Granular Computing and Rough Sets, etc. He has published more than 160 papers in this area, more than nine books and academic works, and nine national invention patents. E-mail: dqmiao@tongji.edu.cn.

Witold Pedrycz is a professor and Canada Research Chair in the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada. He is also with the Systems Research Institute of the Polish Academy of Sciences. He is actively pursuing research in Computational Intelligence, fuzzy modeling, pattern recognition, knowledge discovery, neural networks, granular computing and software engineering. Dr. Pedrycz is also an Editor-in-Chief of Information Sciences and IEEE Transactions on Systems, Man, and Cybernetics part A. He is the past president of IFSA and NAFIPS. He currently serves as an Associate Editor of the IEEE Transactions on Fuzzy Systems and is a member of a number of editorial boards of other international journals. He is a Fellow of the IEEE. E-mail: wpedrycz@ualberta.ca.

Zhaoguo Wang received his master's degree from Department of Electronic and Information Engineering, Tongji University, China, in 2022. His research interests include computer vision and pattern recognition, object detection. E-mail: 1930806@tongji.edu.cn.

Minghui Jiang received his master's degree from of Department of Electronic and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision and pattern recognition, object detection. E-mail: 604235572@qq.com.