

Article

Image Detection Network Based on Enhanced Small Target Recognition Details and Its Application in Fine Granularity

Qiang Fu, Xiaoping Tao, Weijie Deng and Hongliang Liu 

Key Laboratory of Advanced Manufacturing for Optical Systems (Chinese Academy of Sciences), Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, Changchun 130033, China; fuqiang@ciomp.ac.cn (Q.F.); taoxp99@163.com (X.T.)

* Correspondence: hl1367964921@126.com

Abstract: Image detection technology is of paramount importance across various fields. This significance is not only seen in general images with everyday scenes but also holds substantial research value in the field of remote sensing. Remote sensing images involve capturing images from aircraft or satellites. These images typically feature diverse scenes, large image formats, and varying imaging heights, thus leading to numerous small-sized targets in the captured images. Accurately identifying these small targets, which may occupy only a few pixels, is a challenging and active research area. Current methods mainly fall into two categories: enhancing small target features by improving resolution and increasing the number of small targets to bolster training datasets. However, these approaches often fail to address the core distinguishing features of small targets in the original images, thus resulting in suboptimal performance in fine-grained classification tasks. To address this situation, we propose a new network structure DDU (Downsample Difference Upsample), which is based on differential and resolution changing methods in the Neck layer of deep learning networks to enhance the recognition features of small targets, thus further improving the feature richness of recognition and effectively solving the problem of low accuracy in small target object recognition. At the same time, in order to take into account the recognition effect of targets of other sizes in the image, a new attention mechanism called PNOC (protecting the number of channels) is proposed, which integrates small target features and universal object features without losing the number of channels, thereby increasing the accuracy of recognition. And experimental verification was conducted on the PASCAL-VOC dataset. At the same time, it was applied to the testing of the fine-grained MAR20 dataset and found that the performance was better than other classic algorithms. At the same time, because the proposed framework belongs to a one-stage detection method, it has good engineering applicability and scalability, and universality in scientific research applications are good. Through comparative experiments, it was found that our algorithm improved the performance of the mAP by 0.7% compared to the original YOLOv8 algorithm.

Keywords: image detection; one-stage; small target; fine-grain; attention mechanism



Citation: Fu, Q.; Tao, X.; Deng, W.; Liu, H. Image Detection Network Based on Enhanced Small Target Recognition Details and Its Application in Fine Granularity. *Appl. Sci.* **2024**, *14*, 4857. <https://doi.org/10.3390/app14114857>

Academic Editor: Sungho Kim

Received: 27 March 2024

Revised: 28 May 2024

Accepted: 29 May 2024

Published: 4 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Target detection in remote sensing images is one of the basic tasks at present, and it can be subdivided into large target recognition and small target recognition under the background of detection [1–3]. Table 1 illustrates the size limits for size targets.

Table 1. Defining the size of the target (Adapted from ref. [4]).

	Min Rectangle Area	Max Rectangle Area
Small object	0 × 0	32 × 32
Medium object	32 × 32	96 × 96
Large object	96 × 96	∞ × ∞

Through a large number of algorithm competitions and experimental data, it can be identified and concluded that the recognition accuracy of small targets is not comparable to that of large targets. For example, in the proposal of the top-level MS COCO (Microsoft COCO) [5] strength segmentation process described in Table 2, AP^S represents the recognition average precision of small targets, while AP^L represents the recognition average precision of large targets. It can be seen that in the comparison of AP detection, small targets are about 2–3 times lower than large targets.

Table 2. AP detection results for large and small targets.

	AP^S	AP^L
Megvil (Face++)	0.343	0.660
UCenter	0.326	0.648
MSRA	0.343	0.627
FAIR Mask R-CNN	0.328	0.627
Trimps-Soushen-QINIUI	0.310	0.610
bharat umd	0.312	0.606
DANet	0.283	0.591
BUPT-Priv	0.251	0.566
DL61	0.246	0.551
DENet	0.222	0.584

In the proposal for the MS COCO instance segmentation challenge, the AP detection tabulation for small targets was 2–3 times lower than for large targets.

The detection of small targets is not available in many branch tasks, such as identifying different types of aircraft. In a large number of remote sensing images, we can find that it is not easy to create high-resolution scene photos to detect small or distant objects. Generally speaking, based on the average resolution of 0.5–5 m per pixel, the object to be detected is only a few pixels in size. In the process of sampling at the Neck layer of deep learning, these small pixel objects to be measured are often important features that are ignored and thus cannot be taken into account. Based on such characteristics, we will find that even mature samples such as Mask-RCNN predict that small objects will appear even if they can be identified by human eyes [6]. However, they will still be missed by the system, as shown in Figure 1.

In the sample prediction of Mask-RCNN in Figure 1, many small targets that can be recognized by the naked eye will be missed, for example, only six can be detected from hundreds of flying birds, which is attributed to multiple factors. On the one hand, there is occlusion among the birds, thus resulting in incomplete recognition. On the other hand, the size of the birds is also a factor. The birds in the image are small objects, and it has been observed in practice that they are not recognized as effectively as larger objects. So, how to improve the accuracy in identifying small objects is an important research topic.

Based on the above problems, some scholars, such as Professor Tiexin Qin, proposed to replicate small targets several times to improve poor recognition by increasing the number of training samples through recurring features, and corresponding experiments were set up and achieved good results. However, this method does not fundamentally solve the problem of obtaining fewer features for small target objects using conventional feature extraction methods, so the effect is not stable.

Some scholars, such as Professor Duan K, also proposed setting different styles of measurement boxes and raising the threshold of small targets by reformulating the calculation method of similarity. Indeed, through the efforts of a large number of scholars, they have made good progress in the corresponding fields [7]. However, even if the scheme proposed by scholars has good robustness in the identification of small targets, the effect of application in the fine-grained detection of lower layers is not satisfactory.



Figure 1. Examples of identifying small targets with a large number of missing targets.

Fine-grained classification is a small branch of object recognition, which is the subdivision that distinguishes a common category, such as the classification of a dog, the model of a ship, or the identification of aircraft in the same category [8,9]. Although deep learning has good experimental results for many visual tasks, its application in fine-grained classification is not satisfactory, which is largely due to the above-mentioned inability to effectively come up with enough distinguishing features [10–12]. Moreover, the transformation of different poses can also obscure many necessary features.

In the common fine-grained recognition idea, some scholars think that the fine-grained foreground is very similar, but it will be disturbed by the background, which leads to the problem of recognition. Therefore, their research is on how to separate the foreground and background to a greater extent. Some scholars believe that we should rely on human prior knowledge, that is, to classify the poses that humans are good at. Recognition and machine learning coordinate to form an automatic classification.

Some scholars, such as Professor Kong, T and Sun, F, are also committed to the pooling layer of machine learning, thus hoping to use higher-order pooling to improve the accuracy of large recognition. With the popularity of the Transformer, the attention mechanism becomes hot, constantly focusing on the next feature point based on the existing feature, to carry out detailed identification [13].

In response to the aforementioned technical issues, this article proposes to design a new network structure DDU (Downsample Difference Upsample) to enhance the features of small targets, thus fundamentally solving the problem of recognition difficulties. DDU is based on differential and resolution changing methods in the Neck layer of deep learning networks to enhance the recognition features of small targets, thus further improving the feature richness of recognition and effectively solving the problem of low accuracy in small target object recognition. The enhanced network approach is based on the effect of upsampling and downsampling to perform appropriate differentiation, thereby achieving the effect of shielding irrelevant features and focusing on necessary features. At the same time, in order to prevent the blocked features from having other factors that are conducive to recognition, the existing attention mechanism has been improved, thus forming a new attention mechanism PNOC (preserves the number of channels). In the dimension of channel attention, the effect of avoiding the common two full connections without losing the

number of channels has been achieved. While achieving feature fusion, the disadvantage of avoiding key features caused by dimensionality reduction has been avoided.

We have made the following improvements and extensions in this article:

1. In the network framework of the Neck layer, the DDU network has been designed, which mainly obtains the features of small targets based on up- and downsampling, thus enhancing the recognition features of small targets, further improving the feature richness of recognition, and effectively solving the problem of low recognition accuracy of small targets.
2. This article also utilizes the improved attention fusion method PNOC to integrate the small target features extracted by the DDU module with the original feature targets sampled without differential operation, thereby making the data usable.
3. We compared our proposed method with other common algorithms in the PASCAL-VOC dataset and completed ablation experiments that incorporated the DDU and PNOC modules into the algorithm, as well as experiments involving only the DDU module or only the PNOC module and experiments without either the DDU or PNOC modules. These experiments aimed to validate the rigor of the algorithm theory. This paper found that our algorithm has good performance.
4. Subsequent experiments were conducted in the MAR20 [14] fine-grained identification database, and corresponding comparative experiments were conducted with other first-order algorithms, and it was found that our algorithm still has good performance and experimental robustness in the fine-grained database.

2. Related Work

As a hot research topic in the field of computer vision, small target detection has been proposed by many scholars in recent years, which can be divided into the following methods.

Firstly, feature learning based on different dimensions is mainly aimed at solving the problem of small targets with few features. Based on this, scholars have put forward feature pyramid-based and receptive field-based methods, in which feature pyramid-based methods are represented by FPNs (Feature Pyramid Networks). The main idea is to combine the low-level and high-level semantics with more spatial information to enhance the characteristics of the target [15]. The receptive field-based approach is represented by the Trident Network, and the idea is to assign different receptive fields based on the size of the target. Specifically, the dilated convolution with different dilation rates is used to form three branches of objects with different receptive fields that are responsible for different scales [16].

Secondly, some scholars put forward the GAN-based method, which solves the problem that the small target itself has few identification features, and its thinking and practice are also simple and plain: using GANs (Generative Adversarial Networks) to generate high-resolution pictures or high-resolution features. They are not limited to SOD-MTGAN methods for generating high-resolution features, and the general idea is to use the trained detector to obtain subgraphs containing targets. For example, Faster R-CNN then uses the generator to generate the corresponding high-definition image, and based on the discriminator, it is responsible for whether the generated image is real or fake [17–19]. At the same time, this part is carried out from the detector to further predict the category and position of the object. Common ways to generate high-resolution features are Perceptual GANs and BFFBB GANs (Better to Follow, Follow to Be Better Generative Adversarial Networks) [20,21]. Its core ideological history takes the input image shrunk by two times and the features extracted through feature extraction as the low-resolution feature and the features extracted from the original image as the corresponding real high-resolution feature; the generator based on a low false high-resolution feature is generated by the resolution feature. The discriminator is responsible for distinguishing the false and true high-resolution features.

Thirdly, some scholars put forward context-based thinking, which uses the dependency between the environmental information of small objects and other easily detected objects

to enhance detection visibility. The essence of problem solving is to make up for the lack of features of small objects [22,23]. Common methods are inside-OutsideNet and Pyramid box, where inside-OutsideNet extracts information about the global context of an object through four inverted RNN (Recurrent Neural Network) pyramid boxes [24,25]. Pyramid box aims to use the prior information that the head and body always appear together with the face, because the first two are relatively easier to detect than the face, and then use the detected head and body to position the face for further detection. The highly dependent relationship is based on the currently hot attention mechanism Transformer combined with implicit modeling of the relationship between two objects and the use of this dependency relationship to strengthen the characteristics of each objective target.

Next, some scholars have proposed certain solutions for specific datasets, but this method is only useful for specific datasets, such as MS COCO datasets. In image-level and instance-level imbalance, the image-level imbalance is that only 51.8% of pictures in MS COCO have small objects [26]. Therefore, the outcome is to set a Stitcher as the feedback signal for the proportion of loss of small objects in the total loss. When the proportion is lower than the threshold, the four pictures are stitched together into one picture for input, thus increasing the number of small objects. Instance-level imbalance means that the pixel area of small- and medium-sized objects in the MS COCO only accounts for 1%, and augmentation is solved using the copy–paste method. However, these two thoughts do not fundamentally solve the problem, which has strong limitations [27].

Last, some scholars put forward the improvement for the current targets. Special design for small objects refers to the method that has a special processing method for the small objects. Specifically, a way to describe how dense the anchor is on the original image is defined for the FaceBoxes. Specifically, it is defined as follows:

$$A_{density} = A_{scale} / A_{interval} \quad (1)$$

A_{scale} represents the scale of the anchor, $A_{interval}$ is the distance between the central points of the anchor, and as for $A_{density}$, the larger of the size, the denser the anchor in the original image [28].

The authors calculated the density of objects of different sizes and found that the distribution of anchors for small objects was too sparse, thus increasing the number of anchors. However, S3FD's approach is more direct, thus directly reducing the IoU (Intersection over Union) threshold of the positive sample of the object. If the matched anchors still cannot be effectively increased, the previous N anchors are selected as the matched anchors among all the anchors whose IoU is greater than 0.1. Some scholars also designed a new evaluation standard, DotD, to replace IoU as a positive and negative assignment. This method mainly calculates the function of the distance between the center point of a bbox and makes use of the insensitivity of the index to the migration of the bbox so that a small amount of migration will not lead to a drastic decline of the index. At the same time, some scholars have developed a new way to detect small objects by using a specially trained guess path, specifically the SNIP (Selective Kernel Pruning) algorithm, which is proposed based on the training method of the image pyramid. The method of detecting large objects using small-size images and small objects using multiscale images ensures that the objects in the input classifier are close to 224×224 . Thus, it is consistent with the training scale on ImageNet to improve the detection performance of small objects [29].

3. Textual Method

The algorithm framework in this paper is divided into four system parts, namely Backbone, Neck, Head, and Loss. As shown in Figure 2, the Neck layer extracts the DDU structure of small target features, and the optimized attention mechanism PNCO module of feature fusion is the innovation point of this paper.

The following part will briefly introduce the Backbone and Head layer and then focus on the Neck layer optimized in this paper as follows:

The network mechanism proposed in this method is based on the universality and real-time consideration of engineering applications, so the thinking of first-order object detection has been adopted. Therefore, a new deep learning network structure for target recognition has been proposed based on YOLOv8 [30,31]. The Backbone of this network uses the CSP (CrossStagePartial) framework. The existence of this framework not only improves the learning ability of the CNN but also reduces the calculation cost of the model to a certain extent. At the same time, different from other combinations of three convmodules and n DarkNets, its network is composed of two convmodules and n Darknets and connected through Split and Concat. Meanwhile, the ConvModule here is composed of Conv-Bn-Relu. Compared to the ConvModule composed of Conv-Relu-Bn, the structure adopted in this paper is more reasonable.

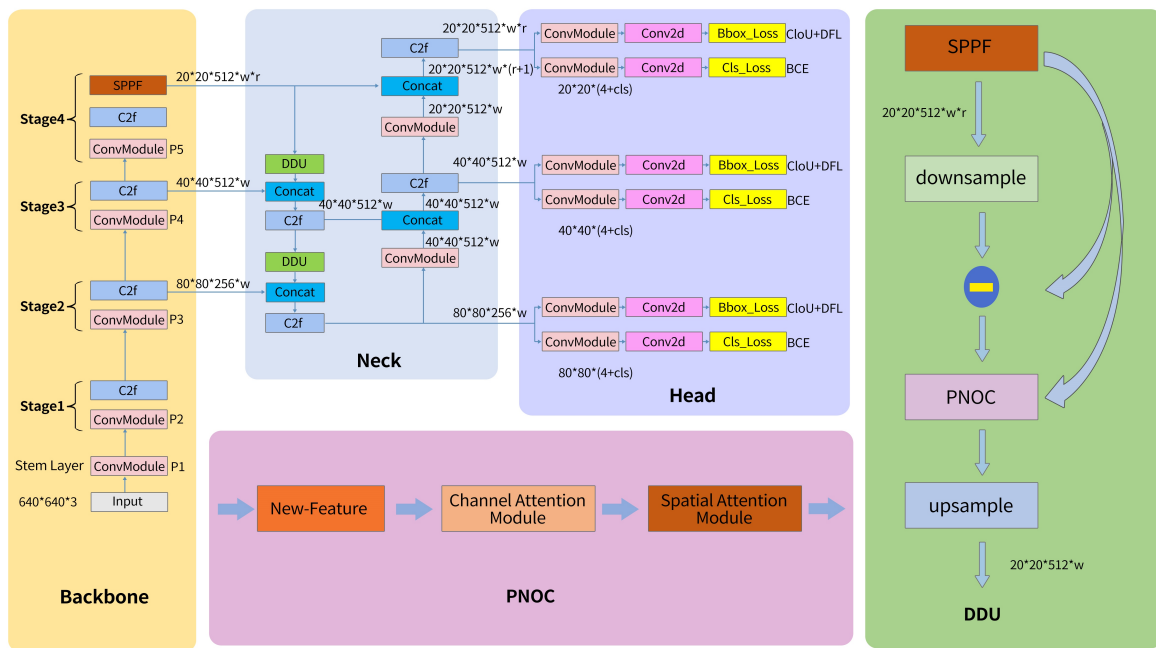


Figure 2. The algorithm framework diagram of this article.

For the detection head of this network, the decoupled head was selected, which is embodied in that the classification and detection head are separate. For the loss function, positive and negative samples were allocated. We used the Task-Aligned One-Stage Object Detection Task-Aligned Assigner in the training model to select positive samples according to the weighted scores of classification and regression. The formula is shown as follows:

$$t = s^{\alpha} \times u^{\beta} \quad (2)$$

The s represents the prediction of tags in the class big score. The u represents the prediction of the IoU and the ground truth bounding box. In addition, our model classification uses binary crossentropy, whose equation is as follows:

$$Loss_n = -w[y_n \log x_n + (1 - y_n) \log(1 - x_n)] \quad (3)$$

where w is the weight, y_n is the annotation value, and x_n is the model prediction value.

The regression branch adopts the distributed Focal Loss DFL and the Complete IoU $CIoU$, where the DFL is used to expand the probability around the object y , and the specific equation is as follows:

$$DFL(\varphi_n, \varphi_{n+1}) = -((y_{n+1}y) \log(\varphi_n) + (y - y_{n-1}) \log(\varphi_{n+1})) \quad (4)$$

The formulas for φ_n and φ_{n+1} are shown below:

$$\varphi_n = \frac{y_{n+1} - y}{y_{n+1} - y_n} \quad \varphi_{n+1} = \frac{y - y_n}{y_{n+1} - y_n} \quad (5)$$

The *CIoU* and *DIoU* involved here add an impact factor on the basis of Loss while considering the predicted aspect ratio and ground truth boundary box, and the specific equation is as follows:

$$CIoU_{Loss} = 1 - IoU + \frac{Distance_C^2}{Distance_C^2} + \frac{v^2}{(1 - IoU) + v} \quad (6)$$

where v is the parameter to measure the consistency of the aspect ratio, which is defined as follows:

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w^p}{h^p} \right)^2 \quad (7)$$

where w is the weight of the bounding box, and h is the height of the bounding box.

For the network at the Neck layer, generally speaking, the deeper the network is, the more feature information it obtains. However, the deeper network will reduce the position information of objects, and too much convolutional behavior will lead to the loss of large information of small objects. Therefore, it is crucial to merge the necessary feature pyramid with the features that enhance the small target information and classify them into target features. In this model, the FPN is used for upsampling from top to bottom. The purpose of such convolution is to increase the feature information in the bottom feature graph. At the same time, a DDU network designed by ourselves was added to resample the features of small target objects, and then the two features were truly combined, thus ensuring the accurate prediction of all kinds of small images. For the fused part, a new attention mechanism fusion method different from Resnet was adopted. This method takes into account other image information while retaining the features of small objects. This increases universality.

3.1. DDU Network Structure

Assuming that the data processed by SPPF is $20 \times 20 \times 512 \times w \times r$, first downsample the data to obtain an easily identifiable feature set, and then complement the feature with the SPPF set so as to obtain the features of small targets that are not easy to obtain. This feature is fused with the original SPPF feature using the attention mechanism of PNOC (preserves the number of channels), and the obtained data are upsampled so that the output data of $20 \times 20 \times 512 \times w$ are obtained [32–34].

The core function of DDU (Downsample Difference Upsample) is to spare no effort to retain the features of small targets and use this as the fundamental basis for detecting small targets. Usually, when constructing image pyramids, we will find that the downsampled images have strong robustness against noise, and at the same time, the risk of overfitting can be reduced in deep learning training. At the same time, the operation is simplified. Although the size of the receptive field is increased, the actual experiment proves that the simple downsampling will not greatly improve the actual detection performance. On the contrary, in the detection of small targets, features based on upsampling are easy to lose their detailed features, thus making features that are not easy to extract more scarce. Therefore, the DDU method designed in this paper first performs subsampling to obtain obvious features of large target objects, which play a small role in the detection process of small targets and even cause interference. Therefore, this part of the features obtained are compared with the original image that has not been subsampled so as to obtain features of small targets that are not suitable for extraction. In simple terms, the feature of the original image is a large set B, and in simple terms, the feature of the original image is a large set F. Through downsampling, the feature set BF of the large target is extracted, which is relatively easy to extract. Then, F and BF are supplemented to obtain features that

can be used for small targets. Next, in order to unify the size of the image, upsampling is performed, as shown in Figure 3.

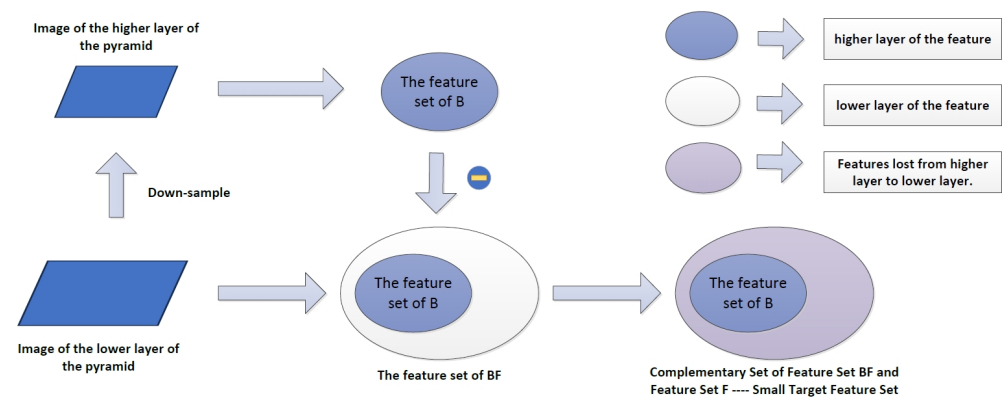


Figure 3. Graphic representation of extracting small target feature sets under DDU framework.

3.2. PNOC Network Structure

Through the above DDU network structure, the features of small targets can be directly obtained, and then how to integrate its features well with the features originally used to identify good results is the key to our experiment. Taking into account many methods proposed by previous scholars, the fusion algorithm of this time is intended to achieve an overall consideration. Under the condition of ensuring the stability of the features originally used for identification, the new features were reasonably fused to reduce the negative interference of the new features on the old features. The fusion method designed in this paper adopts the fusion method of attention mechanism, which is very popular now.

Common attention mechanisms are divided into channel attention mechanisms, spatial attention mechanisms, and mixed attention mechanisms. The new attention mechanism designed in this paper belongs to the mixed attention mechanism. Considering that the features applied by this attention mechanism are the integration of purified small target features and global features, dimensionality reduction in the traditional attention mechanism should be avoided as much as possible so as to improve the protection of learning channels and further optimize the effect of the attention mechanism on channels. At the same time, because the features processed are highly functional, therefore, regardless of the negative effects brought by efficiency and effectiveness and based on this, the channel attention mechanism network has been designed as follows:

1. The input features through the global average pooling: the features will evolve from $H \times W \times C$ matrix to $H \times W \times 1$ vector.
2. Perform Softplus operation on the obtained $H \times W \times 1$ vector, thus obtaining the weight of each channel.
3. Multiply the normalized weights and the original input feature channels to generate the weighted feature map.

On the one hand, selecting the Softplus curve can alleviate the problem of gradient disappearance; on the other hand, it can make each channel have a weight so as to avoid the existence of a large number of limit values when there is an input of large or small data under extreme conditions. The specific channel attention network is shown in Figure 4.

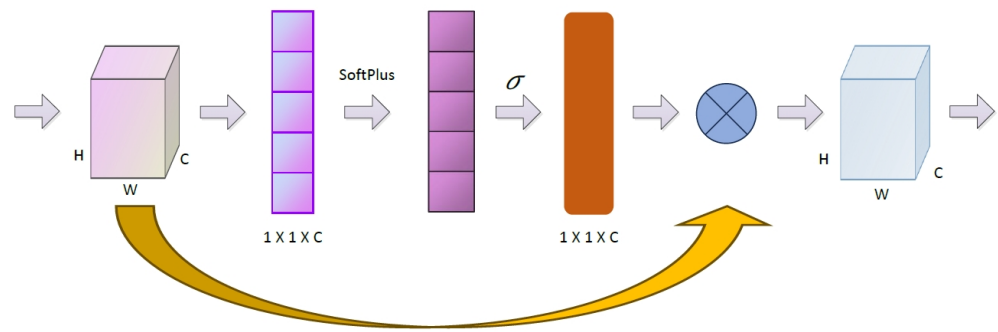


Figure 4. Channel attention module in PNOC network.

Different from the channel attention mechanism, where spatial attention is concentrated in the part of effective information, which is a strong supplement to channel attention [35], the specific network construction has been as follows:

1. Create attention graph $M_S(F) = R * W * H$ using convolutional layers to generate space.
2. Perform two pooling operations to aggregate mapping information of features and record them as F_{avg}^S and F_{max}^S , both of which belong to $R * H * W$, where F_{avg}^S represents average pooling feature and crosschannel feature.
3. Then, using standard convolution layers to connect and convolution, a spatial attention graph of two-dimensional space is generated. In short, the formula is

$$M_S(F) = \sigma(f^{7 \times 7}([AvgPool(F); MaxPool(F)])) \quad (8)$$

$$= \sigma(f^{7 \times 7}([F_{avg}^S; F_{max}^S])) \quad (9)$$

where σ denotes the sigmoid function, and $f^{7 \times 7}$ represents a convolution operation with the filter size of 7×7 . The specific structure is shown in the Figure 5.

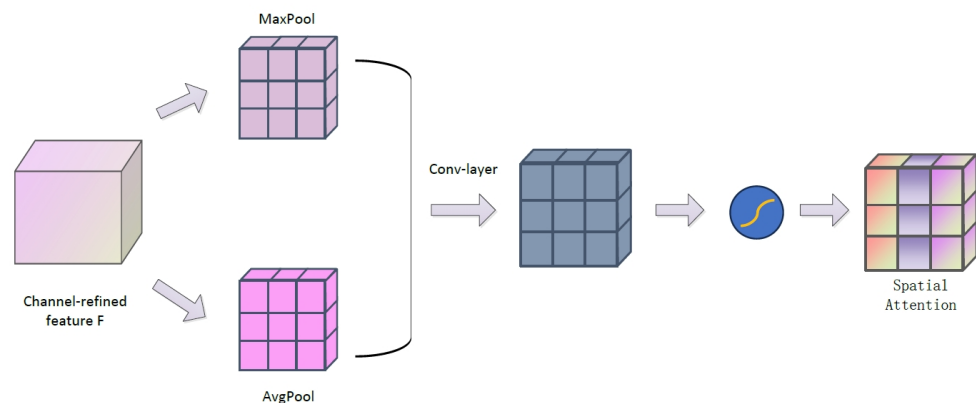


Figure 5. Spatial attention module in PNOC network.

The integrated mixed attention mechanism: Through the above channel attention mechanism, we can clearly focus on where, and through the spatial attention mechanism, we can clearly focus on what. How to superposition the two is a problem that needs to be discussed; the common superposition method is the parallel arrangement or sequential arrangement. Through previous experiments, we found that this algorithm is more suitable for sequential sorting, and in the sequential sorting, we found that when the priority of the space lags behind the priority of the channel, the effect is more excellent. Therefore, the mixed attention mechanism was formed based on the superposition of the sequential ordering of the channel attention and then the space attention, as shown in Figure 6.

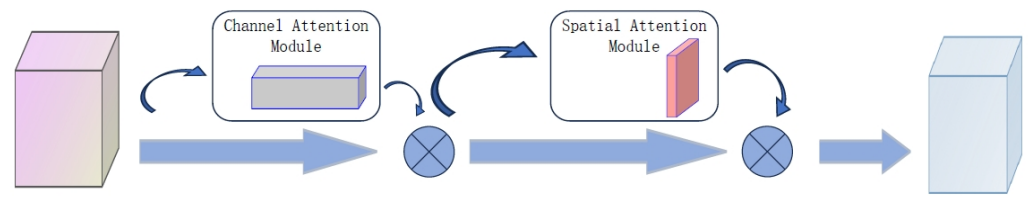


Figure 6. PNOC network structure.

4. Experiment and Analysis

In this chapter, we introduce the process of model training, verification, and testing on the dataset, including the architecture of the model and the data enhancement technology used in the training process. Figures 7 and 8 show the training process and evaluation flow chart of the PASCAL-VOC dataset and the MAR20 dataset.

The PASCAL-VOC dataset [36], with the full name Pattern Analysis, Statical Modeling and Computational Learning, is a large-scale data set of object detection, segmentation, key point detection, and caption. VOC is divided into two versions: voc2007 and voc2012, with VOC2007 including 9903 images divided into training sets (5011 images) and test sets (4952 images) [37,38]. In addition, the 2012 version has added 17,125 additional training images as a supplementary training set. Common usage methods include:

1. Only train with VOC2007 and test with VOC2007.
2. Only train with VOC2012 and test with VOC2012. This usage is rarely used, as it is commonly used in conjunction with VOC2007. (Train with VOC2007's train+val and VOC2012's train+val, and then use VOC2007's test. This usage is commonly seen in papers as 07+12, and researchers can test the results on VOC2007 themselves, because VOC2007's test is public).
3. Use VOC2007's train+val+test and VOC2012's train+val training, and then use VOC2012's test, which is commonly seen in papers as 07++12. This method needs to be submitted to the VOC official server for evaluation results, as the VOC2012 test has not been published.

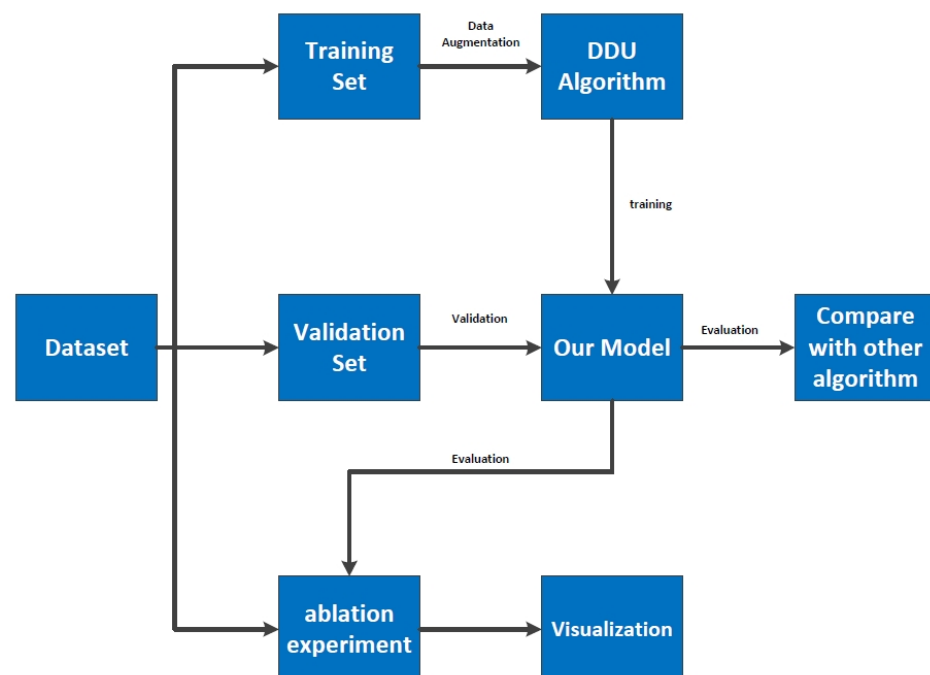


Figure 7. Application of PASCAL-VOC dataset in our network.

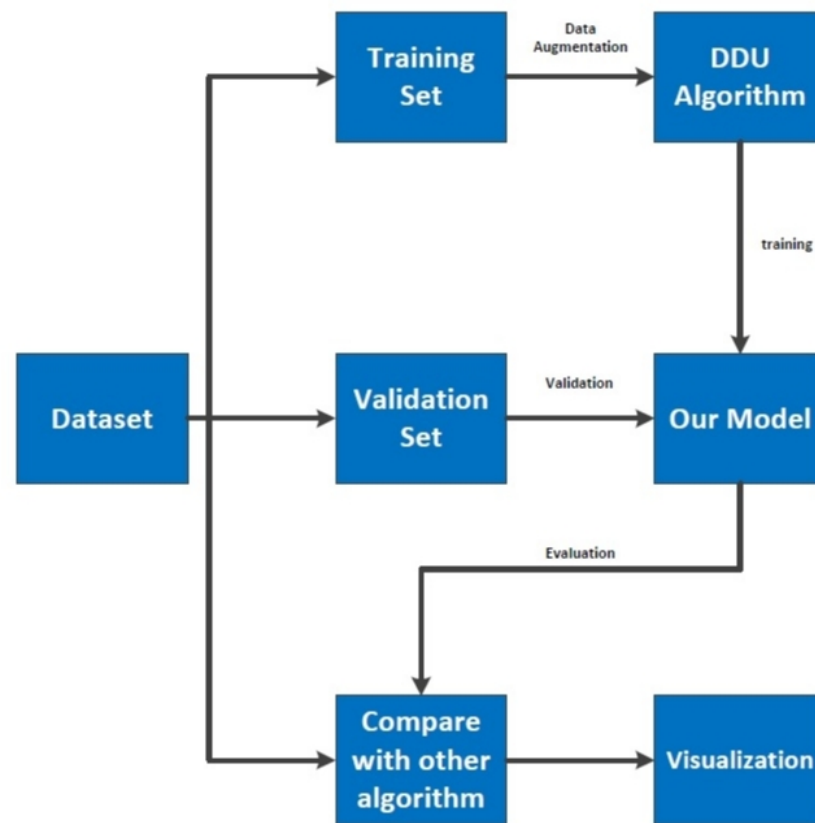


Figure 8. Application of MAR20 dataset in our network.

The method used in this experiment is 07 + 12. The experiment has 20 classes and contains 22,136 training images (voc2007 and 2012 trainval), as well as 4952 test images (voc2007test). We chose the average precision (AP) and average recall (AR) as evaluation indicators.

The specific experimental data of the proposed algorithm and other comparison algorithms in the PASCAL-VOC dataset are shown in Table 3.

Table 3. The experimental data of this algorithm and other algorithms are compared in the PASCAL-VOC dataset (IOU = 0.5) [1].

Method	Backbone	Input Size	mAP
Two-Stage Network			
OHEM	VGG16	1000 × 600	74.6
IPG RCNN	IPGNet101	1000 × 600	84.8
CoupleNet	ResNet101	1000 × 600	82.7
FPN-Reconfig	ResNet101	1000 × 600	82.4
Fast-RCNN	Faster RCNN101	None	68.4
Faster-RCNN	Faster RCNN101	None	70.4
One-Stage Network			
SSD300	VGG16	None	72.4
SSD512	VGG16	512 × 512	74.9
YOLO	GoogLeNet	448 × 448	57.9
YOLOv2	Dark-Net19	352 × 352	44.0
YOLOv3	Dark-Net53	320 × 320	51.5
YOLO-ultralytics	YOLOv8	640 × 640	75.1
Ours	YOLOv8	640 × 640	75.8

The statistical data in the table reflect an objective fact: the algorithm proposed in this paper demonstrates excellent performance in recognizing small targets within the PASCAL-VOC dataset. Compared to some of the two-stage networks, despite having a smaller input size, our algorithm exhibited a significant increase in accuracy. For example, in comparison with the OHEM algorithm, this results in an improvement of 1.2%. Among the enumerated one-stage network methods, our algorithm demonstrated a more precise recognition effect compared to the closest-performing SSD512 algorithm across different backbone frameworks, thus leading to an enhancement of 0.9%. The reason SSD512 is considered the method closest in performance to ours is because of two main factors: First and foremost, both algorithms belong to the category of one-stage methods, which enhances the credibility of the comparison. From the results, it is evident that our algorithm's improvement is indeed significant. However, it is important to note that the input size of the SSD512 algorithm is smaller than that of our algorithm, thus giving our algorithm an advantage. Therefore, based on the conventional way of comparing state-of-the-art methods, disregarding the input image size and focusing solely on the results, our algorithm is considered advanced. On the other hand, the reason SSD512 is considered the method closest in performance to ours is also attributed to its smaller input size compared to our algorithm. Additionally, since our paper is derived from the YOLO-ultralytics framework, it falls under the same line of thought as YOLO-ultralytics. However, it is crucial to mention that the method mentioned in the paper as the closest is based on a completely different approach or computational method. Therefore, despite YOLO-ultralytics outperforming SSD512, the authors still refer to SSD512 as the method closest in performance to ours in the conclusion. The reason this paper innovates on top of YOLOv8 is that even the unoptimized YOLOv8 algorithm still demonstrates advancement in accuracy compared to previous versions. Therefore, this paper chooses to build upon the already excellent framework and innovate further, thus aiming to advance accuracy. The conclusions drawn from the comparison are as follows: In comparison with other frameworks within the YOLO series, our algorithm exhibits varying degrees of improvement, thus notably achieving a 0.7% enhancement compared to the YOLOv8 algorithm. Based on these data, the performance enhancement is primarily attributed to the design of the small target feature module DDU and the integration of the novel attention mechanism PNOC algorithm that fuses features.

For this purpose, we conducted corresponding ablation experiments under the experimental conditions of 100 epochs, with image dimensions fixed at 640×640 and other parameters unchanged. The experiments were aimed at verifying the impact of the DDU module and PNOC module on accuracy, recall, mAP_{50} , and mAP_{50-95} . The results, as shown in the Table 4, demonstrate the effectiveness of the DDU and PNOC modules.

Table 4. The ablation experiment of the algorithm in this paper.

Method	P	R	mAP_{50}	mAP_{50-95}
YOLOv8	71	59.2	66	46.5
YOLOv8+DDU	72.4	59	66.4	46.9
YOLOv8+PNOC	70.6	58.4	65.6	46.1
YOLOv8+DDU+PNOC	71.9	59.3	66.7	47.1

Through the data statistics of the ablation experiment, it can be concluded that the designed algorithm does have corresponding effects, which are specifically as follows: In terms of small target detection item mAP_{50} alone, the accuracy of adding DDU and PNOC modules increased by 0.3% over the module only adding DDU, 1.1% over the module only adding PNOC, and 0.7% over the baseline. It should be noted that the results of Table 4 are based on Table 3, but Table 3 shows the values determined without any change in accuracy after multiple iterations, while Table 4 shows the comparative data of the relevant ablation experiments conducted after 100 epochs.

The number of epochs for this experiment was set to 100, and through experimental comparisons, it was observed that the network performance significantly improved with

the addition of the DDU and PNOC modules compared to the original code or the case of adding only one new module. Additionally, in this experiment, it was noted that the results of adding only the PNOC module decreased by 0.4 compared to the results of the original code. This suggests that attention mechanisms are more suitable for being used in conjunction with sets that have clear features of small targets. Otherwise, the meaningless stacking of attention mechanisms on the original feature set may have a counterproductive effect on the results.

Figure 9 is a selection of some noticeable differences in the visual detection results.

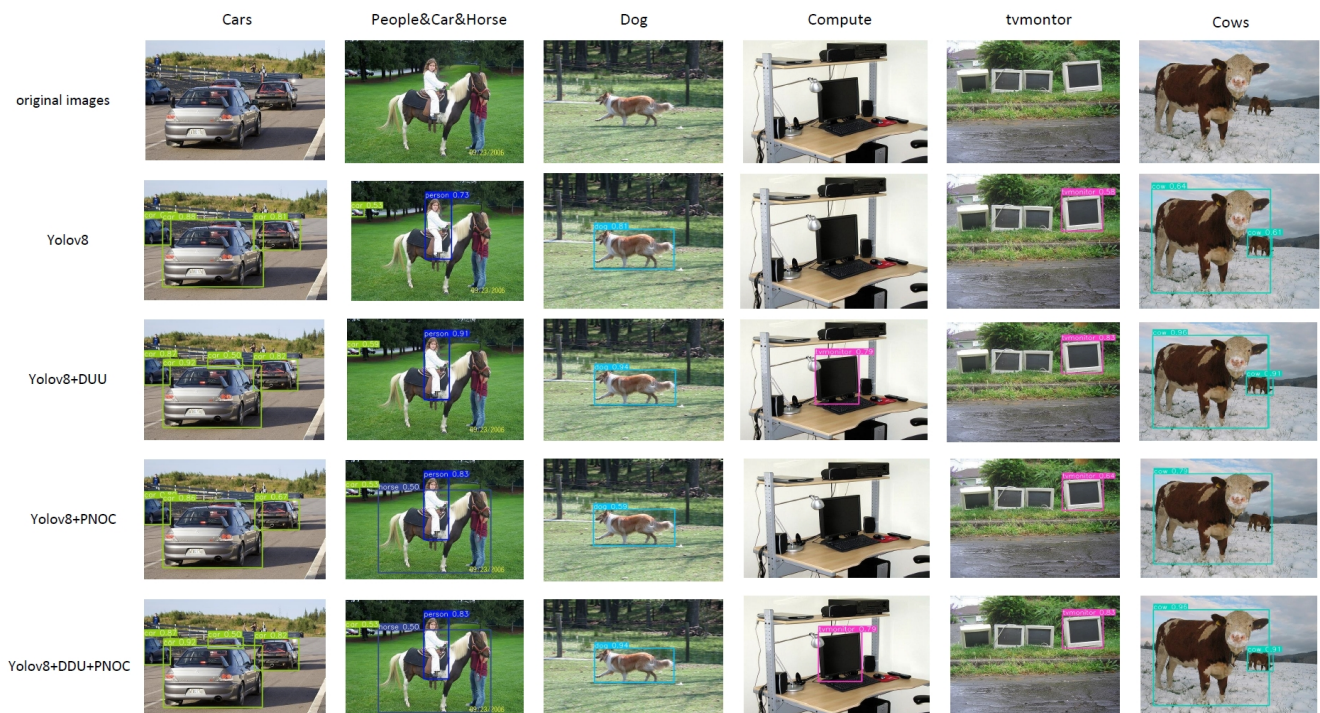


Figure 9. Visualization results of the VOC dataset.

Through visual comparison of the test graphs in different scenarios, we have demonstrated the advanced nature of the algorithm presented in this paper while also identifying some issues. Based on this, we have chosen to describe in detail certain situations in which the performance was good yet with unresolved issues, thus providing a meaningful background for future research. Specifically, we verified the algorithm using an image named “Cars” to test its recognition performance in complex backgrounds with multiple identical target objects overlapping and interacting. The experiments showed that the original YOLOv8 code and YOLOv8+PNOC missed some targets, while the algorithm proposed in the paper and YOLOv8+DDU performed better. We also tested the “People & Car & Horse” image to assess the recognition performance when cars are in the distant background and to test the recognition of both people and horses when they overlap and obstruct each other. The results show that although the four scenarios had different recognition accuracies for distant objects, they all showed good outcomes. However, in the case of obstruction between people and horses, the YOLOv8 original code and YOLO+DDU could not distinguish between the two types of objects. In contrast, YOLOv8+PNOC with attention mechanisms and the proposed algorithm were able to recognize features more comprehensively, but further optimization is needed to effectively recognize heavily obscured people. The verification of images such as “Dog,” “Cows,” and “tv-monitor” was based on the exploration of recognition performance for nonhuman faces and objects without prominent features. The experiments proved that the proposed algorithm performed well in nonhuman face categories, thus achieving high precision when successfully identifying

objects. Choosing the “Computer” image for testing was done to verify how different algorithms performed when encountering objects not in the training set. The experiment showed that the proposed paper mistook a computer screen for a “tv-monitor,” which, given their similarities, is an understandable misclassification for objects not learned before. Overall, from a subjective visual evaluation, the proposed algorithm performed well in other scenario comparisons of the ablation experiment and unearthed some issues that provide meaningful directions for further research.

The MAR20 dataset has the following characteristics [39]: (1) MAR20 is the largest military aircraft target recognition dataset with remote sensing images, including 3842 images, 20 types of military aircraft models, and 22,341 instances, and each target instance has two labeling methods—horizontal boundary box and directed boundary box. (2) Because all fine-grained categories belong to the aircraft category, different types of aircraft often have similar characteristics, thus resulting in the higher similarity between different types of targets. (3) Due to the influence of factors such as climate, season, illumination, occlusion, and even atmospheric scattering in the process of remote sensing image acquisition, there are large intraclass differences among the targets of the same model. This feature is consistent with our fine-grained application testing. In the MAR20 dataset, the specific aircraft models referred to as A1–A20 are listed as shown in Table 5.

Table 5. Aircraft models A1–A20 in the MAR20 dataset.

A1	A2	A3	A4	A5	A6	A7	A8
SU-35	C-130	C-17	C-5	F-16	TU-160	E-3	B-52
A9	A10	A11	A12	A13	A14	A15	A16
P-3C	B-1B	E-8	TU-22	F-15	KC-135	F-22	FA-18
A17	A18	A19	A20				
TU-95	KC-1	SU-34	SU-24				

This test verified the algorithm through 20 different types of accuracy in the statistical data set and compared it with other classical algorithms as Table 6:

Table 6. The experimental data of this algorithm and other algorithms are compared in the MAR20 dataset [14].

Method	Our Paper	FCOS	ATSS	Faster R-CNN	TSD	Cascade R-CNN	Double-Head
One-Stage Network				Two-Stage Network			
A1	84.9	71.1	69.9	79.7	87.2	79.6	82.3
A2	96.2	79.4	83.7	81.6	86.7	81.4	81.2
A3	96.0	67.1	70.3	84.7	93.9	82.2	79.0
A4	94.5	46.4	50.2	72.0	71.8	66.7	63.4
A5	77.9	65.1	59.1	75.0	75.7	76.5	76.8
A6	95.9	84.8	85.1	88.0	93.6	87.9	87.9
A7	97.8	72.1	82.6	87.1	96.5	88.7	85.5
A8	96.0	83.3	80.0	89.7	91.3	86.9	83.6
A9	94.8	80.5	84.8	89.6	89.2	89.2	85.0
A10	98.0	84.5	90.0	90.6	98.1	90.7	90.5
A11	85.2	57.8	61.4	82.5	87.3	81.8	79.9
A12	91.1	71.3	82.7	86.4	87.8	85.1	85.4
A13	85.4	58.5	68.1	66.6	67.1	66.5	66.0
A14	95.7	75.3	78.1	85.8	89.6	78.6	84.6
A15	54.9	33.8	40.5	45.8	34.5	42.7	42.2
A16	94.7	87.0	87.8	91.0	91.0	88.2	88.4

Table 6. Cont.

Method	Our Paper	FCOS	ATSS	Faster R-CNN	TSD	Cascade R-CNN	Double-Head
One-Stage Network				Two-Stage Network			
A17	97.2	87.1	89.5	90.2	95.5	90.1	89.8
A18	82.3	38.6	36.1	54.2	58.5	42.6	44.1
A19	80.6	69.2	67.7	76.6	79.6	78.3	74.4
A20	86.5	66.4	77.5	78.6	80.1	76.7	77.2
mAP	89.3	68.8	72.2	79.6	82.7	78.0	77.4

Through the data analysis in the table, it can be seen that the proposed algorithm has a better effect than other classical algorithms at the fine-grained application level, for example, compared to the TSD (task-aware spatial disentanglement), which was closest to its accuracy algorithm, the mAP accuracy was improved by 6.6%.

Figure 10 is a selection of some obvious differences in the visual detection results.



Figure 10. Visualization results of the MAR20 dataset. The results from part of the test set of the Mar20 dataset are shown in (a–j).

In summary, through the visual experiments in Figure 10, we have not only verified the superiority of our algorithm but also provided thought-provoking insights for future fine-grained detection. There are some shortcomings in Figure 10f–j that require further improvement. The specific analysis is as follows:

In Figure 10, there is a phenomenon where A1 was incorrectly recognized as A19. However, considering the similar morphology of A1 and A19 themselves and the low

resolution of the image, this error can be explained. Nevertheless, it suggests room for improvement through a combination of super-resolution and recognition. Both Figure 10f,j exhibit cases of missed detection. Upon analyzing the specific reasons, it was found that the majority of missed models were A5. By comparing the dataset, it can be observed that the training data for A5 are relatively limited. This issue can be effectively addressed by expanding the database. Figure 10g,i have reasons for incorrect recognition, especially in the case of i, where traces of aircraft having stopped were recognized as aircraft. Apart from resolution, the contrast between the recognition target and its surroundings also plays a role. This provides valuable background support for future research.

However, it is worth noting that the cases listed from a to e demonstrate accurate recognition without omissions. This portion serves as evidence of the superiority of our algorithm in fine-grained detection. Moreover, even in the case of profiles of aircraft that are prone to erroneous recognition, like the one in Figure e, no significant interference with aircraft recognition occurred when there was an adequate amount of training data. Based on these findings, our algorithm shows promising practical applications in the field of engineering for fine-grained detection.

5. Conclusions

In this paper, we proposed a new network structure called DDU (Downsample Difference Upsample), which utilizes the methods of upsampling and downsampling, as well as complementary thinking, in the Neck layer of a deep learning network to enhance the recognition features of small targets and further improve the richness of recognition features. As a result, it addresses the issue of low accuracy in recognizing small target objects and effectively enhances the feature richness of recognition. Additionally, in order to address the recognition of objects of different sizes in image recognition, we introduced a new attention mechanism called PNOC (protecting the number of channels) to ensure that small target features and common objects were not lost during the channel integration process, thereby improving the recognition accuracy. We conducted experimental validation on the PASCAL-VOC dataset. Through experimental comparisons, we found that our accuracy surpassed other one-stage classical algorithms. Ablation experiments demonstrated positive improvements in recognition for both of the designed modules. Compared to the original YOLOv8 code, our approach achieved a 0.7% increase in accuracy. Furthermore, we applied our method to the fine-grained MAR20 dataset and observed superior performance compared to other classical algorithms, with a notable 6.6% improvement over the closely related TSD algorithm. Additionally, our proposed framework falls under the category of hierarchical detection methods, which exhibit excellent timeliness and demonstrate good adaptability and universality in scientific research applications.

This paper proposed a simple sampling algorithm to address the issue of easily lost small target features by comparing multilayer information discrepancies. It then focused on obtaining specific features for these features and utilized the attention mechanism to increase the weight of easily lost parts, thereby partially overcoming the problem of feature loss. The algorithm explores work that predecessors have not attempted in the feature extraction part while also leveraging existing prior knowledge. Benefiting from the accumulation of previous research achievements, the algorithm demonstrates feasibility and interpretability during the theoretical reasoning phase prior to experimental implementation. Through experiments comparing algorithms with the PASCAL-VOC dataset and MAR20 dataset, the theory and experiments have been correspondingly validated, thereby enhancing the interpretability of deep learning. It is evident that this paper exhibits a certain degree of innovativeness in the overall model.

Author Contributions: Methodology, Q.F. and H.L.; software, Q.F. and H.L.; validation, H.L.; data curation, H.L.; writing—original draft preparation, H.L.; writing—review and editing, Q.F., X.T. and W.D.; visualization, Q.F., H.L. and X.T.; supervision, Q.F. and W.D.; project administration, W.D.; funding acquisition, X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (62375260) and Jilin Province Innovation and Entrepreneurial Talent Project (2023QN17). This research was funded by Youth innovation promotion association cas (No. 2021221).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available at <https://doi.org/10.11834/jrs.20222139>, reference number [15].

Acknowledgments: We would like to thank the anonymous reviewers for their valuable suggestions, which have helped us to improve the manuscript greatly.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xiao, J.; Zhao, T.; Yao, Y.; Yu, Q.; Chen, Y. Context Augmentation and Feature Refinement Network for Tiny Object Detection. 2021. Available online: <https://paperswithcode.com/paper/context-augmentation-and-feature-refinement> (accessed on 23 June 2023).
2. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
3. Chen, Y.; Zhang, P.; Li, Z.; Li, Y.; Zhang, X.; Meng, G.; Xiang, S.; Sun, J.; Jia, J. Stitcher: Feedback-driven data provider for object detection. *arXiv* **2020**, 2, 12. arXiv:2004.12432.
4. Xin, H.; Chen, Z.; Wang, B. PCB electronic component defect detection method based on improved YOLOv4 algorithm. *J. Phys. Conf. Ser.* **2021**, 1827, 012167. [CrossRef]
5. Microsoft COCO: Common Objects in Context. Available online: https://github.com/pjreddie/darknet/tree/master/scripts/get_coco_dataset.sh (accessed on 12 July 2023).
6. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended feature pyramid network for small object detection. *IEEE Trans. Multimed.* **2021**, 24, 1968–1979. [CrossRef]
7. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
8. Feng, D.; Harakeh, A.; Waslander, S.L.; Dietmayer, K. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2021**, 23, 9961–9980. [CrossRef]
9. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7036–7045.
10. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
11. Kim, S.W.; Kook, H.K.; Sun, J.Y.; Kang, M.C.; Ko, S.J. Parallel feature pyramid network for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 234–250.
12. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
13. Kong, T.; Sun, F.; Tan, C.; Liu, H.; Huang, W. Deep feature pyramid reconfiguration for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 169–185.
14. Mar20. Available online: <https://gcheng-nwpu.github.io/> (accessed on 17 January 2023).
15. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
17. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, 172, 114602. [CrossRef]
18. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, 44, 1922–1933. [CrossRef] [PubMed]
19. Li, H.; Deng, L.; Yang, C.; Liu, J.; Gu, Z. Enhanced YOLO v3 tiny network for real-time ship detection from visual image. *IEEE Access* **2021**, 9, 16692–16706. [CrossRef]
20. Chen, H.; Zhou, G.; Jiang, H. Student Behavior Detection in the Classroom Based on Improved YOLOv8. *Sensors* **2023**, 23, 8385. [CrossRef]
21. Ma, N.; Su, Y.; Yang, L.; Li, Z.; Yan, H. Wheat Seed Detection and Counting Method Based on Improved YOLOv8 Model. *Sensors* **2024**, 24, 1654. [CrossRef] [PubMed]
22. Li, S.; Huang, H.; Meng, X.; Wang, M.; Li, Y.; Xie, L. A glove-wearing detection algorithm based on improved YOLOv8. *Sensors* **2023**, 23, 9906. [CrossRef] [PubMed]
23. Quan, Y.; Wang, P.; Wang, Y.; Jin, X. GUI-Based YOLOv8 License Plate Detection System Design. In Proceedings of the 2023 5th International Conference on Control and Robotics (ICCR), Tokyo, Japan, 23–25 November 2023; pp. 156–161.

24. Healey, C.; Enns, J. Attention and visual memory in visualization and computer graphics. *IEEE Trans. Vis. Comput. Graph.* **2011**, *18*, 1170–1188. [[CrossRef](#)] [[PubMed](#)]
25. Mushtaq, M.; Akram, M.U.; Alghamdi, N.S.; Fatima, J.; Masood, R.F. Localization and edge-based segmentation of lumbar spine vertebrae to identify the deformities using deep learning models. *Sensors* **2022**, *22*, 1547. [[CrossRef](#)] [[PubMed](#)]
26. Sha, G.; Wu, J.; Yu, B. Detection of spinal fracture lesions based on improved Yolov2. In Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 27–29 June 2020; pp. 235–238.
27. Yang, Z.; Shao, Y.; Wei, Y.; Li, J. Precision-Boosted Forest Fire Target Detection via Enhanced YOLOv8 Model. *Appl. Sci.* **2024**, *14*, 2413. [[CrossRef](#)]
28. Lin, S.; Hou, W. Efficient Sampling of Two-Stage Multi-Person Pose Estimation and Tracking from Spatiotemporal. *Appl. Sci.* **2024**, *14*, 2238. [[CrossRef](#)]
29. Arendt, B.; Schneider, M.; Mayer, W.; Walter, T. Environmental Influences on the Detection of Buried Objects with a Ground-Penetrating Radar. *Remote Sens.* **2024**, *16*, 1011. [[CrossRef](#)]
30. Pandey, S.; Chen, K.F.; Dam, E.B. Comprehensive multimodal segmentation in medical imaging: Combining yolov8 with sam and hq-sam models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 2592–2598.
31. Wang, Z.; Liu, Y.; Duan, S.; Pan, H. An efficient detection of non-standard miner behavior using improved YOLOv8. *Comput. Electr. Eng.* **2023**, *112*, 109021. [[CrossRef](#)]
32. Xue, L.; Yan, W.; Luo, P.; Zhang, X.; Chaikowska, T.; Liu, K.; Gao, W.; Yang, K. Detection and localization of hand fractures based on GA_Faster R-CNN. *Alex. Eng. J.* **2021**, *60*, 4555–4562. [[CrossRef](#)]
33. YOLO-ultralytics. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 22 January 2023).
34. Yuan, G.; Liu, G.; Wu, X.; Jiang, R. An Improved YOLOv5 for Skull Fracture Detection. In *Exploration of Novel Intelligent Optimization Algorithms, Proceedings of the ISICA 2021, Guangzhou, China, 20–21 November 2021*; Springer: Singapore, 2021; pp. 175–188.
35. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
36. PASCAL VOC. Available online: <http://host.robots.ox.ac.uk/pascal/VOC/voc2012/> (accessed on 23 January 2022).
37. Brahmabhatt, S.; Christensen, H.I.; Hays, J. StuffNet: Using ‘Stuff’ to improve object detection. In Proceedings of the 2017 IEEE Winter Conference on applications of computer vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 934–943.
38. Zhang, X.; Zhao, C.; Luo, H.; Zhao, W.; Zhong, S.; Tang, L.; Peng, J.; Fan, J. Automatic learning for object detection. *Neurocomputing* **2022**, *484*, 260–272. [[CrossRef](#)]
39. Wenqi, Y.; Hong, C.; Meijun, W.; Yanqing, Y.; Xingxing, X.; Xiwen, Y.; Junwei, H. MAR20: Remote Sensing Image Military Aircraft Target Identification Dataset. *J. Remote Sens.* **2022**, 1–11. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.