ITC 4/52

Information Technology and Control
Vol. 52 / No. 4 / 2023
pp. 1025-1044
DOI 10.5755/j01.itc.52.4.33460

Adaptive Clustering Object Detection Method for UAV Images Under Long-tailed Distributions

Received 2023/02/22 | Accepted after revision 2023/07/25

HOW TO CITE: Li, G., Wang, X., Li, Y., Li, Z. (2023). Adaptive Clustering Object Detection Method for UAV Images Under Long-tailed Distributions. *Information Technology and Control*, 52(4), 1025-1044. https://doi.org/10.5755/j01.itc.52.4.33460

Adaptive Clustering Object Detection Method for UAV Images Under Long-tailed Distributions

Guoxiang Li
Network and Information Technology Center, Guangxi University of Finance and Economics,
Guangxi Nanning, 530001, China

Xuejun Wang
School of sports economics and management, Guangxi University of Finance and Economics,
Guangxi Nanning 530001, China

Yun Li
School of Big Data and Artificial Intelligence, Guangxi University of Finance and Economics,
Guangxi Nanning 530001, China

Zhitian Li
College of Electronics and Imformation, Guangxi University for Nationalities, Guangxi Guilin 540001, China

Corresponding author: 47047792@qq.com

The target detection algorithm for common objects has achieved good results, but the detection accuracy and speed of the target detection algorithm for Unmanned Aerial Vehicle (UAV) need to be improved. Unmanned Aerial Vehicle (UAV) images are characterized by small targets, difficult to identify in the background image, clustering and sparse distribution of targets, etc. Many researchers have proposed the clustering target detection method (ClusDet) for Unmanned Aerial Vehicle (UAV) images. However, due to the large differences in target scales and uneven distribution of targets in UAV images, showing long-tailed distribution, the traditional ClusDet algorithm tends to truncate large and medium targets in the process of clustering; in the detection process, the fixed-threshold NMS method in the ClusDet algorithm is difficult to adaptively detect targets of different sizes, clustering and mutual occlusion. To address the above problems, this paper proposes an adaptive clustered target detection algorithm based on Unmanned Aerial Vehicle (UAV) images under long-tail distribution. The method is divided into three sub-networks: the adaptive clustering sub-network, which outputs several segmented images of small target clustering regions by extracting potential small target clustering regions in Unmanned Aerial Vehicle (UAV) aerial images; the segmentation and filling sub-network, which fills the images with disproportionate aspect ratio for the output of the adaptive clustering network to keep the size of the images within the reasonable range required by the detection network; and the detection sub-network, which detects the targets within the reasonable range required by the detection network by introducing attention mechanism, using variable threshold NMS, and training using sample balancing strategy effectively improve the detection accuracy of targets in the clustered region. Trained in VisDrone 2019 dataset, the simulation results show that the Unmanned Aerial Vehicle (UAV) image adaptive clustering target detection method based on long-tailed distribution has a large improvement in the detection accuracy of small targets, and can effectively improve the detection accuracy of the model for targets in the aggregation region, while the model has good generalization ability.

KEYWORDS: UAV aerial image; small object; adaptive clustering; NMS.

1. Introduction

UAVs gradually play an irreplaceable role in today's society, Unmanned Aerial Vehicle (UAV) images are more complex and richer because of the limitations of shooting equipment and environmental factors. In-depth mining of the potential information of UAV images is of great significance for the deeper application of UAVs in various fields of society.

The research on Unmanned Aerial Vehicle (UAV) image target detection has gradually been paid attention to by more and more scholars in recent years. In 2019, the ClusDet [30], [5] network proposed by Yang et al. is committed to solving the problem of detecting crowded areas of Unmanned Aerial Vehicle (UAV) images, which integrates 4 sub-networks, and uses iterative and congested area networks and scale estimation subnetworks to transform the size of each aggregation area into an appropriate range. The overall structure of the network is complex, requires high hardware resources, and the detection speed is slow. In order to better deploy the object detection algorithm on UAV-related embedded equipment, Zhang et al. proposed the Slim YOLOv3 [39] algorithm, which was tested on the VisDrone 2018 [10] object detection test set, and the detection accuracy of the algorithm was comparable to that of YOLOv3 [40] [23], the number of parameters of the network was reduced by 92%, and the detection speed was increased by two times. In 2021, Piciarelli et al. proposed an algorithm for real-time tracking and detection of multi-scale targets [2], and experimental results show that its performance reaches the most advanced algorithm performance. In 2022, Maskeliunas et al. proposed a Pareto-optimized deep learning algorithm for building detection and classification in a congested urban environment [18], which improves the detection accuracy under dynamically changing weather conditions as well as the influence of such ambient "noise". Yin et al. proposed deep learning-based compressed sensing (CS) algorithms [35], and the experimental results show the state-of-the-art performance while maintaining an efficient running speed [35]. Yi et al. proposed a composite transformer network for urban scene segmentation of UAV images, and experimental results shows that state-of-the-art results [34]. Aiming at the research on Unmanned Aerial Vehicle (UAV) image small target detection, the above algorithms are improved from clustering, clustering, detection speed and accuracy.

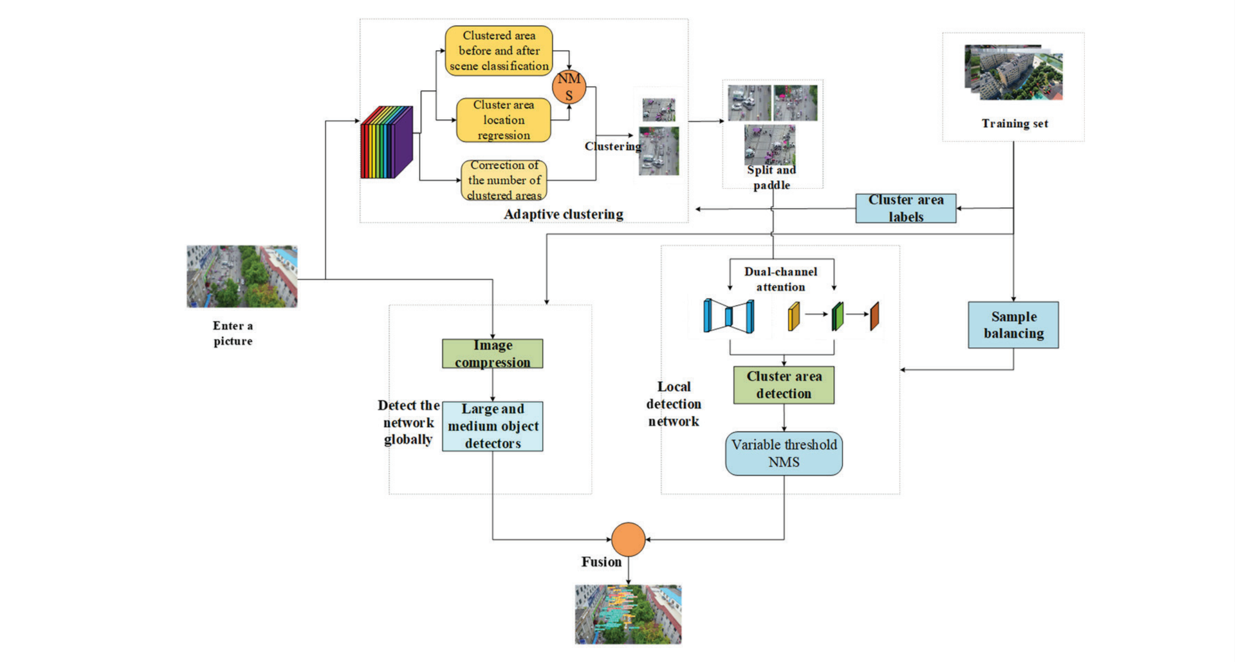
The traditional ClusDet algorithm can detect small targets in the process of clustering, but it is easy to truncate large and medium targets. In the detection process, the fixed-threshold NMS [20] method in the ClusDet algorithm is difficult to adaptively detect targets of different sizes, clusters, and mutual occlusion. However, the delay cannot guarantee the real-time detection Therefore, this paper proposes an adaptive clustering object detection [20] method for Unmanned Aerial Vehicle (UAV) images under long-tailed distributions (ACOD-LTD). The method is divided into three parts: adaptive clustering subnetwork, segmentation and population subnetwork, local detection, and global detection network. The adaptive clustering sub-network divides the target area of Unmanned Aerial Vehicle (UAV) picture clustering, and divides the original image according to the division results to obtain several target clustered pictures. The segmentation and filling sub-network correct the size of the split image to adapt to the input standard of the later detection network. The local detection network and the global object detection network detect the objects of the segmented picture and the original image, and finally use the fusion algorithm to fuse the detection results.

2. System Framework

The adaptive clustering target detection algorithm of Unmanned Aerial Vehicle (UAV) images based on long-tail distribution is divided into three stages, as shown in Figure 1: (1) The adaptive clustering sub-network extracts potential small target aggregation areas in Unmanned Aerial Vehicle (UAV) aerial images through operations such as feature extraction, front and rear scene classification, position regression and quantity correction, and then segments the aggregation area. (2) The segmentation and filling sub-network divide the oversized segmented pictures again, and fills the out-of-proportion pictures; The purpose of this operation is to keep the image size within a reasonable range required by the detection network. (3) The detection subnetwork is divided into local detection network and global detection network. Among them, the local detection network detects the pictures output by the segmentation and filling sub-network, improves the detection ability of the model for small targets by introducing the channel attention mechanism and the spatial attention mechanism, improves the recall of the detection end for the detection of objects of various scales by using the variable threshold NMS algorithm proposed in this paper, and reduces the influence of the long-tail distribution on the model accuracy due to the existence of the data set by training the network with the sample balancing strategy. The global detection network mainly detects two types of targets: the first type is the target truncated by the adaptive clustering subnetwork when the picture is segmented; The second category is the target whose distribution is sparse in the original map and has not been extracted by the adaptive clustering

Figure 1

Adaptive clustering target detection algorithm for Unmanned Aerial Vehicle (UAV) images based on long-tail distribution



subnetwork. The detection results of the global detection network complement the detection results of the local detection network, and the results of the two are fused to obtain the final prediction results.

3.1. Adaptive Clustering Network

The aerial image dataset has the following characteristics: the size of the aerial image dataset is large, the pro1portion of the labeled targets in the dataset occupies no more than half of the area of the picture, and most of the targets in the image are distributed in the form of clusters. According to the characteristics of the dataset, the adaptive clustering target detection algorithm in this paper uses an adaptive clustering network to segment the aggregation area, and the network performs regression prediction on the potential area of the clustered target in the image after training and optimization, and then divides these regions, and then sends the segmented picture to the subsequent detection network for detection. The purpose of clustering the image is to refine the detection of the target aggregation area, so as to improve the target detection accuracy and detection efficiency.

The structure of the adaptive clustering network is shown in Figure 2: feature extraction is performed on the input pictures, the above branch network generates the suggested clustering area, and the suggested clustering area is generated on the feature map through the dichotomous and cluster location regression of the front and back scenes; The following branch network predicts the number of clustered areas and determines the final segmentation based on the output.

3.2. Feature Extraction

The feature extraction network uses the improved DetNet59 [4], [9], [31], [33] as the backbone network, which combines the spatial pyramid structure (FPN) [14], [36], [15], and channel attention mechanism [17], [41], [6]. The network structure is shown in Figure 3, and the C4, C5, and C6 layers on the left part of the figure do not use down sampling operations, the purpose is: to keep the feature map "16x" high resolution, the high-resolution feature map can retain more small targets, which is conducive to the regression of the target position. However, the receptive field of the image is reduced by not using down sampling, so the network introduces hole convolution to expand the receptive field of the feature map, in addition, in order to reduce the amount of network parameters, the number of channels of the output feature map of each layer of the network is fixed at 256. Then, the feature map output of different layers is fused by using the feature pyramid network [16], [26], [8], and the P2 and P3 layer feature maps retain more large target features, and the P4, P5 and P6 layer feature maps retain more small target features. Finally, after each layer of the feature pyramid output, the channel attention network is cascaded, and all the feature maps are output in parallel.

"C1" in Figure 3 represents the output of the first convolutional block, "4x" in the figure represents a quarter of the original image resolution of the output resolution of the feature map, "P2" represents the result of the fusion of "C2" with the previous feature map, and so on.

Figure 2

Adaptive clustering network framework diagram

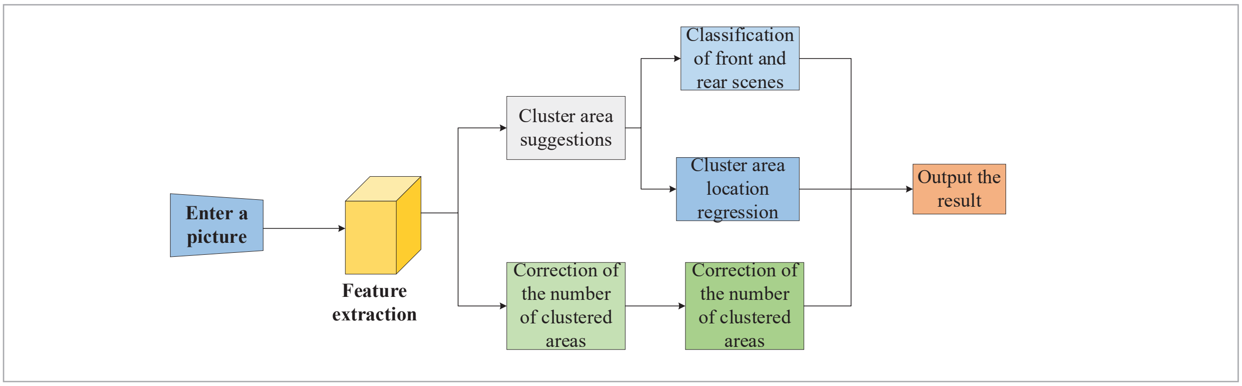
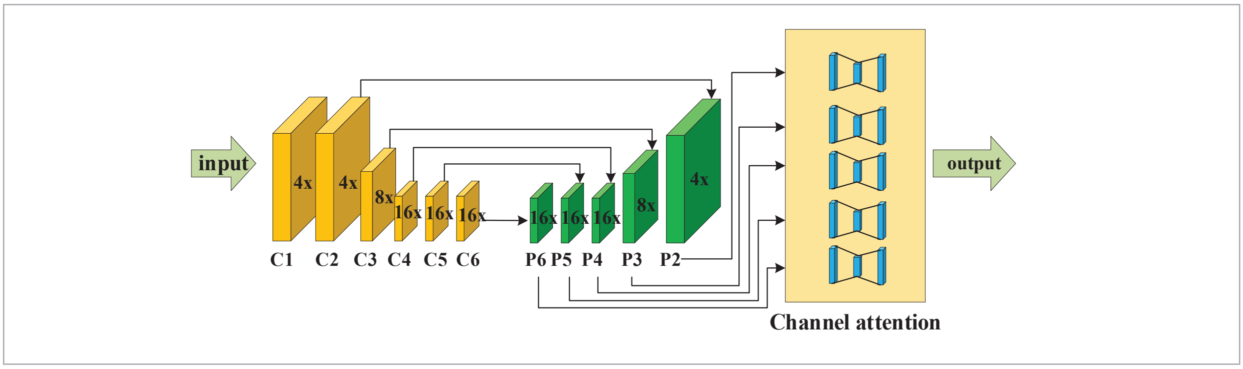

Enter a picture
Feature extraction
Cluster area suggestions
Classification of front and rear scenes
Cluster area location regression
Correction of the number of clustered areas
Correction of the number of clustered areas
Output the result

Figure 3

Schematic diagram of feature extraction network structure

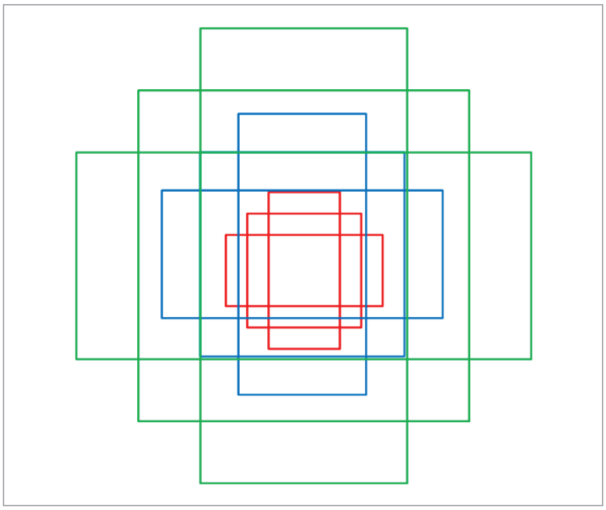


3.3. Cluster Area Suggested Networks

The clustered region suggestion network draws on the idea of RPN [3], [48] in Faster R-CNN [22], [37], [27], the object of RPN network regression and classification is each target, and the clustered region suggests that the object of network regression classification is each clustered region. According to the feature map of 256 M×N output in the feature extraction stage, the 3×3 convolution kernel with an expansion rate of two was used to further expand the receptive field of the feature map, and the output remained at 256×M×N. Think of these feature maps as M×N 256-dimensional vectors, each vector corresponds to nine candidate regions in the original figure, as shown in Figure 4, nine regions are generated by the specified three sizes in a total of 1:1, 1:2 and 2:1 in three ratios, so there are a total of M×N × 9 regions. Binary classification and regression operations were performed on these areas to obtain prediction information for clustered areas. The specific operation is: two branches of the feature map of 256 M×N, one branch to classify the area twice, the purpose of the second classification is: to separate the foreground and background, use the 1×1 convolution kernel to output two confidence scores, a total of 9×2×M×N data; The other branch regresses the clustered regions and uses a 1×1 convolution kernel to obtain 9×4×M×N data. Finally, the regions predicted as foreground are screened out, and the non-maximum suppression algorithm (NMS) is used to fuse these regions to obtain output results.

Figure 4

Nine zones generated by the network



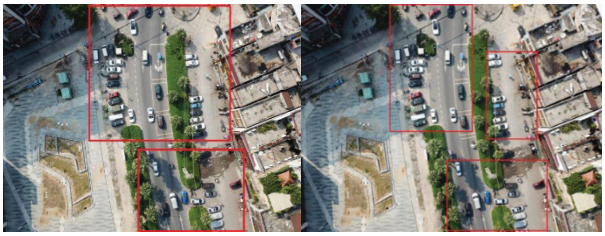
3.4. Correction of the Number of Clustered Areas

Cluster area suggestion network will generate several prediction areas, if according to the recommended cluster area segmented pictures too much, will increase the subsequent detection network detection time, waste more resources, in order to solve this problem, the algorithm proposed to use the addition of cluster area correction network method, this network and cluster area recommended network share feature map of the feature extraction stage, used to correct the number of cluster areas, so that the number of cluster areas to keep within a suitable range.

In the same drone picture, it is not easy to determine the number of clustering areas, taking the artificial labeling dataset as an example, different people use different discrimination criteria to label the clustering area, the number of labeling and the area are therefore different, as shown in Figure 5, the picture on the left is labeled as two clustering areas, the picture on the right is labeled as three clustering areas, the two labeling methods have no advantages and disadvantages, the labeled clustering areas are more reasonable, but if the number of labeled clustering areas is ten, Then this labeling is obviously unreasonable. Therefore, the number of clustered areas per image is around a fixed value.

Figure 5

Comparison diagram of different clustering region division methods



In order to avoid the bias caused by manual labeling, this paper uses an algorithm to label the clustered areas, and the specific process is as follows:

1. Set the threshold value N, which is used to roughly determine whether there are clustered areas in the image. N is a hyperparameter.
2. Select an image, get the number of targets in the image, skip the number of images less than N, and do not do any processing on this image.
3. The mean shift clustering algorithm (mean shift) was used to classify the image targets, and the location coordinates of the cluster area corresponding to the picture and the number of cluster areas were labeled.
4. Iterate through all the training set pictures to get the cluster area labeling files corresponding to all the pictures.

After completing the above procedure, you can get a cluster area labeling file.

The use of algorithm labeling can avoid the error caused by manual labeling, but the problem of algorithm labeling is that when there are individual sparsely distributed targets in the picture, the location of such targets is far away from other clustering areas, and the center of the clustered area generated by the algorithm will also be seriously deviated.

Based on the above analysis, this paper proposes that the clustering area correction network is used to correct the error caused by the algorithm labeling, and the network completes the determination of the number of clustered areas by implementing a classification task. The specific process of network training is as follows: according to the labeling file to obtain the annotation information of the picture, the number of clustering areas in the picture is N, according to N to construct a probability distribution obey the one-dimensional Gaussian distribution of M-dimensional vector, M is the maximum number of clustered areas, M is the empirical value, this paper by the VisDrone 2019 dataset[7] statistics and calculation, take M=10. The Gaussian function is expressed as follows:

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-\mu)^2}{2\sigma^2}). \qquad (1)$$

In Equation (1), take $\mu=N$, $\sigma=1$, sample at the integer position of the function, and construct a probability distribution based on the sampled value. For example, when $N=3$, that is, the number of clustering areas is 3, the Gaussian function constructed is shown in Figure 6, the value is taken at the integer position of the Gaussian function, and the probability distribution after quantization is:

$\boldsymbol{p}$=[0.054 0.24 0.40 0.24 0.054 4.4e-3 1.3e-4 1.5e-6 6.1e-9 9.1e-12]

Figure 6

Gaussian graph

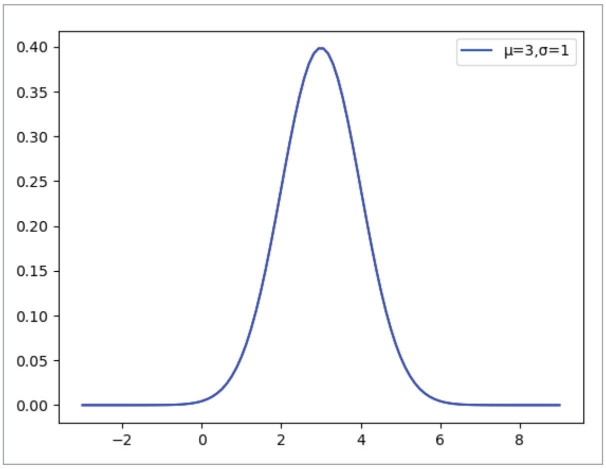
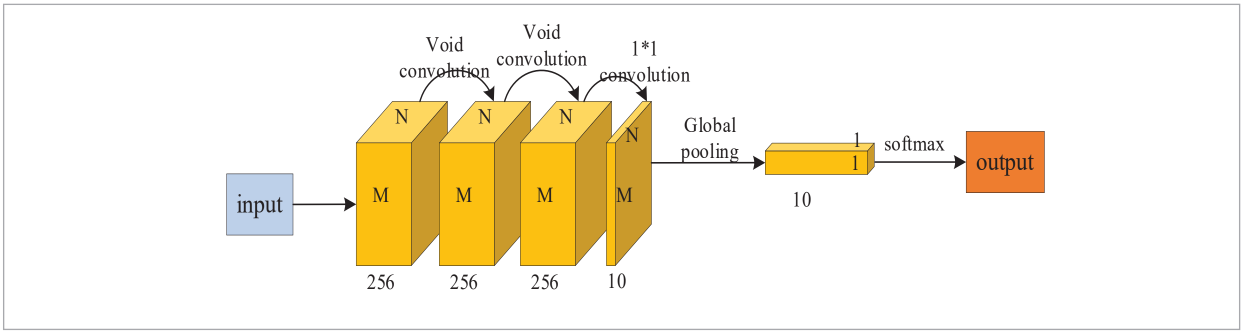


Figure 7

Cluster area modification network structure diagram



Use the softmax function [10] to normalize this probability distribution so that all values add up to 1 and normalize to obtain the following vector:

$\boldsymbol{p}_{true}$=[0.095 0.114 0.134 0.114 0.095 0.090 0.090 0.090 0.090 0.090]

This vector is used as the true label of the modified network for the number of training clustering areas, and the cross-entropy function [11] is used as the loss function during training.

The structure diagram of the cluster area number correction network is shown in Figure 7, and the feature map with an ×expansion rate of three is used for the feature map with an expansion rate of three, followed by a 1×1 convolution kernel for dimensionality reduction to obtain 10 M×N feature maps, then using global maximum pooling, and finally connecting the fully connected layer, and using the softmax function for output. The number of clusters can be determined by correcting the output of the network according to the number of clustering areas, and the output of the adaptive clustering network can be determined by combining the output of the clustering area suggestion network.

3.5. Split and Paddle

There are two problems with images segmented using adaptive clustering networks: First, the image size is too large. For the detection network, the input size of the picture is still difficult to detect small targets; Second, the length and width ratio of the picture is out of balance, and the detection accuracy of the picture will be reduced after it is input into the detection network. To avoid detection differences caused by such extreme inputs, this paper uses a scale estimation strategy to process the size of the segmented image.

The scale estimation strategy is divided into the following steps:

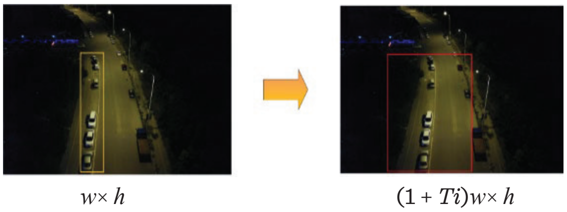
1. Determine the threshold Pm and Ps according to the input scale required by the detector, Pm represents the maximum value of the input size, and Ps represents the minimum value of the input size. where Ps and Pm are hyperparameters.
2. Suppose the length and width of the input image are W and H, respectively. When the long side size w of the picture is greater than Pm, the picture is evenly divided; When the short side size h of the picture is less than Ps, the filling ratio needs to be calculated, and the calculation formula is as follows:

$$Ti = \frac{w}{Pm}, \tag{2}$$

where Ti is the fill ratio, that is, the filling is carried out on the basis of the split picture, and the size of the filled picture becomes (1+Ti) times of the original picture, and Figure 8 shows the comparison before and after filling. (3) Iterate through all the segmented pictures, perform operation (2), and get the processed picture.

Figure 8

Size comparison of clustering regions before and after filling



3.6. Local Detection and Global Detection of Networks

3.6.1. Local Detection of Network Structure

The structure of the local detection network is shown in Figure 9. First, the clustered image input network is used for feature extraction, the extracted feature map is input to the channel attention and spatial attention network, respectively, and then the fused feature map is input to the detection network using FPN, and the variable threshold NMS is used at the output of the detection network for the fusion of the detection results. Among them, the detector can be any object detection network, and the detector used in this paper is a Faster R-CNN-based detection network, and the detection network uses a sample balancing strategy to enhance the data of the training set before training.

3.6.2. Backbone Network Based on Dual-Channel Attention Mechanism

In this paper, attention mechanism [1], [24] is introduced in the detection network, in which the channel attention mechanism mainly focuses on the importance of different channel feature maps. The spatial attention mechanism is to compress the channels of the feature map, and pay more attention to the importance of different spatial regions of the feature map. In this paper, considering that it is extremely difficult to detect small targets in clustered areas, a dual-channel attention backbone network is formed by introducing channel attention mechanism and spatial attention mechanism, which reduces the diffusion of small target features on the feature map, so as to improve the detection accuracy of small targets.

The feature extraction network of the detection network still uses the improved DetNet59, which connects the dual-channel attention network after the feature extraction network and then cascades the FPN network output. The following mainly introduces the implementation process of channel attention module and spatial attention module.

The main integrated operations of the channel attention mechanism are, global maximum pooling [13] and global average pooling [38] of the feature maps of all channels, then using the compression-activation module, and finally using the sigmoid activation function for output, and the network structure of the channel attention mechanism is shown in Figure 10. The specific operation is divided into the following steps:

1 Assuming that the dimension of the input feature map is {H,W,C}, where H represents the height of the feature map, W represents the width of the feature map, and C represents the number of channels of the feature map, after the operation of two branches of global maximum pooling and global average pooling, the feature map of each channel is compressed into 1, that is, the output feature map dimension is {1,1,C};

2 The output feature map is connected to the fully connected layer as input, the two feature maps share the parameters of the fully connected layer,

Figure 9

Local detection network structure diagram

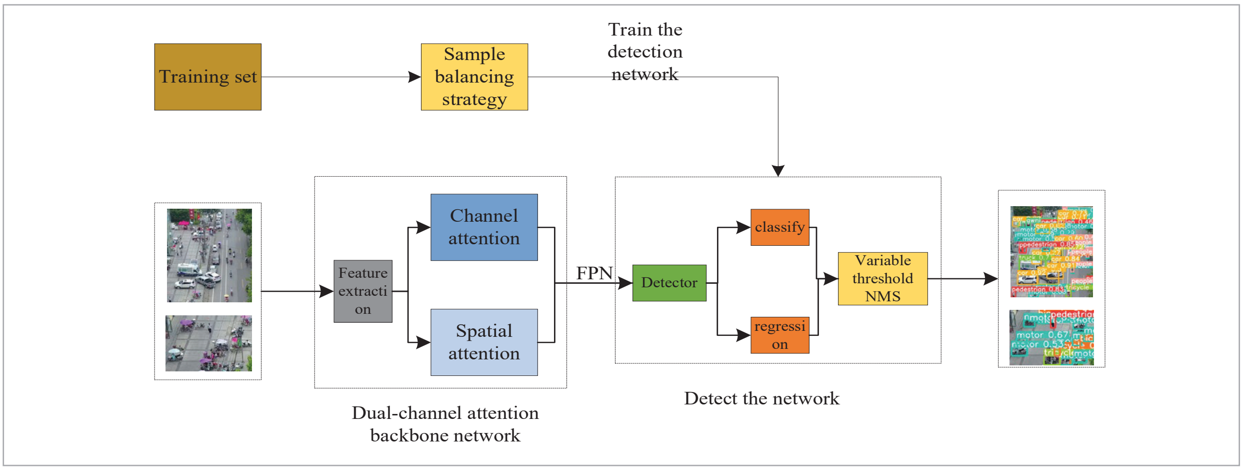
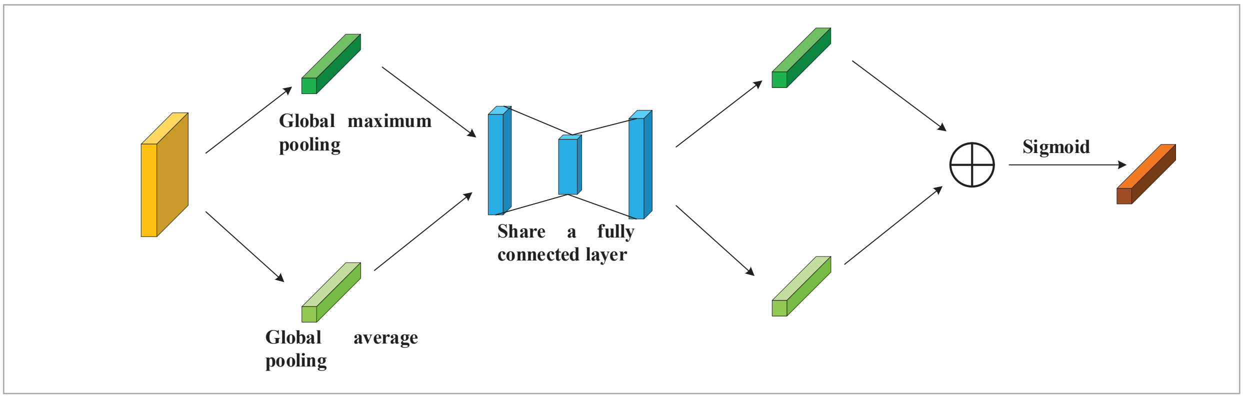


Figure 10

Channel attention



and the compression and activation methods are used in the fully connected layer to reduce the amount of parameters, set the compression ratio to r, and the dimension of the feature map output by the operation is {1,1,C/r};

3 After passing through the fully connected layer, the output feature map has obtained the importance of different channels, and then the two branches are added and fused to obtain a feature map, and finally the sigmoid activation function [19] is used to obtain the output.

The spatial attention network [42] structure is shown in Figure 10, and the dimensions of the feature map are still {H,W,C}. The spatial attention module uses global maximum pooling and global average pooling to extract features from the spatial region of the feature map, outputs two feature maps, then uses convolution operation to fuse the two feature maps, and finally uses the sigmoid activation function to output the feature map. The specific operation is as follows:

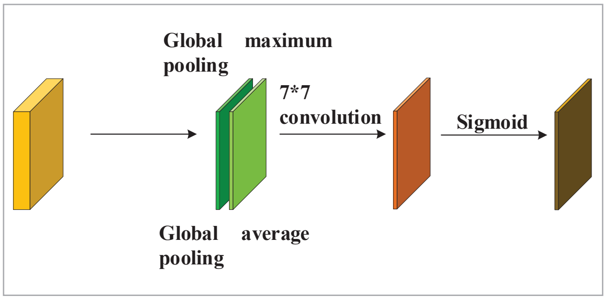
1 The input feature map is regarded as H×W C-dimensional feature vectors, and after the global maximum pooling and global average pooling of these H×W feature vectors, two feature maps with dimensions {H,W,1} are obtained;

2 Splice the two feature maps to obtain a feature map of dimension {H,W,2}, use a convolution kernel of 7×7 to convolve this feature map, and then obtain a feature map with dimension {H,W,1};

3 Finally, the sigmoid activation function is used to output the feature map, which has obtained the information of the importance of different spatial regions.

Figure 11

Spatial attention



3.6.3. NMS Method with Variable Threshold

The non-maximum suppression (NMS) algorithm and the NMS-based improved algorithm [21] are used in the final stage of the detection process, when all the prediction bounding boxes have completed regression, the NMS algorithm is used to fuse each type of target in the picture, in order to prevent multiple prediction boxes from appearing on the same target. The basic steps of the NMS algorithm are as follows:

1 Divide all the prediction bounding boxes into different sets according to categories, and the prediction bounding boxes in each set are arranged in descending order according to the score;

2 Select the prediction bounding box with the largest score from the set, so as to calculate the IoU [45], [29] of the bounding box and other prediction bounding boxes in the set, if the IoU is greater than the set threshold TH, the prediction bounding box with the smaller score is eliminated, and finally the bounding box is retained;

3 Repeat the (2) operation for the remaining prediction bounding boxes in the set until all the prediction boxes in the set have finished filtering;

4 Repeat the operations (2) and (3) for each collection until all collections have completed filtering.

The above is the standard NMS algorithm, and the expression of the suppression function is as follows:

$$S_i = \begin{cases} S_i, & IoU(M, b_i) < TH \\ 0, & IoU(M, b_i) \geq TH \end{cases}, \quad (3)$$

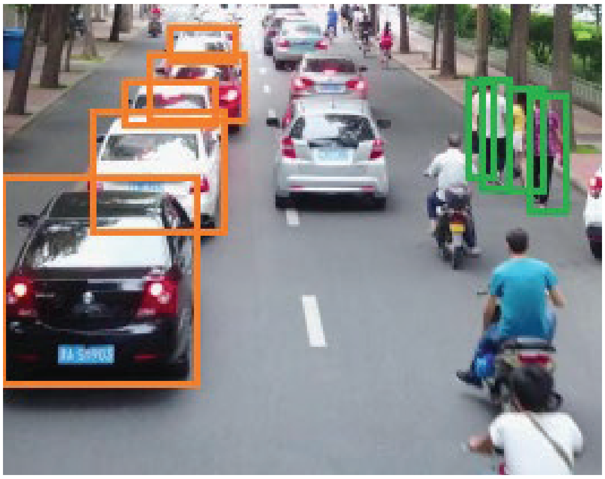
where M represents the prediction bounding box with the largest current score, bi represents the other prediction bounding box, Si represents the score of the other prediction bounding box, and TH is the set threshold.

NMS is used to detect images with sparse target distribution. However, when the distance between multiple targets in the picture is small, the NMS algorithm will mistake the prediction bounding boxes of different targets for the prediction bounding boxes of the same target, thereby removing the prediction bounding box, and the recall of the detection results will be reduced. In the Unmanned Aerial Vehicle (UAV) image, there are a large number of target gathering areas, as shown in Figure 12, the cars in the picture are large targets, although there is a shielding problem between each other, but they are easy to detect targets; The pedestrians on the right are small targets, and there are also occlusion problems between them, such targets are difficult to detect, and the standard NMS algorithm can no longer solve such problems well. In order to solve the above problems, on the basis of the NMS algorithm, this paper proposes a variable threshold NMS algorithm, and the calculation method of the variable threshold is as follows:

$$TH = TH_{low} + \frac{TH_{high} - TH_{low}}{e^{U-I}}, \quad (4)$$

where TH_{low} and TH_{high} are the minimum and maximum thresholds set, and you and I are the union and intersection areas of the current prediction bounding box and other prediction bounding boxes, respectively. When the value of (U-I) is large, it means that the current prediction bounding box predicts a large target, and the large target is an easy target to detect for the detector, and the calculated TH value is small, and when using the NMS algorithm, the current prediction bounding box plays a less inhibiting effect on the fusion of other prediction bounding boxes. When the value of (U-I) is small, it means that the current prediction bounding box predicts a small target, the small target is difficult to detect the target with the detector, and the calculated TH value is large, and when using the NMS algorithm, the current prediction bounding box plays a greater role in inhibiting the fusion of other prediction bounding boxes.

Figure 12
Convergence of objectives of different scales by NMS



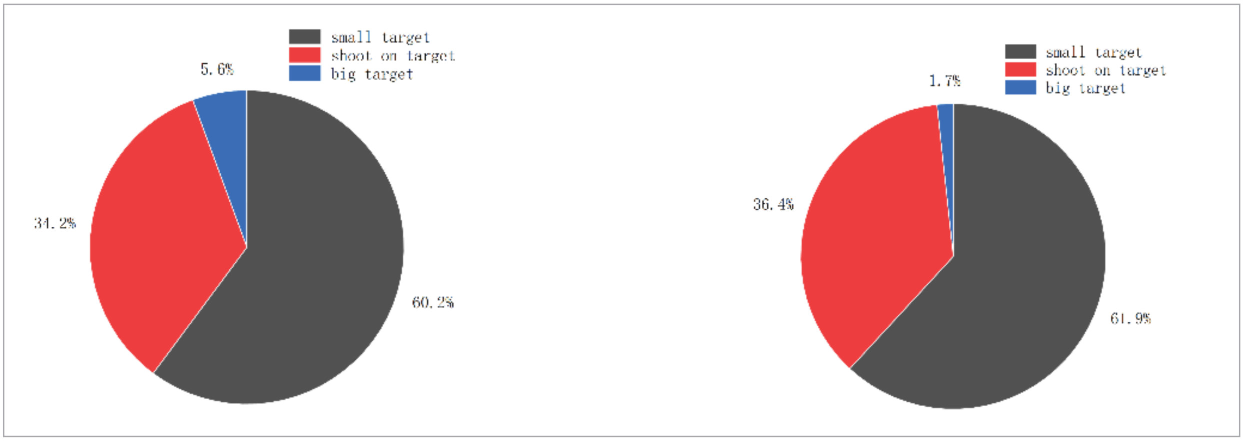
The NMS algorithm with variable threshold can adaptively change the threshold according to the target size, effectively reducing the missed detection rate of small targets by the detector, thereby improving the recall rate and detection accuracy of the detection results.

3.7. Detect the Network Globall

After the adaptive clustering network clusters the original map, it detects the clustered area, but in the process of clustering, large and medium targets have a high probability of being truncated by the clustering network, and these truncated targets become difficult to detect. In addition, there are still sparsely distributed targets in the original picture, and large and medium targets still account for a large proportion, as shown in Figure 13 for the proportion of large, medium and small targets in the VisDrone 2019 training set and UAVDT dataset, among which, the division criteria of large, medium and small targets refer to the

Figure 13

The proportion of VisDrone 2019 and UAVDT targets in different scales



standards of COCO data, that is, small targets with dimensions less than 32×32, large targets with sizes greater than 96×96, and medium targets with sizes in between. In order to solve the above problems, this paper trains a global detection network for large and medium targets to detect large and medium targets that may be truncated.

The global detection network can use any classical network, this paper chooses to use Faster R-CNN as the global detection network, and the backbone network uses the DetNet59 improved in this paper. This network only detects medium and large targets, so small targets in the dataset are ignored during the training process, and only training is conducted for large and large targets.

4. Experiment

4.1. Data Set

To verify the effectiveness of the method, three different aerial image datasets were used for testing: VisDrone 2019, DOTA, and UAVDT.

The VisDrone 2019 dataset was collected by the AISKYEYE team at Tianjin University's machine learning and data mining laboratory, and the entire benchmark dataset was captured by drones, including 288 video clips, including a total of 261908 frames, and 10,209 still images.

The DOTA dataset was proposed by Wuhan University in 2017, and the images were taken by Google Earth, JL-1 satellites, and GF-2 satellites of the China Resources Satellite Data and Application Center. The dataset includes 15 types of targets such as track and field, football field, plane, boat, swimming pool, etc., with a total of 2806 pictures.

The UAVDT dataset is a UAV image dataset, including target tracking and target detection data, with a total of 40,735 images in the target detection dataset, including cars, buses and trucks

The main annotated targets of the VisDrone 2019 dataset are: common transportation vehicles and people, including pedestrians, people, bicycles, cars, vans, buses, tricycles, covered tricycles, motorcycles, and trucks. There are a total of 10 types of targets, among which only standing and walking people are labeled as "pedestrians", while others are labeled as "people". The target size in a data aggregation is often smaller than that of a regular image, and most of the targets are distributed in an aggregated form. As shown in Figure 1, the images from the VisDrone 2019 target detection training set are 960 in size × 540、1400 × 1050, the shooting scene of the picture is near the mall; The shopping mall and forest in the left image are not the targets for dataset annotation, but they occupy half of the image area; In the right figure, only a few cars distributed in the middle of the figure are the targets of the dataset annotation, and their occupied area is only about a quarter of the entire image.

The DOTA data set is an aerial image data set captured by satellites, including 2806 pictures, each of

Figure1

VisDrone 2019 Target detection training set image

which occupies a large pixel size. Some of the data in the data set is captured by satellites, and the other part is provided by Google Earth. The data shooting angle is overhead above the target, and the shooting height is high. The dataset annotates 15 types of targets, including small cars, large cars, ships, helicopters, ports, etc. The target size varies, with the smallest target occupying less than 10 pixels and the largest target occupying several hundred thousand pixels. As shown in Figure 2, the images in the DOTA training set are 1143 in size × 1235、1552 × 1951。 The main annotation targets in Figure 2 (left) are small and large cars, arranged neatly and concentrated in a certain area, with the target area accounting for no more than half of the entire image area; The main annotation target in Figure 2 (right) is the ship, which is mainly distributed in the bottom left corner of the image, and its area does not exceed a quarter of the image area.

Figure 2

DOTA Training Set

The UAVDT dataset is a dataset mainly used for target tracking, which forms a target detection dataset by capturing images of video frames. The dataset is not divided into training, validation, and testing sets. The UAVDT target detection dataset annotates three types of targets: automobiles, buses, and trucks. As shown in Figure 3, the images in the UAVDT dataset

Figure 3

UAVDT Data Set

are both 1024 in size × 540. The majority of the cars in the left image are distributed in the center of the road, accounting for about half of the image area; Some of the cars in the right image are concentrated on both sides of the road, while the other part is concentrated in the parking lot directly above the image.

4.2. The Sample Balancing Policy Processes the Dataset

After analyzing the three datasets, the following conclusions can be drawn: first, the scales of the various targets of the datasets vary greatly; Second, the distribution of various types of targets in the data is long-tail distribution, that is, the proportion of various target areas in the training set varies greatly. In this paper, a sample balancing strategy is proposed to amplify the area and quantity proportion of small sample data, so as to reduce the influence of long tail [43], [12] distribution on the accuracy of the detection model.

The specific steps of the sample balancing strategy are as follows:

1 The proportion of the number and area of various targets in the statistical data collection;

2 The categories that are smaller than NC and SC, respectively, in the proportion of quantity and dough machine are identified as categories that need to be expanded. NC and SC are hyperparameters;

3 Select a picture, and skip the picture when the total number of targets in the image that needs to be amplified is less than N; When the total number of targets is greater than or equal to N, the mean drift clustering algorithm is used to classify the targets that need to be amplified. N is a hyperparameter;

4 Take the boundary target contained in each category as the boundary, frame this area, stretch this area to the original map area size equally, and generate a new labeling file, in which the truncated target area accounts for more than 50% of the original target area, it is marked as a positive sample, otherwise the target is not marked;

Repeat step (3) until all the images in the dataset have been traversed, and finally the expanded training set is obtained.

4.3. Experimental Content

In order to verify the effectiveness of the adaptive clustering network, this paper uses different image segmentation methods to test on the VisDrone 2019 dataset, and analyzes the test results [28]. To verify the effectiveness of the detection network, this paper uses Faster RCNN as the baseline detection network, and in addition to testing on the VisDrone dataset, this paper also uses the DOTA and UAVDT datasets for training and testing. To verify the effectiveness and robustness of the sample balancing strategy, this paper uses a trained YOLOv5 network to test on the VisDrone 2019 dataset [32].

4.4. Experiment Setup

The computer hardware environment uses Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz × 2 processor, RTX 2080Ti (11 G) × 2 graphics card, 64GB memory; The software environment adopts version 1.8.1 Pytorch, 3.8.12 version Python, 3.0 version Anoconda3 and 10.1 version CUDA; the drone used for shooting is DJI Mavic 2, with 20 million image pixels.

4.5. Experimental Results and Analysis

4.5.1. VisDrone Test Set Test Results

Table 1 shows the experimental results of the model in the VisDrone test set, where the master Dry networks represent the network structure used for feature extraction; The "O", "C" and "OC" of the segmentation mode represent the original figure [47], uniform segmentation, and cluster segmentation, respectively; APs, APm, and APl indicate the detection accuracy of small, medium, and large targets, respectively. In order to verify the detection effect of the clustering detection network, different segmentation methods were set up for comparative verification, in which the segmentation method was the original image, indicating that 548 pictures in the test set were used without any processing. The segmentation method is uniform segmentation, which means that all pictures in the test set are evenly divided into 6 equal parts; When ClusDet clusters the pictures, the number of clusters is fixed as a hyperparameter, and the hyperparameters have been determined before training the network, and the number of clusters of the adaptive clustering network proposed in this paper is output by the network adaptively [44].

Table 1 shows that the use of uniform segmentation can improve the detection accuracy of small targets, but the segmentation process will cause large targets to be truncated, greatly reduce the detection accuracy of large targets, and reduce the recall rate. ClusDet algorithm uses the combination of cluster detection and original image detection, and the clustering method further improves the detection accuracy of small targets, while the original image detection ensures the detection accuracy of large targets, and the average detection accuracy of the method is greatly improved. In this paper, the adaptive clustering network is used to cluster the images, and the improved DetNet59 network and CBAM [25] network are used as feature

Table 1

Test results of the model at VisDrone 2019

method	Backbone network	Split method	Number of images	mAP	AP50	AP75	APs	APm	APl
FRCN+FPN	ResNet50 mistake! Reference source not found	o	548	21.4	40.7	19.9	11.7	33.9	54.7
FRCN+FPN	ResNet101	o	548	21.4	40.7	20.3	11.6	33.9	54.9
FRCN+FPN+EIP	ResNet50	c	3,288	21.1	44.0	18.1	14.4	30.9	30.0
FRCN+FPN+EIP	ResNet101	c	3,288	23.5	46.1	21.1	17.1	33.9	29.1
ClusDet	ResNet50	o+ca	2,716	26.7	50.6	24.7	17.6	38.9	51.4
ClusDet	ResNet101	o+ca	2,716	26.7	50.4	25.2	17.2	39.3	54.9
Adaptive clustering detection method +FRCNN	Improved DetNet59	o+ca	2,965	31.4	54.5	27.3	21.7	40.7	53.6

extraction networks, and the detection network still uses Faster RCNN. The detection results show that the detection effect of large and medium targets is not much different between the proposed algorithm and ClusDet algorithm. For the detection effect of small targets, the detection algorithm in this paper is greatly improved than that of ClusDet algorithm.

4.5.2. DOTA Dataset Test Results

Table 2 shows the detection results of different methods in the DOTA dataset. The data shows that the detection accuracy of medium targets using the uniform segmentation method is higher than that of the clustering method, the reason is: the images in the DOTA dataset are taken by satellites, the image size is large, the area and number of medium targets account for the highest proportion, and the proportion of medium targets in evenly segmented pictures is much higher than that of small targets and large targets, and the recall rate of medium target detection results will also be improved. In general, the average accuracy of the adaptive clustering detection algorithm is not much different from ClusDet, and the detection accuracy of small targets is improved, while the detection accuracy of large and medium targets is reduced.

4.5.3. UAVDT Dataset Test Results

Table 3 shows the detection results of different methods in the UAVDT dataset. The data in the table show that the detection accuracy of the adaptive clustering algorithm for large, medium and small targets is higher than that of ClusDet, and its detection effect is better than that of other methods.

Based on the above experimental results, the adaptive clustering detection algorithm proposed in this paper has a good detection effect on UAV images, greatly improves the effect of detecting small targets in UAV images, and improves the detection accuracy to a certain extent.

4.5.4. Ablation Experiment

Through ablation experiment results, we can prove the effectiveness of the model.

4.5.5.Related Experiments on Variable Threshold NMS

The YOLOv5 algorithm was used as the detection network, and the standard NMS, soft NMS, and variable threshold NMS algorithms were used for testing. As shown in Figures 4-6, the test results using standard NMS, soft NMS, and variable threshold NMS, re-

Table 2

Test results of the model in DOTA dataset

method	Backbone network	Number of images	AP	AP50	AP75	APs	APm	APl
FRCNN+FPN+EIP	ResNet50	2,838	31.0	50.7	32.9	16.2	37.9	37.2
FRCNN+FPN+EIP	ResNet101	2,838	31.5	50.4	36.6	16.0	38.5	38.1
ClusDet	ResNet50	1,055	32.2	47.6	39.2	16.6	32.0	50.0
ClusDet	ResNet101	1,055	31.6	47.8	38.2	15.9	31.7	49.3
Adaptive clustering detection method +FRCNN	Improved DetNet59	1,726	31.4	47.1	39.6	17.3	33.6	47.7

Table 3

The model was tested on the UAVDT dataset

method	Backbone network	Number of images	AP	AP50	AP75	APs	APm	APl
FRCNN+FPN	ResNet50	15,069	11.0	23.4	8.4	8.1	20.2	26.5
FRCNN+FPN+EIP	ResNet50	60,276	6.6	16.8	3.4	5.2	13.0	17.2
ClusDet	ResNet50	25,427	13.7	26.5	12.5	9.1	25.1	31.2
Adaptive clustering detection method +FRCNN	Improved DetNet59	45,736	15.3	29.3	16.2	11.7	26.7	32.3

Table 4
Comparison data of ablation experiment

method	Backbone network	Split method	Number of images	mAP	AP50	AP75	APs	APm	APl
FRCN+FPN	Improved Det-Net59	o	548	23.2	41.5	22.4	13.1	35.7	54.8
FRCN+FPN+EIP	Improved Det-Net59	c	3,288	22.1	45.3	20.6	15.4	31.3	31.1
GloDetecNet +Adaptive ClusDet	Improved Det-Net59	o+ca	2,965	30.1	52.7	25.8	19.6	39.5	52.3
GloDetecNet+Adaptive ClusDet+LocDetecNet	Improved Det-Net59	o+ca	2,965	31.4	54.5	27.3	21.7	40.7	53.6

Figure 4
Recall based on NMS

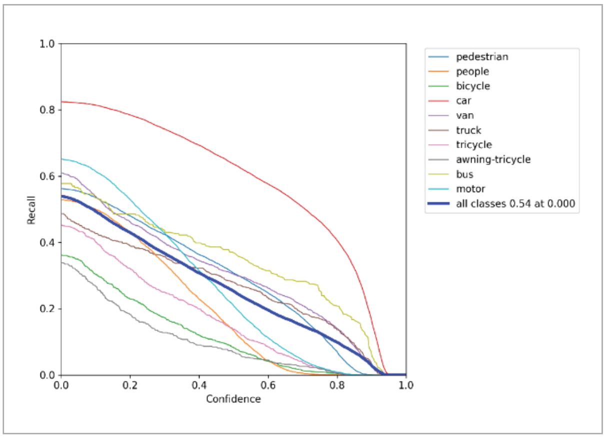


Figure 5
Recall based on softer NMS

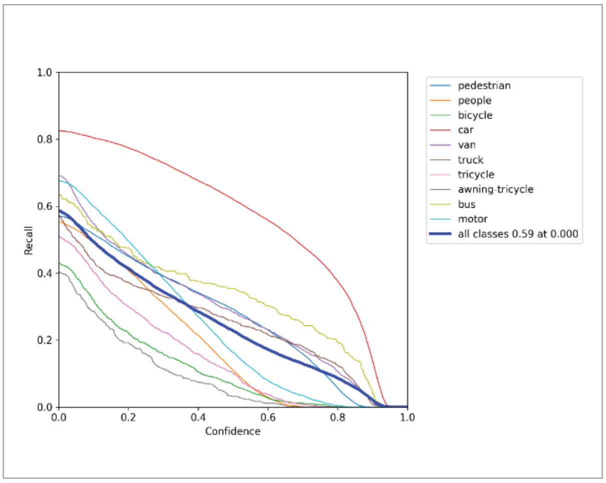
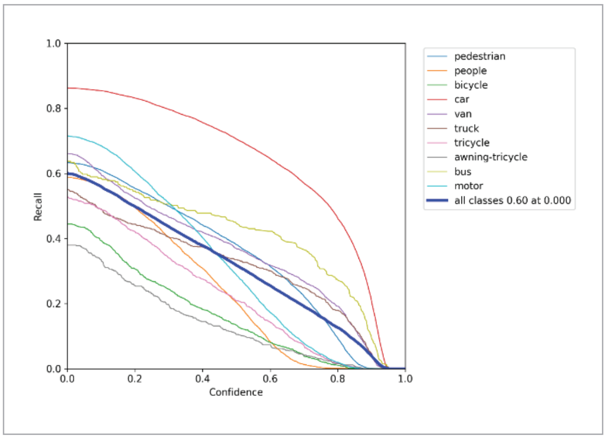


Figure 6
Recall based on Variable threshold NMS



spectively show that the recall rate detected using the variable threshold NMS algorithm has a certain improvement compared to the recall rate detected using the standard NMS algorithm.

4.5.6. Related Experiments on Sample Balance Strategy

To verify the improvement effect of sample balance strategy on network models. This article sets up a comparative experiment, and the experimental process is as follows: the network is trained using the original training set and the training set using the sample balance strategy, and then tested on the VisDrone 2019 test set. The detection results are shown in Figures 7-8.

Figure 7

Detection results without using sample balance strategy

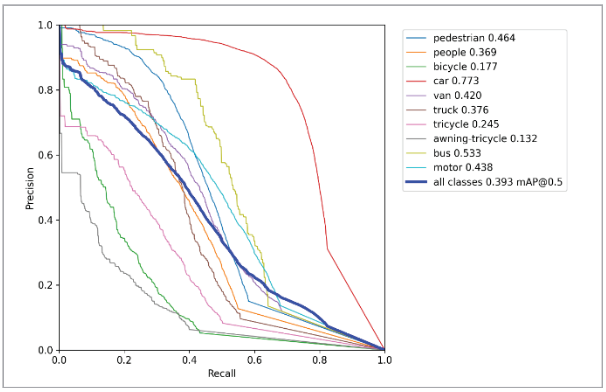
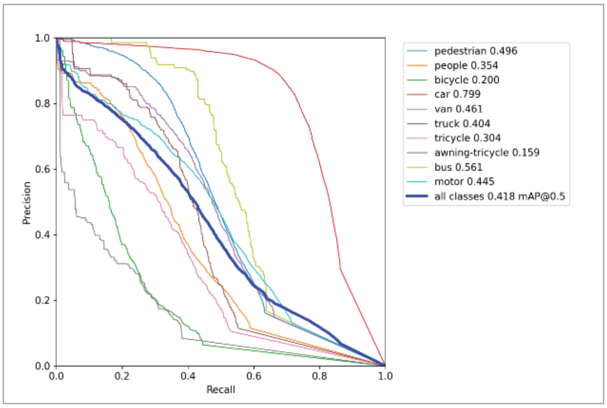


Figure 8

Detection results using sample balance strategy



Calculate the proportion of various target numbers and areas in the VisDrone 2019 training set, with the largest proportion being cars and the smallest proportion being covered tricycles. The provided figures show the accuracy recall (P-R) curves of various target test results. As shown in the figures, the detection accuracy of the car reached 77.3%, while the average detection accuracy of the covered tricycle was only 13.2%, indicating that the existence of a long tail distribution training set has a significant impact on the accuracy of the model. The test results show that the model using the sample balance strategy has improved the detection accuracy of cars by 2.6% to 79.9%, and the detection accuracy of tricycles with canopies has increased by 2.7% to 15.9%. In addition to a decrease in the detection accuracy of 'people', the average detection accuracy of other types of targets has also significantly increased, mAP@0.5 An increase of 2.5%.

As shown in Figure 9, the change curve of relevant parameters in the training set training network process after using the sample balance strategy, the loss function in the figure shows a downward trend and gradually tends to be flat, and the accuracy, recall and average precision curves also tend to be flat, which indicates the effectiveness of the sample balance strategy.

As shown in Figure 10, the detection results of model testing fully demonstrate that the sample balance strategy can effectively improve the detection accura-

Figure 9

Curve chart of changes in relevant parameters during training

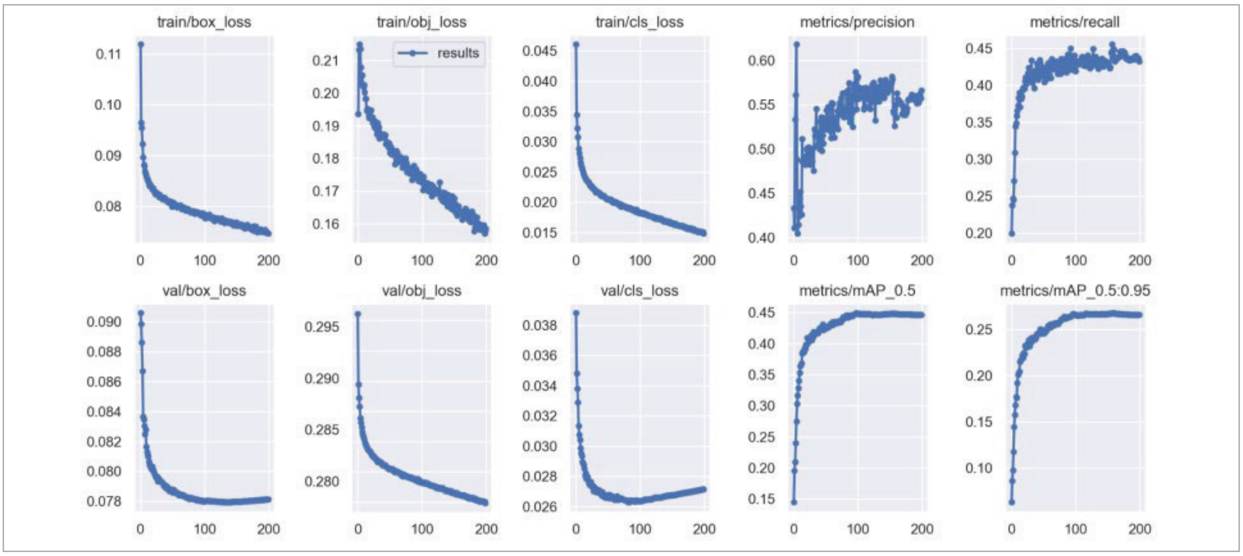
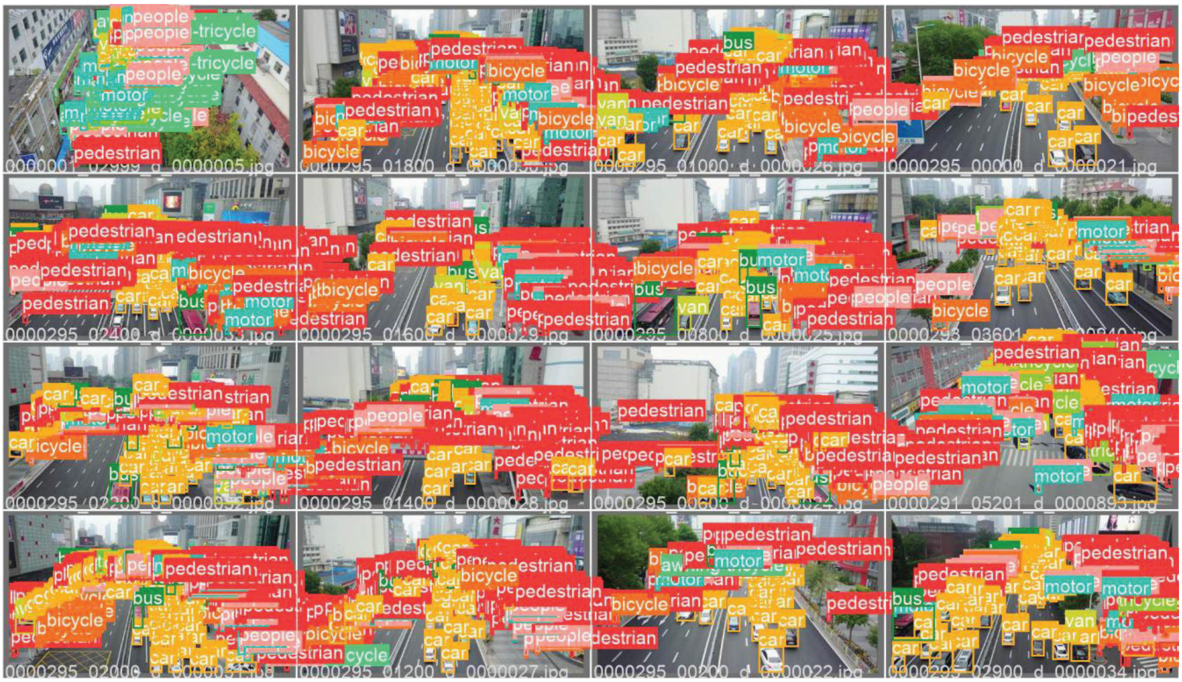


Figure10

Test examples under model testing


0000295_01800_d_0000020.jpg
0000295_01000_d_0000026.jpg
0000295_00000_d_0000021.jpg
0000295_02400_d_0000033.jpg
0000295_01600_d_0000029.jpg
0000295_00800_d_0000025.jpg
0000295_02200_d_0000032.jpg
0000295_01400_d_0000028.jpg
0000295_00600_d_0000024.jpg
0000291_05201_d_0000893.jpg
0000295_02000_d_0000031.jpg
0000295_01200_d_0000027.jpg
0000295_00200_d_0000022.jpg
0000295_02900_d_0000034.jpg
pedestrian
people
bicycle
car
van
tricycle
bus
motor

cy of small sample targets, such as bicycles and other small sample targets in the figure.

To further validate the effectiveness of the model in real drone images, we used drones to capture some images, annotated them, and created a test set.

The test results of drone aerial images are as follows.

The scene in Figure 11 (left) is a basketball court, where pedestrians can be detected correctly. However, in the upper part of the picture, the background environment becomes a construction site, and some vehicles and pedestrians are not detected. This result indicates that the difficulty of object detection in complex background environments is still significant. The scene in Figure 11 (right) is a campus, with a clear background and a larger target size. Targets such as vehicles, pedestrians, and motorcycles can be correctly detected.

Figure 11

Test Result (1)

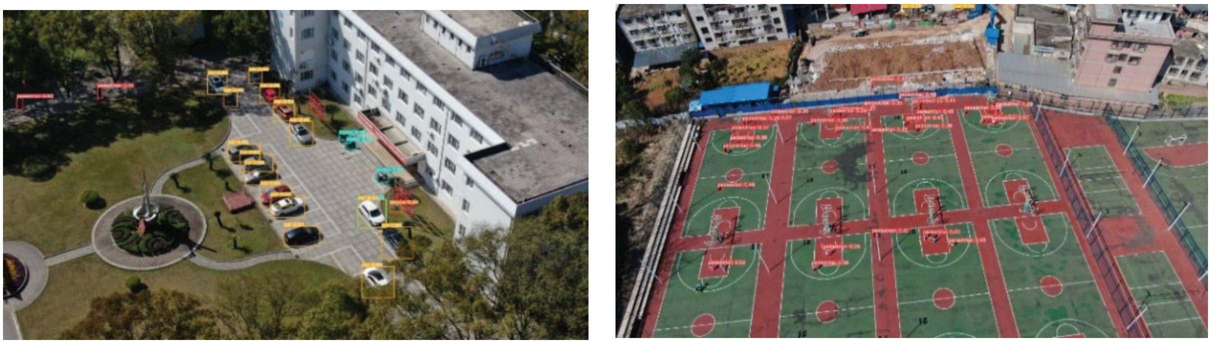
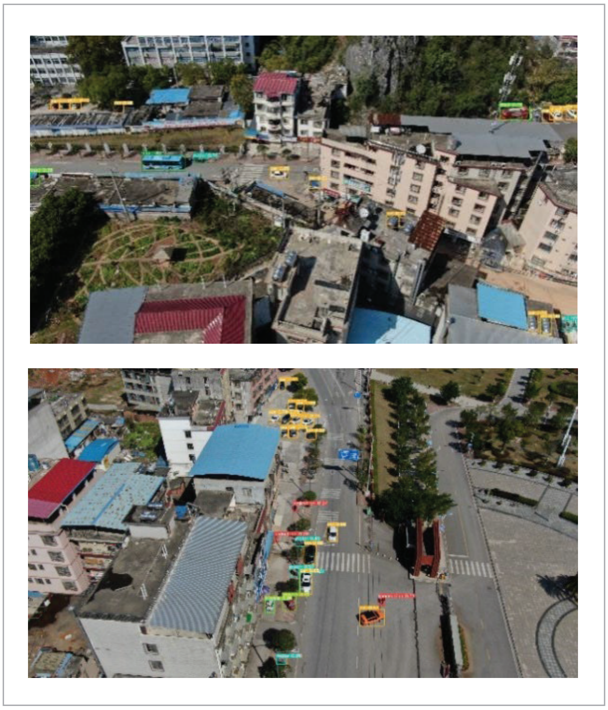


Figure12
Test Result (2)



The background in Figure 12 (top) is relatively complex, with sparse target distribution and high detection difficulty. The water storage tank in the lower right corner of the image is detected as a car, and the detection results of other targets are all correct. The background in Figure 12 (right) is simpler than that in the left image. Most of the targets are concentrated on the road, and most of them have been detected. However, some sparsely distributed targets have been misdetected, such as the shadow of the pedestrian below the image being detected as a pedestrian.

By testing on real drone images, the network proposed in this article can accurately detect most of the targets. However, there are still problems: firstly, when detecting images with complex backgrounds, there may be missed detections; Secondly, when detecting areas with sparse distribution of image targets, false positives may occur. By using data augmentation strategies and optimizing the backbone detection network to train better models, the missed detection rate and false detection rate can be reduced to a certain extent, but this phenomenon cannot be completely eliminated.

5. Conclusion

In order to improve the detection accuracy of small targets in UAV images, an adaptive clustering detection method is proposed. This method divides the whole detection process into three parts: first, this paper proposes an adaptive clustering algorithm to segment the image; Second, the segmentation and filling methods are used to correct the segmented image; Third, the detection network is used to detect the segmented image and the original image, in which the detection network can be any object detection algorithm, and the backbone network of the detection network is used to improve the detection accuracy of the model by introducing attention mechanism, NMS with variable threshold, and using sample balancing strategy. The simulation results show that the adaptive clustering target detection method of Unmanned Aerial Vehicle (UAV) image based on long-tail distribution greatly improves the detection accuracy of small targets, which can effectively improve the detection accuracy of the model for targets in the agglomeration area, and the model has good generalization ability. This article proposes a sample balance strategy that to some extent changes the uneven distribution of data, but the enhanced dataset still has a long tail distribution. It is hoped that researchers can propose better data preprocessing methods to reduce the deviation of algorithm models and improve detection accuracy.

References

1. Aurelio, Y. S., De Almeida, G. M., de Castro, C. L., Braga, A. P. Learning from Imbalanced Data Sets with Weighted Cross-entropy Function. Neural Processing Letters, 2019,50,1937-1949. https://doi.org/10.1007/s11063-018-09977-1
2. Avola, D., Cinque, L., Diko, A., Fagioli, A., Foresti, G. L., Mecca, A., Pannone, D., Piciarelli, C. MS-Faster R-CNN: Multi-stream Backbone for Improved Faster R-CNN Object Detection and Aerial Tracking from UAV Images. Remote Sensing, 2021,13, 1670. https://doi.org/10.3390/rs13091670
3. Aziz, L., Salam, M. S. B. H., Sheikh, U. U., Khan, S., Ayub, H., Ayub, S.Multi-level Refinement Feature Pyramid Network for Scale Imbalance Object Detection.

IEEE Access, 2021, 9, 156492-156506. https://doi.org/10.1109/ACCESS.2021.3130129

4. Bodla, N., Singh, B., Chellappa, R., Davis, L. S. Soft-NMS--Improving Object Detection with One Line of Code. In Proceedings of the IEEE International Conference on Computer Vision, 2017, 5561-5569. https://doi.org/10.1109/ICCV.2017.593
5. Cai, Z., Vasconcelos, N. Cascade r-cnn: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 6154-6162. https://doi.org/10.1109/CVPR.2018.00644
6. Cao, F., Liu, H. Single Image Super-resolution via Multiscale Residual Channel Attention Network. Neurocomputing, 2019, 358, 424-436. https://doi.org/10.1016/j.neucom.2019.05.066
7. Carlotto, M. J. A Cluster-based Approach for Detecting Man-made Objects and Changes in Imagery. IEEE Transactions on Geoscience and Remote Sensing, 2005,43, 374-387. https://doi.org/10.1109/TGRS.2004.841481
8. Chen, S., Ma, W., Zhang, L. Dual-bottleneck Feature Pyramid Network for Multiscale Object Detection. Journal of Electronic Imaging, 2022, 31, 013009-013009. https://doi.org/10.1117/1.JEI.31.1.013009
9. Comaniciu, D., Meer, P. Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002,24,603-619. https://doi.org/10.1109/34.1000236
10. Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., et al. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, 213-226.
11. Kanai, S., Fujiwara, Y., Yamanaka, Y., Adachi, S. Sigsoftmax: Reanalysis of the Softmax Bottleneck. Advances in Neural Information Processing Systems, 2018,31.
12. Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R. S., Indyk, P., Katabi, D. Targeted Supervised Contrastive Learning for Long-Tailed Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2022,6918-6928. https://doi.org/10.1109/CVPR52688.2022.00679
13. Li, Y. L., Wang, S.HAR-Net: Joint learning of Hybrid Attention for Single-stage Object Detection, 2020, 29, 3092-3103. https://doi.org/10.1109/TIP.2019.2957850
14. Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., Sun, J. Detnet: A Backbone Network for Object Detection, 2018, 1804, 06215.
15. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106
16. Lyu, W., Lin, Q., Guo, L., Wang, C., Yang, Z., Xu, W. Vehicle Detection Based on an Imporved Faster R-CNN Method. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2020, 104, 587-590. https://doi.org/10.1587/transfun.2020EAL2071
17. Ma, J., Chen, B. Dual Refinement Feature Pyramid Networks for Object Detection, 2020, 01733.
18. Maskeliūnas, R., Katkevičius, A., Plonis, D., Sledevič, T., Meškėnas, A., Damaševičius, R. Building Façade Style Classification from UAV Imagery Using a Pareto-optimized Deep Learning Network. Electronics, 2022, 11, 3450. https://doi.org/10.3390/electronics11213450
19. Narayan, S. The Generalized Sigmoid Activation Function: Competitive Supervised Learning. Information Sciences, 1997,99, 69-82. https://doi.org/10.1016/S0020-0255(96)00200-9
20. Neubeck, A., Van Gool, L. Efficient non-maximum suppression. In 18th international conference on pattern recognition,20063,850-855. https://doi.org/10.1109/ICPR.2006.479
21. Ning, C., Zhou, H., Song, Y., Tang, J. Inception Single Shot Multibox Detector for Object Detection. In 2017 IEEE International Conference on Multimedia Expo Workshops, 2017, 549-554.
22. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems, 2015, 28, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031
23. Thipsanthia, P., Chamchong, R., Songram, P. Road Sign Detection and Recognition of Thai Traffic Based on YOLOv3. In International Conference on Multi-disciplinary Trends in Artificial Intelligence, 2019, 271-279. https://doi.org/10.1007/978-3-030-33709-4_25
24. Wang, X., Zhu, D., Yan, Y. Towards Efficient Detection for Small Objects Via Attention-guided Detection Network and Data Augmentation. Sensors, 2022, 22, 7663. https://doi.org/10.3390/s22197663
25. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1
26. Wu, Y., Tang, S., Zhang, S., Ogai, H. An Enhanced Feature Pyramid Object Detection Network for Autono-

mous Driving. Applied Sciences, 2019, 9, 4363. https://doi.org/10.3390/app9204363

27. Xu, X., Zhao, M., Shi, P., Ren, R., He, X., Wei, X., Yang, H. Crack Detection and Comparison Study Based on Faster R-CNN and Mask R-CNN. Sensors, 2022, 22, 1215. https://doi.org/10.3390/s22031215
28. Yan, B., Fan, P., Lei, X., Liu, Z., Yang, F. A Real-time Apple Targets Detection Method for Picking Robot Based on Improved YOLOv5. Remote Sensing, 2021, 13, 1619. https://doi.org/10.3390/rs13091619
29. Yan, J., Wang, H., Yan, M., Diao, W., Sun, X., Li, H. IoU-adaptive Deformable R-CNN: Make Full Use of IoU for Multi-class Object Detection in Remote Sensing Imagery. Remote Sensing, 2019, 11, 286. https://doi.org/10.3390/rs11030286
30. Yang, F., Fan, H., Chu, P., Blasch, E., Ling, H. Clustered object detection in aerial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 8311-8320. https://doi.org/10.1109/ICCV.2019.00840
31. Yang, J., Xie, X., Shi, G., Yang, W. A Feature-enhanced Anchor-free Network for UAV Vehicle Detection. Remote Sensing, 2020,12, 2729. https://doi.org/10.1109/34.1000236
32. Yao, J., Qi, J., Zhang, J., Shao, H., Yang, J., Li, X. A Real-time Detection Algorithm for Kiwifruit Defects Based on YOLOv5. Electronics, 2021, 10, 1711. https://doi.org/10.3390/electronics10141711
33. Yao, S., Chen, Y., Tian, X., Jiang, R. GeminiNet: Combine Fully Convolution Network with Structure of Receptive Fields for Object Detection. IEEE Access, 2020,8,60305-60313. https://doi.org/10.1109/ACCESS.2020.2982939
34. Yi, S., Liu, X., Li, J., Chen, L. UAVformer: A Composite Transformer Network for Urban Scene Segmentation of UAV Images. Pattern Recognition, 2023, 133, 109019. https://doi.org/10.1016/j.patcog.2022.109019
35. Yin, Z., Shi, W., Wu, Z., Zhang, J. Multilevel Wavelet-based Hierarchical Networks for Image Compressed Sensing. Pattern Recognition, 2022, 129, 108758. https://doi.org/10.1016/j.patcog.2022.108758
36. Yuan, Y. L., Luo, Y. B., Zhang, C., Zhu, Y. X. FPN Analysis and Processing of the Science Grade CCD Alta U9000. Advanced Materials Research, 2011, 301, 1007-1010. https://doi.org/10.4028/www.scientific.net/AMR.301-303.1007
37. Zeng, L., Sun, B., Zhu, D. Underwater Target Detection Based on Faster R-CNN and Adversarial Occlusion Network. Engineering Applications of Artificial Intelligence, 2021, 100, 104190. https://doi.org/10.1016/j.engappai.2021.104190
38. Zhang, B., Zhao, Q., Feng, W., Lyu, S. AlphaMEX: A Smarter Global Pooling Method for Convolutional Neural Networks. Neurocomputing, 2018, 321, 36-48. https://doi.org/10.1016/j.neucom.2018.07.079
39. Qiu, S. Global Weighted Average Pooling Bridges Pixel-Level Localization and Image-level Classification, 2018, 1809, 08264.
40. Zhang, P., Zhong, Y., Li, X. SlimYOLOv3: Narrower, Faster and Better for Real-time UAV Applications. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, 37-45. https://doi.org/10.1109/ICCVW.2019.00011
41. Zhang, X., Dong, X., Wei, Q., Zhou, K. Real-time Object Detection Algorithm Based on Improved YOLOv3. Journal of Electronic Imaging, 28, 5, 053022-053022. https://doi.org/10.1117/1.JEI.28.5.053022
42. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y. Image Super-resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the European Conference On Computer Vision, 2018, 286-301. https://doi.org/10.1007/978-3-030-01234-2_18
43. Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., Jia, J.Psanet: Point-wise Spatial Attention Network for Scene Parsing. In Proceedings of the European Conference on Computer Vision, 2018, 267-283. https://doi.org/10.1007/978-3-030-01240-3_17
44. Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., Jia, J. Psanet: Point-wise Spatial Attention Network for Scene Parsing. In Proceedings of the European Conference on Computer Vision, 2018, 267-283. https://doi.org/10.1007/978-3-030-01240-3_17
45. Zhao, J., Zhang, X., Yan, J., Qiu, X., Yao, X., Tian, Y., Zhu, Y., Cao, W. A Wheat Spike Detection Method in UAV Images Based on Improved YOLOv5. Remote Sensing, 2021, 13,3095. https://doi.org/10.3390/rs13163095. 46. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D. Distance-IoU loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 34, 12993-13000. https://doi.org/10.1609/aaai.v34i07.6999
46. Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., Zuo, W. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. IEEE Transactions on Cybernetics, 2020, 52, 8574-8586. https://doi.org/10.1109/TCYB.2021.3095305.
47. Zhong, Z., Sun, L., Huo, Q. An Anchor-free Region Proposal Network for Faster R-CNN-based Text Detection Approaches. International Journal on Document Analysis and Recognition, 2018, 22, 315-327. https://doi.org/10.1007/s10032-019-00335-y