IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 13, 2021, accepted July 2, 2021, date of publication July 7, 2021, date of current version July 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3095405

A Small Target Detection Method Based on Deep Learning With Considerate Feature and Effectively Expanded Sample Size

JUN ZHANG, YIZHEN MENG, AND ZHIPENG CHEN
Department of Computer Science, Tangshan Normal University, Tangshan 063000, China

Corresponding author: Yizhen Meng (ru780065@163.com)

This work was supported in part by the Science and Technology Project of Hebei Education Department under Grant QN2020513, and in part by the Scientific Research Foundation of Tangshan Normal University, China, under Grant 2021B24 and Grant 2020A04.

ABSTRACT As a basic task in the field of computer vision, target detection has been concerned by many researchers. The performance of target detection method is also directly related to the research in many advanced semantic fields. Current general target detection methods are not effective in small target detection, so this paper studies the problem of small target detection and proposes a small target detection method based on deep learning with considerate feature and effectively expanded sample size. Firstly, according to the characteristics of convolutional neural network, we improve the current deep network architecture which performs well in target detection, and introduce considerate multi-feature and multi-scale detection. Then, we transform the high-resolution images obtained on the Internet by combining two groups of sampling method, so that the feature distribution of the high-resolution target is closer to that of the low-resolution target, thus effectively expanding the training data set, solving the problem that small target data is difficult to be labeled and effectively avoiding overfitting. The results show the effectiveness of the improved method in small target detection. In addition, in view of the shortage of small target detection review literature, this paper gives a comprehensive and detailed introduction to the field of small target detection in terms of related work and future work.

INDEX TERMS Deep learning, target detection, feature extraction, sample size, overfitting.

I. INTRODUCTION

Computer vision originated from the neural network technology in the 1980s, and has been rapidly developed in recent years [1]. Computer vision is mainly used for image classification [2], image detection [3] and image segmentation [4] on behalf of human eyes. From the perspective of engineering, it can achieve automation of tasks based on human vision. Target detection is a fundamental computer vision task that combines two tasks, i.e., target localization and target recognition. Its purpose is to find several targets in the complex background of an image, to give an accurate target bounding box, and to determine the category to which the targets in that box belong [5]. The effect of target detection directly determines the effect of many high-level vision tasks such as image semantic understanding and target re-identification, and it has good application prospects in intelligent surveillance systems, medical image analysis, etc., so its research has strong theoretical and application values. In conclusion, target detection has been one of the several research directions that have received much attention in the field of computer vision. As shown Figure 1, in which shows some detection results of plain targets.

Small target detection has a wide range of important applications in many fields [6]. For example, in the field of autonomous driving, the pedestrian target or traffic sign in the high-resolution scene photos collected by the car is too small, but the accurate detection of these small targets is an important prerequisite to achieve safe autonomous driving. In the field of medicine, the successful detection of small masses in medical images is an important prerequisite for the early and accurate diagnosis of tumors. Automatic industrial inspection locating small defects on the surface of materials also shows the importance of small target detection.

The associate editor coordinating the review of this manuscript and approving it for publication was Longzhi Yang.

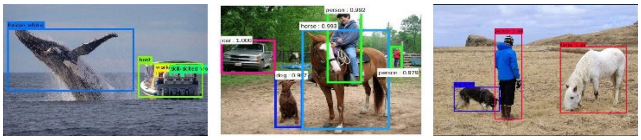


FIGURE 1. The plain target detection instance.

The analysis of satellite images needs to effectively annotate objects such as cars, boats and houses that are too small to be detected. In the criminal investigation images, abnormal small packages, small pedestrians, small pendants inside the car, small signs on the clothes, some indoor decorations, etc., are the key clues to solve the case. Because more complex systems are deployed in the real world, small target detection is of great value. There are two main definitions about small target [7]. The first is absolutely small object. It is specified in the COCO data set [8] that when the number of pixel points of an object is less than 32×32, the object can be regarded as a small object, as shown in Figure 2. The second is a relatively small object, which can be considered as a relatively small object when the target size is less than 0.1 times the size of the original image in terms of the length and width of the original image, as shown in Figure 3.

Most of the early target detection algorithms are built based on manual features. As shown in Figure 4, the basic idea is to first look for the areas where the target may exist in the input original image, then extract the features of each area and send them into the classifier model for judgment, and finally screen the areas considered as targets by the classifier model for post-processing operations to obtain the results. In the absence of an effective image representation at the time, there was no choice but to design complex feature representations and use various acceleration techniques to use up limited computational resources. Since AlexNet proposed by Krizhevsky *et al.* [9] achieved significant improvement in the accuracy of IMAGENET image classification task, various deep learning methods represented by convolutional neural network (CNN) have been widely used in many vision tasks, which also include target detection. Since deep learning methods usually achieve better results than traditional manual feature-based methods, deep learning methods have now become mainstream in the direction of target detection, and most of the research work is centered on CNNs.

However, even though these deep learning-based methods achieve good results on a generic target detection dataset, they still do not solve the problem of small target detection well. For example, Figure 5 is the result of small target detection about *excavator* using the proposed method, while the general target detection method cannot be completed at all.

There are 2 main difficulties in small target detection.

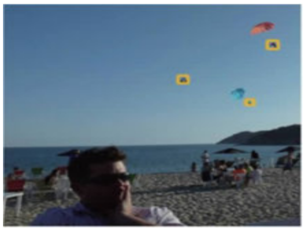


FIGURE 2. The absolutely small target.

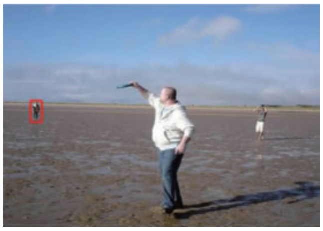


FIGURE 3. The relatively small target.

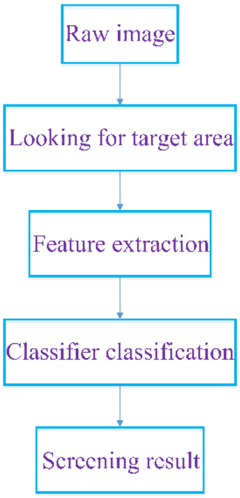


FIGURE 4. The traditional target detection pipeline.

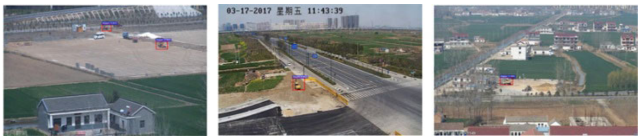


FIGURE 5. Our small target detection partial results.

(1) When the target occupies a very small percentage in the image, the amount of information reflected by the pixels in the corresponding region is very limited. In the extreme case, the small target detection task may even degenerate into a pixel classification task. This makes it difficult to apply some general-purpose target detection algorithms to small target detection, while some algorithms designed specifically for small target detection can only be used for specific application contexts and lack generality.

(2) The small targets in the labeled images are prone to errors when used as training data, and subtle errors can easily have a large impact on the detection results when the targets themselves are already small, and the labor cost of labeling the data is also high [10], so there is not a large and complete dataset for small target detection research so far, which hinders the academic research on small target detection.

To address the above problems, a small target detection method based on deep learning with considerate feature and effectively expanded sample size is proposed in this paper. This method modifies the network structure according to the characteristic of CNN, so that the network can use both low-level and high-level features for multi-feature and multi-scale detection. In addition, this paper uses the high-resolution large target images obtained from online search engine crawlers for training to solve the problem of lacking small target training data and avoid overfitting. Since the data distribution of the high-resolution large-target training image and the low-resolution small-target test image are very different, the problem is analyzed by visualization and then the differences between the training image and the test image are eliminated as much as possible using the downsampling and upsampling methods. The experiments show that the proposed method can indeed solve the problem of small target detection better. Finally, in view of the shortage of small target detection review literature, this paper completes it in two aspects of relevant work and future work. Specifically, the main contributions of this paper include:

(1) To address the shortcomings of Faster-RCNN in small target detection, a generic network structure modification rule is proposed. Concretely, multi-feature and multi-scale detection are performed using low-level and high-level features to improve the accuracy of small target detection.

(2) By downsampling and upsampling the target high-resolution image, the distribution of the data obtained online is as close as possible to the actual test data, solving the problem of lacking small target training data.

(3) Although it is a scientific paper with a research nature, our related and future work, especially in small target detection method about deep learning technology, is comprehensive and detailed.

II. RELATED WORKS

Current target detection methods are mainly divided into two categories: target detection methods based on traditional artificial features and target detection methods based on deep learning. In the following, we will introduce these two aspects separately, in which traditional machine learning methods only list several very representative methods, while for deep learning, we will expand the mainstream methods and related improvements and skills related to small target detection methods in more detail.

A. TRADITIONAL TARGET DETECTION ALGORITHM

1) VJ-DETECTOR

In 2001, Viola and Jones designed an efficient face detector, which was dozens of times faster than other detectors at the time. This was a milestone in the development of face detection and computer vision. In honor of this work, it was named Viola Jones (VJ) detector [11], [12]. The VJ-detector adopts the detection method of sliding window, uses Haar feature to describe each window, and introduces the integral graph to accelerate the extraction of Haar feature, so that the computational complexity of each window is independent of the window size. Adaboost algorithm [13] is combined for feature selection, and the idea of cascading is introduced. In this way, the computation of background window is reduced and the computation of face target is increased, and the computation scale is reduced while the accuracy is improved.

2) HOG FEATURE

Histogram of Oriented Gradient (HOG) feature [14] was originally a local feature proposed by Dalal et al for pedestrian detection. As the name implies, HOG feature accumulates the gradient values in different directions in a certain area of an image to form a histogram, which serves as the feature of this area. HOG feature can better extract the local details of the image, and has good feature invariance in the case of image geometry, deformation, optical distortion, etc. Therefore, HOG feature has been the basis of many target detectors and various computer vision systems for many years.

3) DPM

In view of the poor performance of HOG feature in handling the occlusion problem, Felzenszwalb *et al.* [15] proposed Deformable Part Model (DPM) algorithm in 2008, and then combined with Felzenszwalb *et al.*, made various improvements [16]–[18]. They performed well at that time and won the VOC Challenge 2007, 2008, and 2009 continuously. DPM algorithm adopts a divide-and-conquer idea, which can regard the training and detection process as the collection of learning and detection of each component of the object. It also improves HOG feature, cancelling the block in the HOG feature and retaining only the unit. In the subsequent improvement, it combined some other important technical ideas to improve the accuracy, such as hard case mining, boundary box, regression, etc., which still has a profound impact on the present.

B. TARGET DETECTION ALGORITHM BASED ON DEEP LEARNING

In view of the main technical difficulties of traditional machine learning methods mentioned above in small target detection, many scholars have begun to study the improvement and skills of deep learning-based target detection technology in small target detection methods.

1) PRINCIPAL METHOD

Regional Convolutional Neural Network (RCNN) [3], [8], [9] is a representative work on target detection using deep learning methods. The RCNN proposed by Lin *et al.* [8] is a pioneering combination of candidate region generation and deep learning classification methods. RCNN generates some candidate regions by over-segmentation [10], and then uses CNN to extract features for each candidate region separately, and finally sends them to the classifier to determine the class and regress the edges. This method is very slow because of the repeated convolution of different candidate regions. Learned from the spatial pyramid pooling network (SPPNet) proposed by Viola *et al.* [11] and the localization idea proposed by Viola *et al.* [12], Krizhevsky *et al.* [9] proposed Fast-RCNN. This method introduces region of interest pooling (ROI pooling) on the basis of RCNN, which is actually a single-layer pyramid pooling layer that allows the network to generate features of the same size for different sizes of input images, ensuring the dimensional invariance of the input images. It also extracts the features of the candidate regions directly on the feature map of the whole image by the feature mapping method, which avoids repeated convolution and outperforms RCNN in terms of accuracy and speed. After Fast-RCNN, the main constraint on the speed of this method becomes the over-segmentation used for candidate region generation. Subsequently, Ren *et al.* [19] proposed Faster-RCNN on the basis of Fast-RCNN. This method uses anchor to generate candidate regions and leaves the candidate region generation to the deep network, which further improves the speed and accuracy. So far, the use of deep learning for target detection is unified by RCNN into a deep framework. After RCNN, Redmon *et al.* [25] proposed a faster target detection method, i.e., YOLO (you only look once). YOLO differs from RCNN in that it treats target detection as a regression problem by directly regressing the target bounding box and the class to which it belongs on the divided grid. Because the complex and time-consuming candidate region generation is eliminated, YOLO is very fast, but its detection accuracy is low and its generalization ability is weak for very close or small targets. Combining the anchor idea of RCNN and the regression idea of YOLO, Liu *et al.* [26] proposed the SSD (single shot multi-box detector). SSD has the advantages of accurate localization of RCNN and fast speed of YOLO, and because of the introduction of multi-scale detection [27], it has a better detection effect for targets of different sizes. The detection speed and accuracy are further improved.

These three types of methods mentioned above have good accuracy for the general target detection problem, however, the detection accuracy for small targets is not satisfactory for all of them. In fact, the targets that are not detected by these methods are often not some complex targets, but some smaller targets, such as the *bottle* in the PASCAL VOC dataset [10]. This indicates that it is not the lack of learning and representation ability of the deep network, but the information that can be represented by the small target features extracted by the deep network is too little [28].

In addition, there are some researchers who have specifically studied the detection of small targets. Takeki *et al.* [29] proposed a small target detection method combining image semantic segmentation, which combines the fully convolutional network (FCN) and its variants with the CNN to integrate the results of all three with support vector machine (SVM). However, this method is only applicable to the task of detecting small birds in a pure sky background, and it is difficult to apply to the task of detecting multiple types of targets in complex backgrounds. Chen *et al.* [30] improved RCNN so that RCNN can generate smaller candidate regions, which is a great improvement to RCNN for the task of small target detection, but the complexity of the algorithm is high. Eggert *et al.* [31] also improved RCNN, and they investigated the relationship between feature map resolution and detection effectiveness. In the context of the problem of company logo detection, the improved anchor box generation method is used to improve the effectiveness of RCNN for detection using high-resolution feature map.

It can be seen that although some work has been done for small target detection, those methods can only be used in the context of specific problems or are less effective than the previous three methods for general target detection and lack some generality.

2) TIP1: IOU THRESHOLD MATCHING

Intersection over Union (IOU) refers to the ratio of intersection and union between the predicted bounding box and the real bounding box. that is, the overlap degree [32] between the object bounding box and the ground truth. IOU is defined as a standard to measure the accuracy of object positioning. In target detection, the threshold of IOU is set to 0.5 by default, that is, as long as IOU is greater than or equal to 0.5, it will be considered as a positive sample. If the IOU threshold is set low, the quality of samples is difficult to guarantee; In order to obtain high-quality positive samples, the IOU threshold can be increased. However, the number of samples will decrease, resulting in imbalance [33] of positive and negative samples, and higher IOU threshold is easy to lose small-scale target boxes.

To solve the above problems, literature [34] proposed a multi-stage cascade structure. By continuously increasing the threshold value of IOU, it can ensure the number of samples without affecting the quality of samples. Finally, a high-quality cascade R-CNN detector is trained. The IOU thresholds of candidate boxes are gradually increased (the thresholds are 0.5, 0.6 and 0.7 respectively) in three detection model stages. When the candidate box threshold is close to the training threshold, the sample will get closer to the ground truth value to adapt to the multi-level distribution with each regression. Therefore, the candidate box resampled in the previous stage can be more suitable for the next stage, and samples meeting the corresponding threshold can be obtained while solving the over-fitting in training. Experiments show that the cascade R-CNN structure is used on the reference detector to achieve good detection results on the MS COCO data set, and the detection accuracy for small targets is also improved. At the same time, Liu *et al.* [35] also proposed the idea of improving small target pedestrian detection by increasing IOU threshold. In SSD-based pedestrian detection, a single IOU threshold is used for training to define positive and negative samples. To avoid the limitations of a single-stage detector, the ALF module is proposed. The idea of cascade network is used for multi-step prediction to gradually locate, and a network base on that ResNet-50.

The original image is sampled at 8, 16, 32 and 64 times to extract multi-scale feature maps. The regression anchor frame is used instead of the default anchor frame optimization predictor in each stage, and multiple positioning models are trained by using the continuously increasing IOU threshold to generate more accurate positioning, thus solving the limitation of single-stage detection model SSD on pedestrian detection and improving the detection performance of small-scale pedestrians.

According to the cascade idea, the above two methods can obtain high-quality positive samples by continuously increasing the IOU threshold, which can improve the detection effect of small targets to a certain extent. However, the number of matched anchors decreases, resulting in missed detection with the continuous increase of the IOU threshold. The IOU threshold is reduced from 0.5 to 0.35 in literature [36], and the method of reducing the threshold is used to ensure that each target can have enough anchor frame detection. At the same time, in order to solve the problem that the sample quality cannot be guaranteed due to the increase of positive samples, the method of maximizing background label is proposed. In the lowest level classification, the background is divided into multiple categories instead of two categories. Anchors with IOU greater than 0.1 are sorted, and the background value is predicted three times for each box. The maximum value in the background probability is taken as the final background. By improving the classification difficulty, the problem that the quality of positive samples cannot be guaranteed is solved and the detection accuracy of small targets is improved. However, this method may lead to the problem that the IOU threshold is too low, resulting in too many invalid positive samples, which leads to an increase in the false detection rate.

For different detection tasks, if there is little difference between the scales of the targets to be detected, that is, most of the targets in the data set are of the same scale, the IOU threshold can be appropriately lowered before selection, so as to extract the features of small targets to the greatest extent. In practical application, the detection in the same scene cannot only contain targets of a single scale, and there is a large difference in the scale span of different targets. If the IOU threshold is fixed for unified detection and screening, the problem of sample imbalance will be brought about, and the features of small targets are most likely to be discarded by the strict IOU threshold. Therefore, it is more universal to set dynamic IOU threshold as target detection of different scales. It is dynamically adjusted according to different sample numbers. When the number of negative samples is too high, the IOU threshold is continuously increased to balance the number of samples, thus avoiding missed detection caused by directly setting too high IOU threshold and the trained model has stronger generalization.

3) TIP2: ANCHOR FRAME DESIGN

In the introduction, it is mentioned that most of the traditional target detection algorithms are affected by sliding windows, which need to traverse sliding windows position by position to generate different preset borders. With the emergence of deep learning, anchor frame was initially applied to Faster RCNN [37] model, which to some extent solved the disadvantages of low efficiency caused by traversing sliding windows. When the Faster RCNN model uses RPN (region proposal networks) to generate candidate detection boxes, an anchor with a minimum scale of 128 × 128 and the average size of the candidate box is more than 100 × 100. That is, the set minimum anchor is much larger than the small target to be detected. However, if the input image is considered to be enlarged to match the Anchor in order to detect the small target, the large target may be continuously enlarged so that there is no corresponding Anchor for detection. Therefore, Faster RCNN considers the detection of targets of different scales and the designed Anchor should cover all targets in the training set as much as possible. That is, each target can be matched to one or more Anchors.

With the advent of anchor frame technology, anchor has been widely used in mainstream target detection networks such as SSD and YOLO. In order to better detect small targets, SSD designs anchors with different sizes for different convolution layers. For shallow convolution conv4_3, 6 small anchors with different scales of 60 are used, and for deep convolution conv10_2 and conv11_2, 4 large anchors with different scales of $\{228, 270\}$ are used. Through this method of setting anchor according to the size of the target in training, taking into account the small and dense anchor owned by the small target and the large and sparse anchor owned by the large target, SSD obtains better small target detection results than Faster RCNN. YOLO uses the data of the full connection layer to complete border prediction [38], and regards object detection as a regression problem. However, YOLO will lead to the loss of more spatial information, resulting in inaccurate positioning, and is not good at detecting dense small objects. YOLOv2 [39] abandons the full connection layer and introduces anchor mechanism to predict bounding box. In order to effectively reduce the initial loss, YOLOv2 did not directly use manual design of anchor frame size, but clustered the training set through K-means algorithm [40]. Through clustering, the anchor frame size which is more in line with the target size distribution characteristics in the data set is found, which reduces the difficulty of border regression to a certain extent, converges faster, and is more conducive to network training. YOLOv3 [41] also uses clustering to obtain 9 anchors instead of 5 anchors of YOLOv2, and the size of the small anchor frame on the feature map can be as small as 10 × 13. Through clustering, the complexity of the model and the IOU area are balanced, and the detection performance of small objects is improved. Since then, YOLOv4 has also borrowed the anchor frame mechanism of YOLOv3.

In addition, anchor is also used in the improved algorithm of mainstream target detection framework to improve the accuracy of small target detection. The SNIP framework [42] filters anchor by analyzing the relationship between the small-scale target and the scale of the pre-training model.

If the ground truth box is located in a given candidate area range, it is judged to be a valid box. Otherwise, it is an invalid box; If the overlap between an anchor and an invalid box exceeds 0.3, the anchor is determined to be an invalid anchor. At the same time, SNIP introduces multi-scale training, corresponding to three different resolution images. During training, invalid anchor is not backpropagated, but targets of appropriate size are selectively selected for gradient update. Therefore, small targets always have the opportunity to participate in training within the appropriate scale range, so as to realize the normalization of target scale and features and improve the detection effect of small-scale objects. In order to improve the recall rate of small targets, literature [43] also proposes a new dense anchor frame strategy. Specifically, $A^{density} = A^{\mathrm{scale}}/A^{\mathrm{interval}}$, where $A^{density}$ represents the density of anchor, A^{scale} represents the anchor scale, and A^{interval} represents the anchor interval. A^{scale} is 32×32, 64×64, 128×128, 256×256, 512×512 respectively, and A^{interval} is 32, 32, 32, 64, 128 by default, then $A^{density}$ is 1, 2, 4, 4, 4. Obviously, the density of anchor is different at different scales, and the small-scale anchor frame in shallow network is sparser than the large-scale anchor frame in deep network. Aiming at the problem of unbalanced anchor frame density, the shallow small anchor frame is densified. For example, the 32×32 small-scale anchor is densified four times to ensure that the Anchor of different scales has the same density, so as to improve the recall rate of small-scale targets. Literature [44] also starts from anchor's point of view, the sampling step size of anchor related to feature mapping is reduced by increasing the feature mapping scale in the network model and the anchor density is increased around the original predefined anchor center. The number of anchors matching with the small target ground truth is increased, thus making up for the deficiency of poor detection performance for small targets. Literature [36] points out that the detection effect of small target face based on anchor is not ideal, i.e., there is a mismatch problem between receptive field, designed anchor and small target face, and the size of small target is much smaller than the designed anchor. Since the size of anchor is not continuous while the size of face is continuous, the number of anchor available within a certain set range will be reduced, and too small or too large targets cannot match enough anchor. If anchor is added blindly to detect small targets, the increase in the number of negative samples is not ideal for the detection effect. Therefore, anchors with different scales are set for different feature layers to solve the problem of lack of available anchors, the size of anchor is adjusted at equal proportional intervals to set the size from 16 to 512. The size value of anchor can roughly cover the range of effective receptive fields, ensuring that each feature layer has a corresponding anchor, and satisfying that targets of different sizes can be matched to appropriate anchor for detection.

Anchor frame design is widely used in small target detection. By designing an anchor frame that is more in line with the target size distribution characteristics in the data set, the scale value of the anchor frame matches the range of the effective receptive field as much as possible, thus improving the recall rate of small targets and improving the detection effect of small targets. However, the anchor frame suitable for detecting one small target may not be suitable for detecting other small targets because the small target detection scene is usually complex and there are many types of small targets; However, if the number of anchor frames is increased to detect small targets of different types and sizes, the number of negative samples will increase, resulting in an increase in the false detection rate.

4) TIP3: NARROWING TARGET DIFFERENCES

In 2014, the core idea of GAN (Generative Adversarial Net) put forward by Wang *et al.* [45] originated from Nash equilibrium of game theory. GAN has become a hot research direction in the field of deep learning in recent two years, and is widely used in image super-division reconstruction [46], representation learning [47], style transfer [48] and other tasks. GAN network mainly consists of two parts: Generator (G) and Discriminator (D), each of which has its own role in the game.

For detecting small targets, literature [49] proposes to use Perceptual GAN to enhance the feature expression of small targets. The traditional generator in GAN learns the mapping from noise distribution to data, while Perceptual GAN is responsible for finding structural associations between objects of different scales. In the generator, the original poor small target features are converted into super-resolution expression forms by introducing low-level fine granularity features so that the generator can generate large-scale targets from fake small-scale targets, and reduce the representation differences between objects. Make small objects and large objects have similar feature representations; The discriminator is used to distinguish whether it is a real object feature or a feature generated by the generator. The alternate training of the two sub-networks finally achieves a balance, which improves the detection accuracy of small targets in the detection of the Tsinghua-Tencent 100K traffic sign data set [50] and the Caltech pedestrian data set [51], and which has excellent results.

In addition, literature [52] proposes a multi-task generative countermeasure network MTGAN to detect small targets. This framework can be applied to any existing detector. The generator G generates high-quality images with the help of super-resolution network, and the discriminator D discriminates whether it is a real picture or a super-divided picture. At the same time, the classification loss and regression loss of the discriminator D return to the generator through back propagation, prompting the generator G to have more detailed information of small object images. The two compete with the training method of learning through alternate iteration until the data generated by G is false and true, which makes D unable to distinguish accurately. The AP value of MTGAN in small target detection is increased by 1.5% compared with baseline detectors Faster RCNNand Mask-RCNN [53].

So as to improve the effect of small target detection, images with high resolution and obvious feature information of small targets can be obtained with the help of GAN network, and the scale of data sets can be increased. However, using GAN network to detect small targets may lead to unstable training. Specifically, if some features in the result generated by G are approved by D at a certain time, G will think that the output is correct and will continue to output similar results. In fact, the result generated by G is not good, resulting in incomplete missing features in the final generated result and poor detection effect. Therefore, the GAN network is used for small target detection. It is suitable for scenes with single type of small targets and obvious feature information.

5) TIP4: HYPERPARAMETER TUNNING

Model parameters based on depth learning are mainly divided into parameters and hyperparameters. Parameters are usually automatically obtained from data and do not need to be manually set, while hyperparameters are configuration variables outside the model and usually need to be manually set. Hyperparameters mainly include learning rate, batch size, iteration times (epoch), number of hidden layers, selection of activation function, adjustable coefficient of partial loss function and regularization coefficient, etc. Hyperparameter optimization is a key step in target detection based on depth learning. It is even more necessary to select the optimal hyperparameter by means of parameter optimization combination in small target detection, so as to give full play to the maximum performance and better detection of small targets.

At present, according to the implementation mechanism and advantages and disadvantages, there are four methods of hyperparameter tuning, i.e., manual adjustment, grid search, random search and Bayesian optimization algorithm. Compared with manual adjustment methods that require some knowledge and previous experience, the automatic hyperparameter optimization method can more effectively select the relatively better super-parameter combination for the model. However, most of the existing automatic hyperparameter optimization cannot get rid of the fixed network model structure and data set. The adaptive adjustment has the problem that the optimal hyperparameter combination obtained in one small target detection model may not be applicable to another small target detection model, and the specific model still needs to be optimized.

III. THE PROPOSED METHOD

A. MULTI-FEATURE AND MULTI-SCALE DETECTION

In a multi-layer convolutional neural network, the features in the lower layers tend to represent the detailed information of the texture and edges of the image well, while the higher up the hierarchy, as the receptive field of neurons expands, the features in the higher layers tend to represent the semantic information of the image well, but accordingly some detailed information is ignored [54].

When the target is very small, the semantic information that can be reflected from only pixels is very limited, and targets that are too small do not require neurons with large receptive fields at all, so we generally have to rely more on detailed information from the lower layers in order to identify small targets. To prove this conclusion, we use gradient ascent to reconstruct the image features [55] as a way to visualize how the features of small targets extracted by different layers of the deep network differ. As shown in Figure 6-8, the *excavator* in the border in Figure 6 is a small target to be detected, and we extract its features through the VGG16 network and reconstruct the image with the features extracted from the conv1_2 and conv5_3 layers; Figure 7 is the result of reconstructing the features in the conv1_2 layer of the VGG16 network, and it can be clearly seen as an excavator; while Figure 8 is the result of reconstructing the image with the VGG16 network conv5_3 layer features, and only the outline can be seen. Therefore, for the problem of small target detection, the low-level features of the CNN are often more effective than the high-level features.

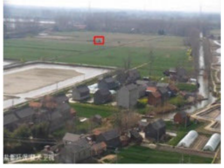


FIGURE 6. The image with small target.

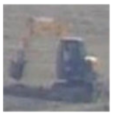


FIGURE 7. The reconstructed result using the conv1_2 layer of VGG16.

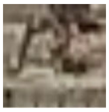


FIGURE 8. The reconstructed result using the conv5_3 layer of VGG16.

In the original Faster-RCNN approach, candidate regions are generated by the RPN, and the features of the candidate regions are obtained by pooling the target regions by the last convolutional layer only, and there are obviously more problems in detecting small targets using such high-level features. Therefore, we introduce multi-feature and multi-scale detection to Faster-RCNN, i.e., instead of relying on the feature map of the last layer alone for detection, we generate candidate regions for multiple scales of feature map in the network. The specific process is shown in Figure 8. The input image is extracted by a CNN, and the multiple feature map of different scales extracted from different layers are sent to their respective RPNs to generate candidate regions, and the RPNs corresponding to different scales are different, because the receptive field of the neurons in the lower layers is small, the corresponding anchor box size should also be small, so the lower the features of the lower layers the smaller the candidate regions are. The specific anchor settings will be explained in detail in the experimental section IV. After getting the generated candidate regions, the feature map mapping is obtained, and then the features are turned into uniform size by ROI pooling and finally fed into the classifier, so that the low-level features can be fully utilized for the detection of small targets. Such a structure is applicable to different feature networks, and we improve Faster-RCNN method using two feature networks, ZF [56] and VGG16 [57], respectively, in the experimental part. For the ZF network, the outputs of the three layers conv1, conv2, and conv5 are fed into the candidate region generation network and the ROI pooling layer for multi-scale detection. For the VGG16 network, the outputs of the five layers conv1_2, conv2_2, conv3_3, conv4_3, and conv5_3 are fed into the candidate region generation network and the ROI pooling layer for multi-scale detection. The other specific parameter settings will be explained in the section IV.

B. EFFECTIVE TRAINING DATA TRANSFORMATION

The improvement of the network structure solves the problem that it is difficult to detect small targets using only high-level network features. Aiming at the problems of small targets that are difficult to label and lack of training samples, we use images obtained from the Internet as the training data, a total of 7804 images. However, images retrieved by search engine keywords tend to be subject to targets. Compared with small targets in the test images, the targets in these images have higher resolution. For example, what we retrieve by *excavator* are generally images with excavator as the main body, and the target of the excavator occupies a very large proportion of the image, while the proportion of the target in the test image is very small. The number of pixels between the two is not the same, and the amount of information reflected is also not the same, so there may be differences in the distribution of data. As shown in Figure 10-12. Figure 10 shows some of the training images of large targets acquired online with high resolution, Figure 11 shows the training images after sampling and processing, and Figure 12 show some of the test images for small target detection.

To show that the distribution between high-resolution (HR) targets and low-resolution (LR) targets does differ, we use the method from T-SNE [58] to reduce the target features in these two types of images. We validate with the original Faster-RCNN structure, and the feature extraction network is VGG16. We train the network with the target low-resolution image, and then test it with the target low-resolution image and the target high-resolution image. The feature vectors of the same size obtained after the ROI pooling layer are reduced in the dimensionality and the visualization results are shown in Figure 13, where the purple dots represent the low-resolution target and the blue crosses represent the high-resolution target, which shows that there is indeed a great difference between the two.

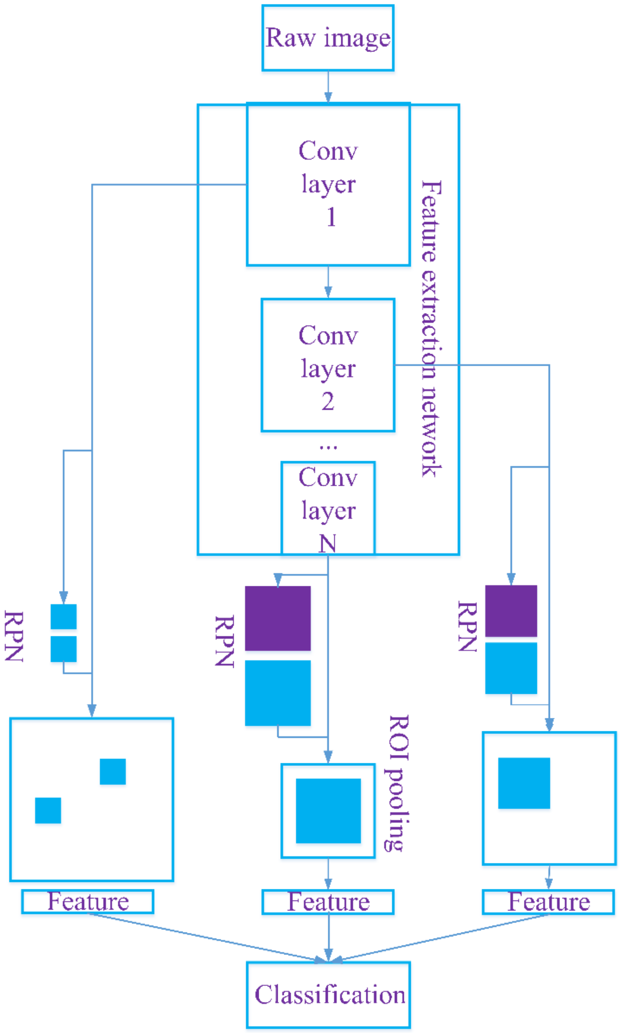


FIGURE 9. The flowchart of proposed method.

FIGURE 10. The image taken from the internet.

To address such a distribution difference, we preprocess the training data using downsampling and upsampling. The used downsampling methods include maximum pooling and average pooling, which reduces the information of the high-resolution images. The used upsampling methods include linear interpolation, region interpolation and nearest neighbor interpolation, which reduces the image to its original size and introduces some noise. From the visual point of view, the sampled training image is more similar to the test image. We have experimented with the combination of these six sampling methods in the section IV and the training images after the best sampling method are shown in the Figure 10. The experiments demonstrate that downsampling and upsampling can effectively improve the detection

FIGURE 11. The trained image has been processed.

FIGURE 12. The tested image in the realistic scene.

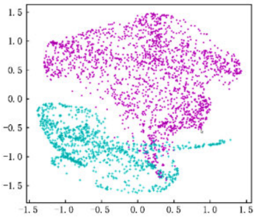


FIGURE 13. The feature distribution of high-resolution targets and low-resolution targets.

accuracy of the model trained with high-resolution target images to detect low-resolution target images. The effects of different downsampling and upsampling methods on the detection results are explained in the section IV.

IV. EXPERIMENT

A. EXPERIMENTAL SETUP

The data set used in the experiment consists of two parts, and the small target for detection is the excavator. One part comes from the images taken by the monitoring cameras on the towers of the base station, with a total of 15,477 images, in which the excavators are usually very small; the other part comes from the high-resolution images of large targets obtained by searching the keyword excavator through the search engine, with a total of 7804 images, in which the excavators are usually large, as shown in Figure 9.

In order to compare the accuracy of the methods, two backbone networks, ZF and VGG16, are used as feature extraction networks for the experiments. Because the detection target is an excavator, the scale parameters of the anchor box are set to 0.7, 1, 1.4. Each scale of feature map corresponds to a different anchor size of the candidate region generation network. For the ZF network: the scale parameters of conv1 layer are 2, 4, 8; the scale parameters of conv2 layer are 4, 8, 16; the scale parameters of conv5 layer are 8, 16, 32. For the VGG16 network: the scale parameters corresponding to the conv1_2 layer are 2, 4; the scale parameters corresponding to the conv2_2 layer are 4, 8; the scale parameters corresponding to the conv3_3 layer are 4, 8; the scale parameters corresponding to the conv4_3 layer are 8, 16; the scale parameters corresponding to the conv5_3 layer are 8, 16. The rest of the parameters are the same as the original Faster-RCNN.

In order to compare the computational complexity of the methods, the following settings were made based on the experimental setup for comparing the accuracy. For the ZF network, the complexity of the methods using only the conv1 feature, the conv2 feature, the conv5 feature and all three at the same time were tested separately. For the VGG network, the complexity of the method with only the conv1_2 feature, the conv2_2 feature, the conv5_3 feature and all three at the same time is tested separately. The average detection time of a single image is used as the evaluation index of the computational complexity, and *s(second)* is used as the unit.

B. THE EXPERIMENT RESULT AND ITS ANALYSIS

The target low-resolution dataset was divided into 2 parts, i.e., 7739 for evaluating the model performance and 7738 for training alone or with the target high-resolution images for training, and the detection results obtained under different feature networks are shown in TABLE 1.

TABLE 1. Comparison of modified method and conventional method.

Method	ZF	MFMS-ZF	VGG16	MFMS-VGG16
High resolution	12.2	20.7	17.2	30.4
Low resolution	50.2	57.4	54.1	58.7
Low & High resolution	46.7	48.7	50.2	51.4

The first column indicates the used network structure, and those with MFMS prefix indicate the improved model using our multi-feature and multi-scale detection. After that, the first row of each column denotes the training data. High resolution denotes the target high-resolution images obtained from the web, Low resolution denotes the 7738 target low-resolution images used for training, and the rest of the values denote the detection accuracy of the model trained under the corresponding data, respectively. The metric is mean average precision (mAP), which in this case is actually the AP of the excavator.

From TABLE 1, the following conclusions can be drawn:

(1) The detection accuracy of small targets can be effectively improved by using the method of multi-feature and multi-scale detection regardless of whether high-resolution images or low-resolution images are used as training data, which indicates that the method of multi-feature and multi-scale detection combining low-level and high-level features in deep networks is indeed feasible.

(2) The detection effect of the model using only high-resolution images as training data is poor, and the detection effect of the model using only low-resolution images as training data is better, while the performance is compromised when the two are combined, which indicates that it is not

possible to train directly using high-resolution images of targets obtained online, and the differences existing between the training set and the test set, i.e., the differences between high-resolution targets and low-resolution targets, must be resolved if we want to use this part of the data.

The average detection time of 7739 test images was used as a metric to evaluate the computational complexity, and the detection results obtained under different feature networks are shown in TABLE 2.

TABLE 2. Performance comparison of different layers.

Layer	ZF	VGG16
Conv1	0.357	0.484
Conv2	0.224	0.354
Conv5	0.072	0.081
All	0.613	0.757

The first row indicates the network structure used in the model, and then the first column of each row indicates which layer of features is used for testing, and ALL indicates that all three features are used. For better illustration, VGG's conv1_2, conv2_2 and conv5_3 are abbreviated as conv1, conv2, and conv5, respectively.

From TABLE 2, the following conclusions can be drawn:

(1) The average detection time using high-level features is less when only one feature is used for detection, which indicates that large low-level features, although suitable for small target detection, bring additional computational overhead.

(2) The increased computational overhead of using multiple features at the same time is still within an acceptable range for tasks with low real-time requirements.

The new training data is obtained by downsampling and upsampling the target high-resolution image, and the detection accuracy of the model trained with the new data is shown in TABLE 3.

The meanings of the characters in TABLE 3 are basically the same as those in TABLE 1. The two suffixes after the training data high-resolution/low-resolution indicate different combinations of downsampling and upsampling operations, and the first suffixes Max and Average denote the two downsampling methods of max pooling and average pooling, respectively, with a pooling operation window of 2 × 2 and a sliding step of 2. The second suffixes Area, Linear, and Nearest denote the three upsampling methods of area interpolation, linear interpolation, and nearest neighbor interpolation, respectively.

From TABLE 3, it can be seen that:

(1) The accuracy of the model trained from the target high-resolution image can be significantly improved by simply downsampling, which indicates that the downsampling approach can eliminate to some extent the effect of data differences between the target high-resolution image and the target low-resolution image.

TABLE 3. Performance comparison of different layers.

Method	MFMS-ZF	MFMS-VGG16
High resolution	20.7	30.4
Low resolution	57.4	58.7
Low resolution +High resolution-Max-Linear	58.5	61.4
High resolution-Max	47.2	51.9
High resolution -Max-Average	47.7	51.5
High resolution -Max-Linear	49.5	54.9
High resolution -Max-Nearest	45.7	49.5
High resolution -Average	43.3	43.4
High resolution -Average-Area	42.7	43.1
High resolution -Average-Linear	43.8	46.0
High resolution -Average-Nearest	36.7	37.9

(2) Max pooling is generally better than average pooling in the context of such a problem.

(3) The use of linear interpolation upsampling on top of downsampling can slightly improve the accuracy of the model, for reasons that cannot yet be explained theoretically, but may be due to the fact that this adds noise and prevents over-fitting to some extent.

(4) The detection accuracy of the model trained by combining the sampled transformed target high-resolution image and the target low-resolution image is higher, and instead of decreasing in accuracy like in TABLE 1, the accuracy has been improved, which shows that the sampling transformation of the target high-resolution image can indeed eliminate the effect of the difference between the target high-resolution image and the target low-resolution image data. In the case that the small target detection data is difficult to label and the training data is lacking, the amount of training data can be increased in this way simply and quickly to improve the detection accuracy. Some of the detection results are shown in Figure 14.

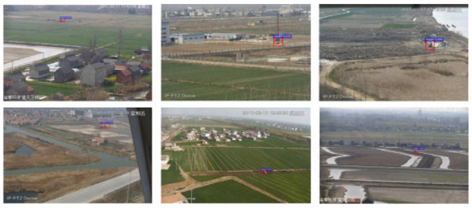


FIGURE 14. Our small target detection partial results.

Similarly, to show that the sampling transformation of the target high-resolution image can eliminate the effect of the difference between the target high-resolution image and the target low-resolution image data, we use the T-SNE [58] method to downscale the features of the target before and after the sampling operation. We perform the validation using the MFMS-VGG16 structure with the best experimental results. The network is trained with the target low-resolution image, and then tested with the target low-resolution image, the target high-resolution image and the target high-resolution image after sampling operation. The feature vectors of the same size obtained after the ROI pooling layer are downscaled by the T-SNE method. The visualization results are shown in Figure 15, where the purple dots represent the low-resolution target, the blue crosses represent the high-resolution target, and the red forks represent the high-resolution target after sampling, which shows that the feature distribution after the sampling operation is indeed closer to the feature distribution of the target low-resolution image than the feature distribution before the sampling operation.

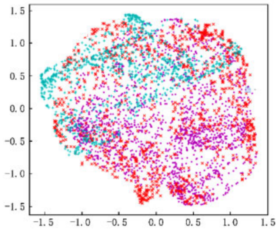


FIGURE 15. The feature distribution of processed high-resolution targets and low-resolution targets.

C. COMPARISON WITH OTHER METHODS

In order to better illustrate the novelty and superiority of our method, this section compares the proposed method with the previous classical small target detection methods. The results of the comparison are shown in TABLE 4. Currently popular small target detection algorithms based on deep learning can be divided into three categories to some extent [64]. The first category is a one-stage algorithm that uses a CNN to directly predict different target categories and positions. One-stage algorithms do not need to use candidate boxes. Instead, it transforms the problem of target frame positioning into a regression problem, so as to generate the category probability and position coordinate value of the target directly. Finally, the final detection result can be obtained directly after a single detection. This kind of algorithm is fast but less accurate, and it is shown in the shaded form in TABLE 4. The second type is the algorithm based on candidate regions and this algorithm is two-stage. It needs to first use the Region Proposal Network (RPN) to generate candidate regions, then conduct classification and regression on the candidate regions, and finally get the final detection results through two stages. This kind of method has high accuracy but slow speed, and it is shown in bold form in TABLE 4. The third type is the multi-stage algorithm, which iterates the two-stage detector and uses different detection heads and preselect boxes in the iteration process. Iteration greatly increases the size of the model, which requires greater computational support during training. Due to hardware constraints, the current three-stage algorithm has higher accuracy and slower speed due to more complex implementation process. It is shown in italics in TABLE 4.

TABLE 4. Performance comparison of different methods.

Method	mAP
Literature [65]	39.1
Literature [66]	50.7
Literature [67]	**53.6**
Literature [68]	**57.7**
Literature [69]	*60.8*
Our method	61.4
Literature [70]	*65.7*

It can be seen from TABLE 4 that the performance of the multi-stage method is better than that of the two-stage method and that of the single-stage method. The method in this paper has achieved a good score of the second place in the comparison method, which proves the effectiveness of the method in this paper. The reason why further investigation is inferior to literature [70] may lie in its higher network deepness.

V. FUTURE WORKS

Small target detection is a difficulty in target detection, and has important application value in real life. For example, in the field of criminal investigation, small packages on the table, small pedestrians in the corner of surveillance video, small marks on clothes, etc., are all clues to solve crimes. Small target detection has important research significance. For the research of small target detection, small target detection is faced with great challenges due to the few features carried by small target itself. This section points out the future research direction in view of the difficulties of small target research:

(1) Small target detection is carried out with traditional methods. Although the method based on deep learning has been the mainstream in recent years, a lot of work shows that because small targets contain little information and lack sufficient semantic information, features extracted by deep convolutional network have insufficient semantic information, but the effect is not very good for small targets. Consider to study some features that are more capable of representing small targets, and combine some non-deep learning methods for feature extraction, such as random forest and local rank of images, which may play a better role.

(2) Introducing attention mechanisms. In this paper, our multi-feature and multi-scale detection network can make good use of the feature information from the shallow layer of

the network, but the shallow feature also contains noise information from the image background. Considering the introduction of attention mechanism for detection, it can help to reduce unnecessary shallow feature information and improve the detection effect of small targets. For example, SENet [59] proposed by Jie *et al.* as an attention mechanism on channels, strengthens the features of important channels while weakens those of non-important channels, and can be flexibly embedded in various network structures to improve the effect. As a lightweight structure, it requires relatively little additional computation. In addition to the attention mechanism on the channel, there is also the attention mechanism of the spatial direction. Through the transformation of the spatial direction, the local spatial features of the target sample are easier to be learned. Compared with the channel direction, the amount of calculation is slightly increased, but higher accuracy can be obtained. It can be considered to combine the two to design the structure flexibly for small targets, so as to obtain lower calculation cost and higher precision.

(3) Minimize the interference caused by complex environment to small target detection. At present, small target detection mostly relies on specific scenes, such as military monitoring [60], aviation, sea surface [61], oil field well site [62] and other complex operation fields. In the case of complex background noise, the information of a small target will be concealed by the noise of other large objects, or it will be integrated with the background and lack obvious image contrast [63], which is also one of the factors causing the difficulty of small target detection.

(4) Research on small target detection based on anchor-free. Although the current anchor-based target detection method has been excellent and widely used in both single-phase and two-phase methods, there are still many shortcomings. Due to the method based on the anchor has a set of pre-defined scale box, lead to less sensitive to small scale target, or need special door frame of the scale of the default in view of the small target, but that very high requirements for hardware, the default frame of the scale of the more negative samples at the same time, easy to cause the imbalance of positive and negative samples which influence the training effect. Therefore, the anchor-free method should be considered for small target detection. Some recent studies have proved that the anchor-free method can achieve the same effect as the Anchor-based target detection method. The application of the anchor-free method to small target detection may also promote the research on small target detection.

(5) The model is lightweight to improve the real-time performance, accuracy and robustness of the detection system. With the development of the times, the demand for detection of small and medium-sized targets in various fields is gradually increasing. However, in the current research, in order to improve the accuracy, the models are often very redundant. For example, the addition of super-resolution modules leads to a significant increase in the amount of computation. Accuracy and robustness Therefore, how to ensure the lightweight of the model without losing accuracy will also become a research hotspot in the future.

(6) Build a more perfect small target detection data set. Although the existing VOC data set COCO data set has been widely recognized by researchers, the development of deep learning methods is always inseparable from data. Samples of small target and the data set is still inadequate, the sample equilibrium, the sample size is not enough, all hinder the development of the small target detection, so still need to consider to set up a special small target detection data sets, or to use some data to enhance way to establish a small target simulation data set, also can yet be regarded as a good way to supplement the training sample.

VI. CONCLUSION

At present, the core problem of small target detection research based on deep learning is how to improve the feature expression of small target to make it contain rich semantic information, which is also the key to improve the performance of small target detection. Compared with the detection performance of large and medium targets, there is still a big gap in the detection performance of small targets. Therefore, based on the analysis of existing target detection methods, this paper proposes a small target detection method based on deep learning with considerate feature and effectively expanded sample size. According to the characteristics of the convolutional neural network, the deep learning structure of the mainstream target detection model is modified, so that the network can use considerate features for target detection, and the accuracy of small target detection task based on low-level features is improved. At the same time, in order to avoid the over-fitting problem caused by the sample size, the data obtained from the Internet are used to train the model. Because the distribution of these training data is different from that of the task test data, two sampling methods are used to transform the high-resolution training image of the target, so that the feature distribution of the training image and the test image are more similar. Experimental results show the effectiveness of the proposed method. In addition, the related work and future work in this paper are also quite comprehensive and detailed

REFERENCES

[1] Y. Bengio, "Deep learning of representations: Looking forward," in *Proc. Int. Conf. Stat. Lang. Speech Process.* Berlin, Germany: Springer, 2013, pp. 1–37.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[3] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1–9.

[4] S. Ding and K. Zhao, "Research on daily objects detection based on deep neural network," in *Proc. IOP Conf., Mater. Sci. Eng.*, vol. 322, 2018, Art. no. 062024.

[5] K. A. Joshi and D. G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *Int. J. Soft Comput. Eng.*, vol. 2, no. 3, pp. 44–48, 2012.

[6] J. Guo and S. Gould, "Deep CNN ensemble with data augmentation for object detection," *Comput. Sci.*, to be published.

[7] J. B. Gali, P. Ray, and G. Das, "Uniformly most powerful CFAR test for Pareto-target detection in Pareto distributed clutter," in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2020, pp. 1–6.

[8] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.

[9] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates Inc., 2012, pp. 1–9.

[10] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, pp. I–I.

[12] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[13] A. Parsons, "The sharing economy: A short introduction to its political evolution," Tech. Rep., 2014.

[14] C. B. Perez and G. Olague, "Learning invariant region descriptor operators with genetic programming and the F-measure," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[15] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[16] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2241–2248.

[17] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[18] Y. D. Chen, R. F. Li, S. Y. Li, and X. Huang, "A combined grammar for object detection and pose estimation," *Chin. J. Comput.*, to be published.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[22] J. Kim and K. Grauman, "Shape sharing for object segmentation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 444–458.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[24] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," 2013, *arXiv:1312.6229*. [Online]. Available: http://arxiv.org/abs/1312.6229

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[27] Q. Chen, S. Feng, P. Xu, L. Li, L. Zheng, and J. Wang, "Scalable object detection using deep but lightweight CNN with features fusion," in *Proc. Int. Conf. Image Graph.*, 2017, pp. 374–385.

[28] S. Mathe, A. Pirinen, and C. Sminchisescu, "Reinforcement learning for visual object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2894–2902.

[29] A. Takeki, T. T. Trinh, R. Yoshihashi, R. Kawakami, M. Iida, and T. Naemura, "Combining deep features for object detection at various scales: Finding small birds in landscape images," *IPSJ Trans. Comput. Vis. Appl.*, vol. 8, no. 1, Dec. 2016.

[30] C. Chen, M. Y. Liu, O. Tuzel, and J. Xiao, "R-CNN for small object detection," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 214–230.

[31] C. Eggert, D. Zecha, S. Brehm, and R. Lienhart, "Improving small object proposals for company logo detection," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2017, pp. 167–174.

[32] M. A. Rahman and Y. Wang, "Optimizing intersection-over-union in deep neural networks for image segmentation," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2016, pp. 234–244.

[33] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 2999–3007, Feb. 2017.

[34] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[35] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 618–634.

[36] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: Single shot scale-invariant face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 192–201.

[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[38] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[39] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[40] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J. Roy. Stat. Soc. C (Appl. Statist.)*, vol. 28, no. 1, pp. 100–108, 1979.

[41] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[42] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection SNIP," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.

[43] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "FaceBoxes: A CPU real-time face detector with high accuracy," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 1–9.

[44] C. Zhu, R. Tao, K. Luu, and M. Savvides, "Seeing small faces from robust Anchor's perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5127–5136.

[45] H. Wang, J. Wang, J. Wang, M. Zhao, and W. Zhang, "GraphGAN: Graph representation learning with generative adversarial nets," *IEEE Trans. Knowl. Data Eng.*, to be published.

[46] M. Zhao, X. Liu, H. Liu, and K. K. L. Wong, "Super-resolution of cardiac magnetic resonance images using Laplacian pyramid based on generative adversarial networks," *Computerized Med. Imag. Graph.*, vol. 80, Mar. 2020, Art. no. 101698.

[47] J. Deng, G. Pang, Z. Zhang, Z. Pang, H. Yang, and G. Yang, "CGAN based facial expression recognition for human-robot interaction," *IEEE Access*, vol. 7, pp. 9848–9859, 2019.

[48] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2849–2857.

[49] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1222–1230.

[50] M. Sudha, "Traffic sign detection and recognition using RGSM and a novel feature extraction method," *Peer–Peer Netw. Appl.*, vol. 14, pp. 2026–2037, Apr. 2021.

[51] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

[52] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 206–221.

[53] A. V. Dalca, J. Guttag, and M. R. Sabuncu, "Anatomical priors in convolutional networks for unsupervised biomedical segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9290–9299.

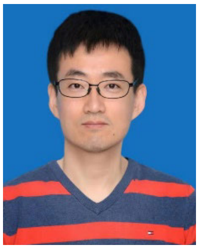
[54] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[55] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5188–5196.

[56] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, to be published.
[58] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 2605, pp. 2579–2605, Nov. 2008.
[59] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[60] R. Collins, "A system for video survellance and monitoring: VSAM final report," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep., 2000.
[61] Y. Zhou, Y. Cui, X. Xu, J. Suo, and X. Liu, "Small-floating target detection in sea clutter via visual feature classifying in the time-Doppler spectra," 2020, *arXiv:2009.04185*. [Online]. Available: https://arxiv.org/abs/2009.04185
[62] H. T. Chou, W. J. Gao, J. Zhou, B. You, and X.-H. He, "Enhancing electromagnetic backscattering responses for target detection in the near zone of near-field-focused phased array antennas," *IEEE Trans. Antennas Propag.*, vol. 69, no. 3, pp. 1658–1669, Mar. 2021.
[63] F. Gao, A. Liu, K. Liu, E. Yang, and A. Hussain, "A novel visual attention method for target detection from SAR images," *Chin. J. Aeronaut.*, vol. 32, no. 8, pp. 1946–1958, Aug. 2019.
[64] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," in *Proc. 9th Int. Conf. Adv. Comput. Inf. Technol. (ACITY)*, Dec. 2019.
[65] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[66] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[67] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020.
[68] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[69] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[70] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021.

JUN ZHANG was born in 1982. He received the B.S. degree from Northeast Normal University, Changchun, China, in 2003, and the M.S. degree from the School of Software Engineering, Tongji University, Shanghai, China, in 2009. He is currently a Lecturer with the Department of Computer Science, Tangshan Normal University, Tangshan, China. His main research interests include target detection, target tracking, and deep learning.

YIZHEN MENG received the B.E. degree from the North China University of Science and Technology, Tangshan, China, in 2003, and the M.S. degree from the School of Software Engineering, Tongji University, Shanghai, China, in 2009. She is currently a Lecturer with the Department of Computer Science, Tangshan Normal University, Tangshan. Her current research interests include computer vision and pattern recognition.



ZHIPENG CHEN received the Ph.D. degree in signal and information processing from Beijing Jiaotong University, Beijing, China, in 2019. He is currently an Associate Professor with Tangshan Normal University, China. His research interests include multimedia, signal processing, digital forensics, and data hiding.