IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 December 2022, accepted 19 December 2022, date of publication 26 December 2022, date of current version 11 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3232293

RESEARCH ARTICLE

Small Target Detection Model in Aerial Images Based on TCA-YOLOv5m

MIN HUANG[1,2], YIYAN ZHANG[2], AND YAZHOU CHEN[1], (Member, IEEE)
[1]National Key Laboratory on Electromagnetic Environment Effects, Army Engineering University, Shijiazhuang Campus, Shijiazhuang 050003, China
[2]School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China

Corresponding author: Yazhou Chen (chen_yazhou@sina.com)

This work was supported in part by the Foundation for Key Laboratories for National Defense Science and Technology under Grant 6142205210301.

ABSTRACT Target detection in aerial images taken by unmanned aerial vehicles is the most widely used scene at present. Compared with ordinary images, the background of aerial images is more complex, and the target size is smaller, which results in inferior detection precision and a high false detection rate. This paper proposes a new small target detection model TCA-YOLOv5m, which is based on YOLOv5m and combines the Transformer algorithm and the Coordinate Attention (CA) mechanism. In this model, the transformer algorithm is added to the end of the backbone of the YOLOv5, which enables the model to mine more features information of images. In the neck layer of the TCA-YOLOv5m, the Path Aggregation Network (PANet) and transformer algorithm are combined to enhance the expression capacity for the feature pyramid and improve the detection precision of occluded high-density small targets, and CA is introduced to more accurately locate targets in high-density scenes. In addition, the TCA-YOLOv5m adds a detection layer to improve the ability to capture small targets. This paper uses VisDrone 2019 as experimental data, and takes experiments to compare the detection precision and detection speed of the proposed model with baseline models. The experiment results indicate that the detection precision of the TCA-YOLOv5m reaches 97.4%, which is 5.2% higher than that of YOLOv5; the value of MAP @ 50 reaches 58.5%, which is 14.8% higher than YOLOv5. The Frames Per Second (FPS) of the TCA-YOLOv5m is 12.96 f/s, which ensures a certain real-time performance. Therefore, the TCA-YOLOv5m is suitable for the task of detecting dense small targets in aerial images.

INDEX TERMS Aerial images, small target detection, TCA-YOLOv5m, transformer algorithm, coordinate attention, path aggregation network.

I. INTRODUCTION

In recent years, deep learning has been broadly applied to target detection with the rapid development of artificial intelligence and machine vision. Unmanned driving, pedestrian detection, face recognition, and aerial images have become hot topics in target detection research. With the popularity of Unmanned Aerial Vehicles (UAV)s, Silva et al. [1] put forward a map-building and sharing framework suitable for the multi-UAV system, in which Edge computing is used to help UAVs navigate autonomously. Aerial images are usually acquired by UAVs and are widely used in military reconnaissance, environmental monitoring, smart cities, and other fields. However, small target detection in aerial images still faces tremendous challenges because it is easily influenced by external factors such as weather, light, and shooting angle. In addition, when there are small targets with high exposure and complex background in the image, the difficulty of detecting small targets will significantly increase.

Machine learning includes shallow learning and deep learning. Shallow learning only contains one or two layers of nonlinear transformation layer, and its feature extraction is relatively simple. It maps input features to the feature

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

space of a particular problem through a single-layer structure. Typical shallow learning feature structures include the Markov model [2], conditional random field [3], maximum entropy model [4], Support Vector Machine (SVM) [5], and Multilayer Perceptron (MLP) [6]. These models are relatively simple and have some limitations in the feature extraction of complex problems. Sutskever et al. [7] put forward the theory of deep learning. Deep learning simulates the mechanism of the human brain and establishes a neural network of autonomous learning on the computer. Deep learning has a deeper network structure, which can extract various data features from input information and has obvious advantages in many fields, such as the application of robot vision [8], natural language processing [9], and speech signal processing [10].

Convolutional Neural Network (CNN) [11] includes multiple convolution layers and pooling layers, and uses a sparse connection method to connect neurons between different layers. The feature information obtained from the image is more abundant through hierarchical calculation, and the output layer completes the classification and regression of the target. CNN makes the extracted features more efficient by using its local connection and parameter-sharing characteristics.

Target detection is mainly used to classify and locate the targets, commonly used in computer vision fields such as autopilot, fire prevention and control, animal protection, and so on. The traditional target detection is to in-depth analyze the feature extracted by the artificial feature detectors. The traditional inefficient method of manually designed feature detectors has been transformed into an efficient deep learning method based on a CNN. Target detection is often used for large and medium targets with sparse distribution in life and natural scenes. For example, one-stage target detection algorithms YOLO series algorithms [12], Single Shot multibox Detector (SSD) algorithm [13], and two-stage target detection algorithms Faster Region-CNN (Faster R-CNN) [14] and Mask R-CNN [15] have gained good effects in commonly used datasets. However, these characteristics of small targets in aerial images, such as large numbers, small proportions, complex backgrounds, weather changeability, and many noise, lead to low recognition accuracy and a high false detection rate.

Aerial images are usually taken from various angles overlooking. Compared with ordinary images in daily life, aerial images contain more complex spatial scenes and more types and quantities of small targets. There are two kinds of ways to define a small target. One is the definition of relative scale. That is, if the width and height of the bounding box of the target object is one-tenth of the width and height of the original image, it will be regarded as a small target. The other is from the definition of absolute pixels, resolution less than 32×32 pixels defined as small targets. Due to the complexity of the actual scenes in aerial images, small targets account for a small proportion of the images. The available feature information is less, making it more difficult to detect. Hence, the study on the automatic detection of small-sized and high-density targets in aerial images has great theoretical research value.

The YOLOv5 is a widely used model in the YOLO series, which meets the lightweight of the model design and is more conducive to environmental deployment. This paper proposes a new small target detection model TCA-YOLOv5m, which is based on YOLOv5m and combines the transformer algorithm and CA mechanism. The major contributions of this paper are as follows:

1) This paper selects, verifies, and analyzes the model through model selection, ablation, and comparative experiments. The model selection experiments compare different network size parameters and images with different resolutions. YOLOv5m is selected as the basic model, and the image resolution is 1536×1536. In the ablation experiments, the paper studies whether to add the transformer algorithm and CA mechanism. In the comparative experiments, compared with the classical models, the TCA-YOLOv5m has the best detection precision.
2) Based on the YOLOv5m, the transformer algorithm is added to the end of the backbone, which can extract more comprehensive feature information and rich context information from the input image.
3) In the neck layer, the PANet and transformer algorithm are combined to enhance the extraction ability of the feature pyramid and improve the detection precision of occluded high-density small targets. The CA mechanism is introduced to obtain the feature map with directional perception and position information by updating the multi-scale fusion feature map.
4) This paper adds another detection layer to the original three detection layers of YOLOv5m. In this way, the proposed model can enhance the capturing power of small targets to increase detection precision.

II. RELATED WORK

The process of traditional target detection mainly includes three steps: firstly, the bounding boxes are selected, then the target features are extracted. Finally, the classifier is designed. The flow of the traditional target detection process is shown in Fig. 1.

FIGURE 1. Traditional target detection process.

The first step is to select bounding boxes. Traditional target detection methods generally adopt the form of a sliding window to obtain bounding boxes. Because the proportion of the detection target in each image is different, the traditional detection randomly obtains various bounding boxes according to setting the different threshold ratios by the width and the height of the sliding window. The sliding

window intercepts different region proposals on the image through different moving positions. Due to the inferiority of sliding windows, there will be a lot of redundancy in intercepted candidate regions.

The second step is to use texture, color, or shape methods to extract features from the images intercepted by each sliding window. Standard feature extraction algorithms include the local binary pattern algorithm [16] and histogram of oriented gradient [17]. Most feature detectors are designed manually. Such feature detectors usually have poor mobility and low robustness and are only suitable for the current scene.

The third step is the design of a classifier, which classifies the detected targets. The designed classifier needs to be studied and trained in advance. Standard classifiers include SVM [18], Bayesian algorithm [19], and K-Means clustering algorithm [20]. Their working principles are to train the labeled pictures with the designed model and send the test dataset to the classifier for classification after training. The selection and use of classifiers are an essential part of the traditional detection algorithm. Choosing the appropriate classifier is vital in improving the accuracy of target detection and classification results.

Traditional target detection algorithms have many limitations, including low robustness, poor transfer ability, complex computation, slow speed, and high time complexity on small target detection tasks. Therefore, scholars introduce deep learning frameworks for target detection tasks.

In 2012, Krizhevsky et al. [21] proposed to widen and deepen the convolution layer based on CNN, introduced non-saturating neurons, and called multiple GPUs to reduce time consumption, and offered the Dropout method to prevent datasets from over-fitting and reduce false detection rate. In 2014, Girshick et al. [22] proposed R-CNN with CNN feature region for target detection and semantic segmentation. Since then, target detection based on deep learning has started an unprecedented development.

Currently, target detection is mainly divided into two categories: region proposal-based algorithms and regression-based algorithms. The method based on region proposal is also called the two-stage algorithm. In the first stage, candidate regions are generated according to the contained targets. In the second stage, network models are selected for classification according to the features extracted from bounding boxes. The algorithm based on regression is also called the one-stage algorithm. One-stage algorithm directly predicts the probability of categories, calculates the offset of input images without generating candidate regions, and realizes end-to-end network architecture.

Girshick's team proposed a two-stage algorithm R-CNN [22], target detection includes roughly three steps:

1) The region proposals are created by selecting the search method for the detection target.
2) The region proposals are adjusted to a fixed size and input to the CNN step by step. At the same time, feature vectors are extracted from the fixed length of each region's proposals.
3) All feature vectors are classified by multiple SVM.

However, the repeated calculation of region proposals leads to extremely low detection efficiency of R-CNN. In 2015, He et al. [23] proposed the Spatial Pyramid Pooling (SPP) module, which broke the limitation of inputting fixed-size pictures. At the same time, to avoid redundancy of feature extraction, feature mapping was only done once, which made a qualitative leap in detection speed. Grishick [24] proposed the Fast R-CNN algorithm using Region of Interest (ROI) pooling based on SPP-Net and R-CNN. The Fast R-CNN algorithm mainly solved the problem of repeated convolution of bounding boxes in the R-CNN, which used the softmax function [25] to calculate the class probability instead of the original SVM classifier. Although Fast R-CNN had improved the detection speed, it still used the R-CNN method in region proposal selection, and the training time was too long. Ren et al. [26] proposed the Faster R-CNN network to improve the region proposal generation method. The time for generating region proposals was shortened from 2-3s to less than 0.1 s using the region proposal network instead of the selective search algorithm, dramatically improving the algorithm's real-time performance.

The one-stage target detection algorithms directly predict the category probability and calculate the offset of the border from the input image. Therefore, the detection speed is greatly improved, but the detection precision is lower than that of the two-stage target detection algorithms. The one-stage target detection algorithms are mainly YOLO series algorithms and SSD algorithms.

In 2015, Redmon et al. [27] proposed the YOLOv1, a typical target detection method based on regression. It used a CNN as the backbone to directly predict the bounding box and the detected target's class probability. In 2016, Redmond and Farhadi [28] proposed YOLOv2 based on YOLOv1, which can detect more than 9000 targets. Compared with YOLOv1, the YOLOv2 network had tremendous changes. For example, to prevent data from over-fitting, batch normalization was added to each convolution layer to play a specific regularization effect. The mAP value increased by 2% compared with the YOLOv1. Subsequently, Redmon et al. and Bochkovskiy et al. researched YOLOv3[29] and YOLOv4[30] networks in 2018 and 2020 to achieve the balance between detection precision and speed.

Liu et al. [31] put forward the SSD, which core is multiscale feature maps. SSD uses feature maps after multiple convolution layers to locate and detect targets. When the convolution layer convolves the input image to obtain the feature map, rectangular frames with different sizes are predefined at each position of the feature map. These rectangular frames contain the position of the frames and the target detection scores. By comparing the predicted rectangular frames with the actual object rectangular frames, the best predicted rectangular frames are output, which improves the accuracy of target detection.

Zhao and Li [32] put forward a new clustering algorithm based on YOLOv3 to predict the width and height of the

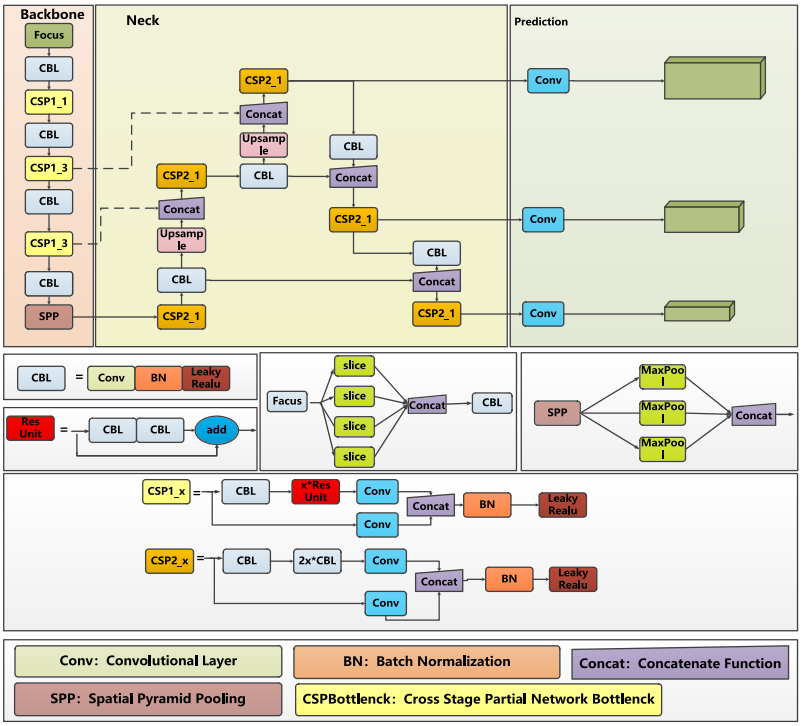

FIGURE 2. Network architecture diagram of YOLOv5m model target detection algorithm.

bounding box instead of the K-means algorithm. Compared with the YOLOv3, the new model's precision was improved by about 0.53%. Aiming at the slow speed of detecting remote sensing images with Faster R-CNN, Wang et al. [33] used a dense connection network to replace VGG Net as a backbone based on the SSD. The feature pyramid was added to the dense connection module to replace the multi-scale feature map. Compared with Faster R-CNN, the mAP value of SSD increased by 14.46%, and the time spent detecting a single image reduced by 45.7ms. Ju et al. [34] took YOLOv3 as the basic model, took eight times the output subsampled feature map and then took two times of upsampling, and spliced it with the subsampled feature map output by the residual module in Darknet 53 to form a subsampled feature fusion target detection layer. Compared with the YOLOv3, the mAP increased by 6.55%. Ye et al. [35] proposed the idea of adaptive spatial feature fusion. The deep and shallow feature maps were combined by using the semantic information of the deep feature and the feature information of the bottom edge and texture. The model used the K-Means algorithm to generate anchor boxes. Compared with the original YOLOv3, the mAP increased by 1.63%. These algorithms improve the accuracy of small target detection in aerial images to a certain extent. But under the influence of objective factors such as weather and illumination, the extracted features of the model still include a large number of redundant features related to complex backgrounds, and the detection precision still needs to be improved.

III. METHODOLOGY

A. TCA-YOLOv5m MODEL

This paper studies based on YOLOv5 and selects YOLOv5m as the basic model, as shown in Fig. 2. Because the size of small targets in aerial images is too tiny, the transformer algorithm is introduced. First, the transformer algorithm is added at the end of the YOLOv5's backbone to obtain feature maps with richer global information. Second, the transformer and PANet structure are fused in the neck layer to get the multi-scale fusion feature map. The CA mechanism is introduced to update the multi-scale fusion feature map, and the multi-scale fusion feature map with direct perception and position information is obtained, improving the model's target detection precision. Finally, a detection layer is added to the original three detection layers of the model. The detection layer is generated from the low-level, high-resolution feature maps, improving the model's ability to capture smaller targets. Adding a detection layer increases the computation and storage cost, but the detection performance is much better than that of the YOLOv5m model. The TCA-YOLOv5m is shown in Fig. 3. The specific improvement works are as follows:

1) The backbone is an essential part of YOLOv5. The function of the backbone is to extract image features from input images, which lays a foundation for the subsequent network to locate target positions and classify targets. The TCA-YOLOv5m uses Focus as the basic network and uses C3 to improve the

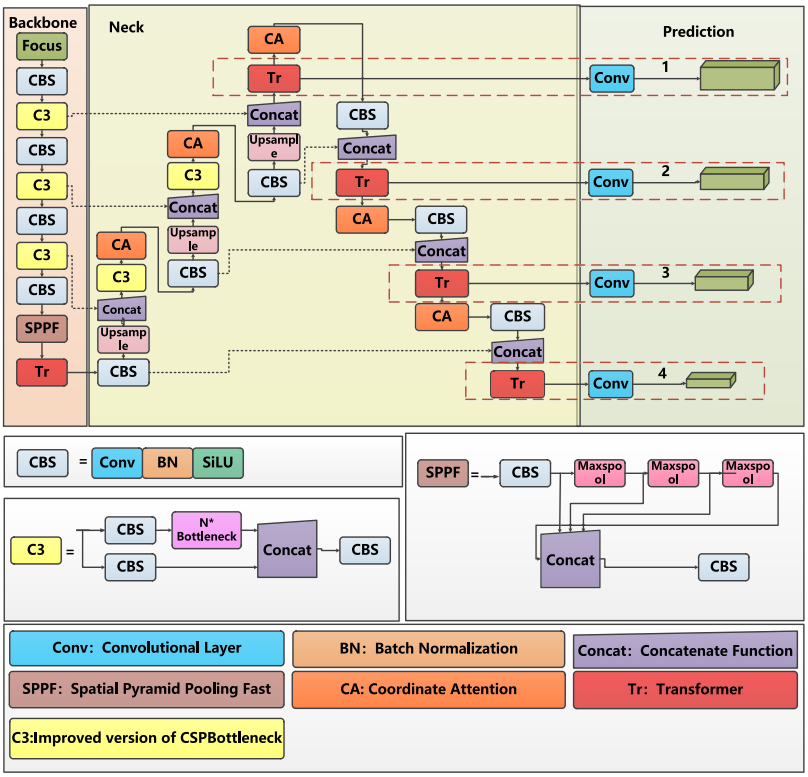

FIGURE 3. Network architecture diagram of small target detection algorithm in the aerial images of TCA-YOLOv5m.

ability of image feature extraction. Focus splits the high-resolution image into several low-resolution images and then performs the convolution operation to get the double-subsampled feature map. The function of the C3 is to enhance the learning ability of CNN and maintain accuracy while keeping the lightweight network structure. Specifically, the input channel is divided into two branches. In one branch, the convolution operation and residual error processing are performed on the image feature map first, and then the convolution operation is performed. The other branch directly convolves the image feature map. Then, the two branches are fully connected and output through the SiLU activation function. Spatial Pyramid Pooling Fast (SPPF) structure extracts and fuses the features of images through the maximum pooling and improves the receptive field of the network. The transformer algorithm is added at the end of the backbone, which extracts the image's global and rich context information.

2) In the YOLOv5 model, the neck layer adopts the multi-scale feature fusion algorithm of the PANet structure, which adds the bottom-up feature fusion layer compared with FPN. The TCA-YOLOv5m integrates the transformer algorithm and PANet structure to reduce the loss of feature information and improve target detection precision. Firstly, feature maps of different scales are extracted from the backbone. Then, the first feature fusion is realized by transverse connection with the subsampled structure, and the second feature fusion of the subsampled layer and the upsampling layer of the same scale is realized by transverse connection. Finally, the transformer algorithm processes the feature maps to obtain the multi-scale fusion feature graph with rich image feature information. In addition, the CA mechanism is used to get the multi-scale fusion feature map with directional perception and position information. It can better reduce the loss of target feature information in the process of multiple subsampled.
3) The YOLOv5 model detects targets of different sizes by using three feature maps with different scales obtained by eight times, sixteen times and thirty-two times subsampled. In the feature extraction pyramid, the receptive field with a subsampled of thirty-two times is the largest, and the larger the area of the mapped full-size image is, the more suitable it is for predicting large targets. Similarly, the subsampled of sixteen times and the subsampled of eight times are more suitable for predicting medium and small targets. The proportion of targets in aerial images is small, so the detection layer for small targets is added in the TCA-YOLOv5m. The images are processed by four times subsampled, and then sent to the feature fusion

network. This feature map has a small receptive field and rich target information. After multi-scale fusion, it can better learn target features, enhance the capture power of the network to smaller targets and improve the detection effect of targets. As shown in Fig. 3, sequence number 1 in the prediction layer is the detection layer added to the model.

B. TRANSFORMER ALGORITHM

The transformer is an algorithm for sequence-to-sequence tasks, first proposed in 2017. The transformer is the first transformation model that entirely relies on the multi-head attention mechanism to calculate input and output representation instead of using a sequence-aligned recurrent neural network (RNN) or CNN.

Multi-head attention is a mechanism to improve the performance of the attention layer. Due to the limited feature subspace, the single-head attention module will limit the ability to pay attention to multiple specific locations. The multi-head attention mechanism is realized by assigning different representation subspaces to the attention layer without affecting other attention with the same status. Different attention modules use different query vectors, key vectors, and value matrix queries.

The dot product operation between the vectors query vector and key vector generates the attention weight. Due to the significant variance of elements in the matrix in the calculation process, the Softmax function becomes very steep, which affects the gradient stability. Therefore, the scaling factor $\sqrt{\mathrm{d_k}}$ is first used to scale, then the Softmax function normalizes the attention weight, and finally, the normalized weight is assigned to the corresponding element in the value matrix vector to produce the final output vector.

The calculation process of the multi-head attention mechanism is as follows formula (1).

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

In the formula, SoftMax represents the normalized exponential function. Q, K, and V are the query, key, and value matrices. d_k represents the vector dimensions of Q and K.

For the standard transformer algorithm, the input is a patch sequence, that is, a two-dimensional matrix [num_token, token_dim], where num_token represents the number of patches and token_dim denotes the dimensions of patches. For image data, the data format [Height, Width, Channel] is a three-dimensional matrix, which the transformer encoder cannot parse, so it is necessary to transform the data through the Embedding layer first, as shown in Fig. 4.

Firstly, a picture is separated into several patches according to a given size, and then each patch is mapped to a one-dimensional vector by linear mapping. Finally, each patch data shape is flattened according to H and W dimensions by convolution operation to obtain a two-dimensional matrix. Before entering the transformer Encoder, the two-dimensional matrix must add class_token and position embedding. Insert a vector specially used for target classification, the class_token, into the previously obtained patch. This patch is a trainable parameter, and the data format is the same as other patches. It is spliced with the previously generated vector and converted into a two-dimensional matrix. The function of Position Embedding is to fix the position coding and fuse the global features of other patches, which is not based on the image content but directly superimposed on the patch.

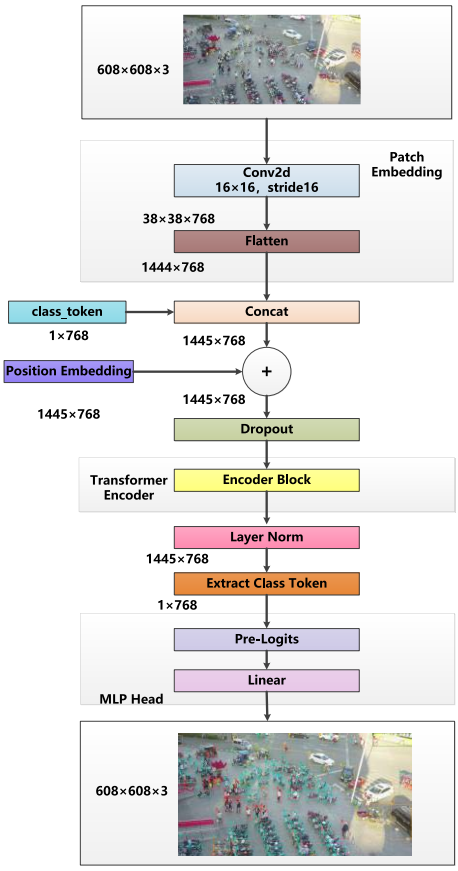

FIGURE 4. Transformer architecture diagram.

Input the patch with class_token and Position Embedding into the transformer encoder module, as shown in Fig. 5. The module consists of two sub-layers, the first is the multi-head attention mechanism, and the second is MLP Block. Apply layer norm before each sub-layer and Dropout after each sub-layer. The first sub-layer is to normalize the hidden layer. At the same time, the patch is processed by the multi-head attention mechanism, and the results are combined in sequence. The second sub-layer is to discard sub-layers randomly to prevent model over-fitting. Residual connection is used in each sub-layer to avoid the problem that the gradient disappears due to the increase of network depth. The patch is output after being repeated several times.

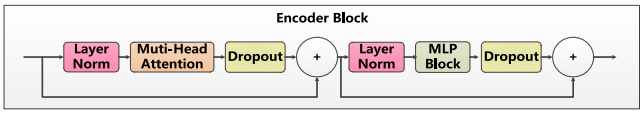

FIGURE 5. Encoder block layer.

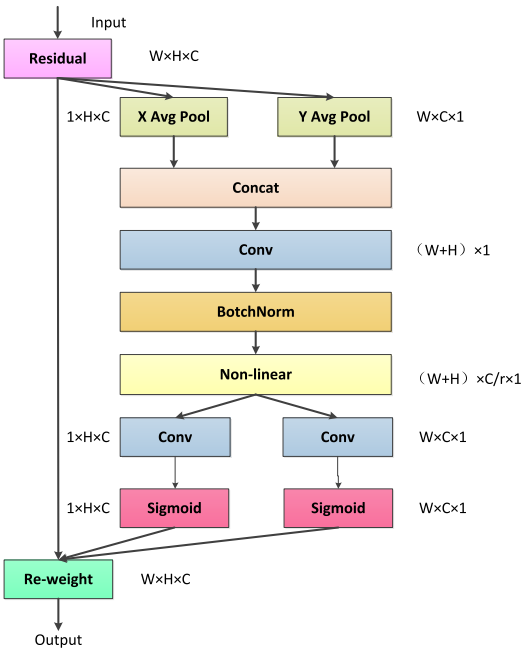

FIGURE 6. CA mechanism structure.

C. CA MECHANISM

Studies on lightweight target detection networks show that adding an attention mechanism can bring significant performance improvement to the model. Common lightweight network attention mechanisms are Squeeze-and-Excitation Net (SENet), Convolutional Block Attention Module (CBAM), and CA mechanism. The SENet attention mechanism performs attention or gating operations in the channel dimension to obtain image feature values. The CBAM extracts image feature values from the channel and spatial attention, and combines them to obtain the final image feature values.

Compared with SENet and CBAM, the CA mechanism can capture both the cross-channel information and the information on directional perception and location perception so that the detected target can be located and recognized more accurately. The CA mechanism is flexible, lightweight, and portable. The CA mechanism structure is shown in Fig. 6.

IV. EXPERIMENTS

A. EXPERIMENTAL ENVIRONMENT

All experiments are carried out on a computer with Windows 10 64-bit operating system, Intel (R) Core (TM) i9-10900X processor (CPU), and NVIDIA GeForce GTX 3080 video card (GPU). The experiments use Python 3.7 as the programming language, and the development tools use PyCharm and Anaconda 3.

B. EXPERIMENTAL DATA

The experiments use the VisDrone2019 dataset. The dataset is collected using UAVs in different scenes, weather and light conditions. It covers the landscapes of more than a dozen cities in China, including 10,000 pictures and 2.6 million annotated information. At the same time, the resolution of pictures in the VisDrone2019 dataset is as high as 2000 × 1500. The dataset includes ten small target categories: car, pedestrian, bus, bicycle, tricycle, awning-tricycle, van, truck, people, and motor. There are 6471 pictures and their labels in the training dataset, 548 pictures and their labels in the verification set, and 1610 pictures in the test set.

The models are trained, verified, and tested under the same super parameters in experiments. Among them, the training epochs is set to 50, the warmup_epochs is set to 3, and the initial learning rate is 0.01. mAP@50, mAP@75, mAP@50:95, precision, parameters, and FPS are used as the evaluation indexes of model performance. mAP@50 and mAP@75 represent the average detection precision of all target categories when IoU thresholds are 0.5 and 0.75, respectively. Among them, mAP@50 reflects the comprehensive classification ability of the algorithm for different types of targets, and mAP@75 can better reflect the regression ability of the algorithm for target bounding boxes. mAP@50:95 represents the average of the detection precision for all 10 IoU thresholds, with IoU thresholds ranging from 0.5 to 0.95 at a step size of 0.05. Generally speaking, the higher the IoU threshold, the higher the requirement for the regression ability of the model. FPS refers to how many pictures the model can detect per second, which is used to measure the real-time performance of the model. Because the resolution of UAV aerial images is high, and FPS is directly related to the resolution of the detection images. In the same model and environment, the higher the resolution of the input image, the lower the FPS. Therefore, FPS in this paper is obtained by detecting 1536 × 1536 high-resolution images. Parameters represent the number of parameters for model training.

C. MODEL SELECTION EXPERIMENTS

To use different scenarios, YOLOv5 adjusts the overall size of the network model by adjusting the network depth_multiple and width_multiple parameters. In the experiments of this paper, YOLOv5s is the smallest network proposed by YOLOv5, with the smallest number of parameters and computational complexity, and the detection speed is fast. However, the detection precision is not as good as that of YOLOv5m, YOLOv5l, and YOLOv5x with a larger network scale. Therefore, it needs to select a model appropriate for small target detection in aerial images taken by UAVs under different network scales, which should have essential detection speed and high detection precision.

In this paper, the comparative experiments before and after improvement are carried out according to the parameters of different network scales (m, l) and the parameters of input

TABLE 1. Influences of different image resolutions and different network sizes on training effect.

Model	mAP@50	mAP@75	mAP@50:95	Precision	Parameters	FPS
YOLOv5m (608)	37.1%	18.6%	20.3%	90.7%	79.688M	44.84f/s
YOLOv5m (1024)	41.3%	20%	21.4%	92.7%	79.688M	33.22f/s
YOLOv5m (1536)	43.7%	27.2%	26.3%	92.2%	79.688M	24.21f/s
TCA-YOLOv5m (608)	39.2%	18.5%	20.2%	94.4%	104.188M	19.61f/s
TCA-YOLOv5m (1024)	50.8%	30.8%	30.4%	96.2%	104.188M	16.50f/s
TCA-YOLOv5m (1536)	58.5%	35.8%	34.7%	97.4%	104.188M	12.96f/s
TCA-YOLOv5l (608)	41.6%	21.3%	22.5%	96.8%	230.352M	12.71f/s

Note: 608 means that the resolution of the input images is 608 × 608, 1024 means that the resolution is 1024 × 1024, and 1536 means that the resolution is 1536 × 1536.

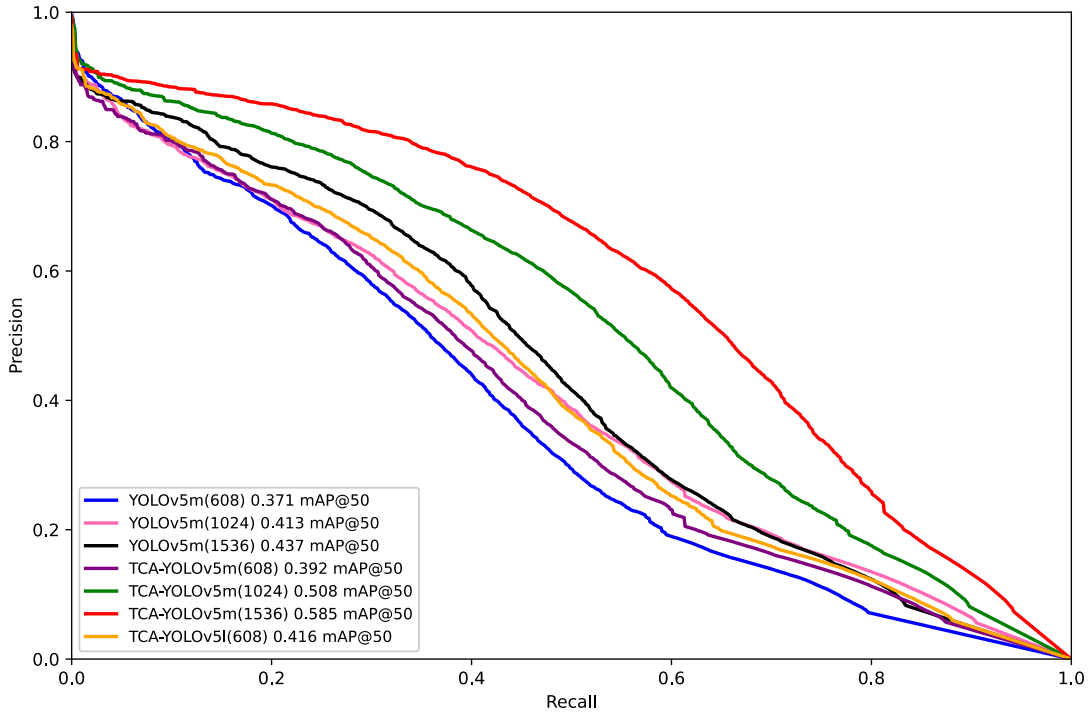

FIGURE 7. P-R curve of mAP@50 for images with different network parameters and different resolutions.

image resolution. The results of the experiment are shown in Table 1 and Fig. 7.

D. ABLATION EXPERIMENTS

In this paper, the effects of different models on small target detection performance are evaluated by ablation experiments under the same experimental conditions to prove the effectiveness of the transformer algorithm and CA mechanism. In the ablation experiments, YOLOv5m as the basic model, the resolution of the input images is 1536 × 1536, and the results after 50 epochs of training are shown in Table 2, Fig. 8.

E. COMPARATIVE EXPERIMENTS

This paper compares various popular target detection algorithms to prove the superiority of the TCA-YOLOv5m. The results are shown in Table 3. The TCA-YOLOv5m is compared with Faster R-CNN, YOLOv3, YOLOv3-SPP, YOLOv4 and YOLOv5. YOLOv3-SPP uses spatial pyramid pooling in SPPNet to fuse features with the backbone. Compared with YOLOv3, the mAP @ 50 of YOLOv3-SPP is improved by 4.7%, which shows that adding the SPPNET module can improve detection precision.

V. RESULTS AND ANALYSIS

A. ANALYSIS OF MODEL SELECTION EXPERIMENTS RESULTS

According to Table 1, Fig. 9, Fig. 10, and Fig. 11 in the model selection experiments, it can be found that the detection precision can be significantly improved when the network structure and scale have not changed and the input image resolution is high during training. However, the input of high-resolution images increases the amount of

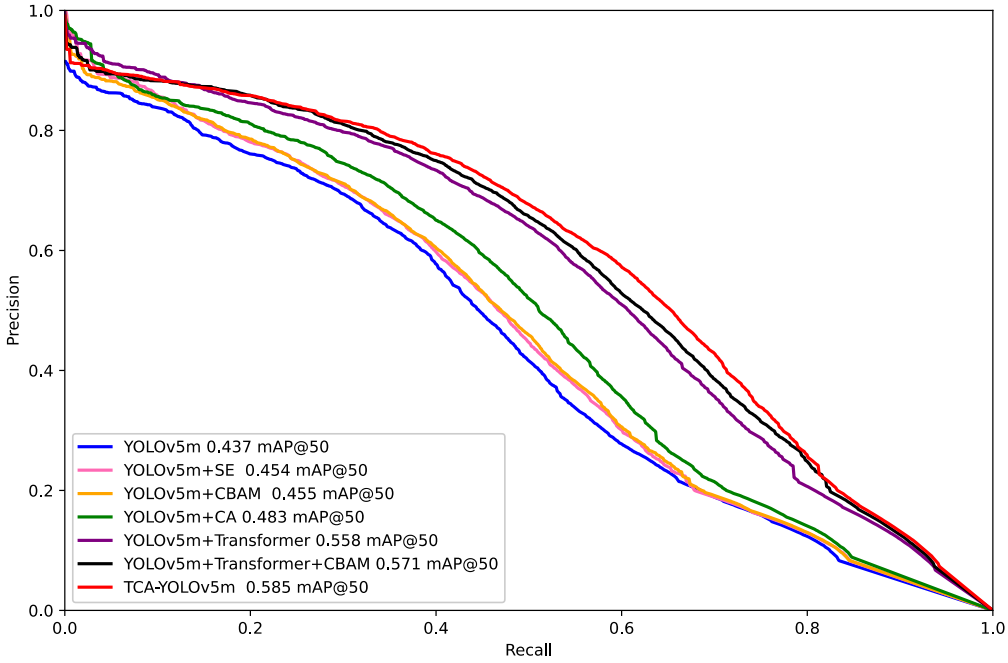

FIGURE 8. Comparison of the P-R curve of mAP@50 in each model.

TABLE 2. Results of Ablation experiments.

Model	mAP@50	mAP@75	mAP@50:95	Precision	Parameters	FPS
YOLOv5m	43.7%	27.2%	26.3%	92.2%	79.688M	24.21f/s
YOLOv5m+SE	45.4%	26.7%	26.6%	94.4%	81.5056M	33.33f/s
YOLOv5m+CBAM	45.5%	27.2%	26.8%	95.1%	81.506M	30.96f/s
YOLOv5m+CA	48.3%	29.7%	28.8%	95.8%	81.452M	32.05f/s
YOLOv5m+Transformer	55.8%	33.4%	32.8%	95.7%	76.212M	13.30f/s
YOLOv5m+Transformer+CBAM	57.1%	34.3%	33.5%	95.4%	104.26M	11.99f/s
TCA-YOLOv5m	58.5%	35.8%	34.7%	97.4%	104.188M	12.96f/s

TABLE 3. Comparative experiments of different detection algorithms.

Model	mAP@50	mAP@50:95	FPS
Faster R-CNN	18.3%	9.7%	11.06f/s
YOLOv3	33.7%	16.7%	8.16f/s
YOLOv3-SPP	38.4%	22%	9.458f/s
YOLOv4	27.9%	16.4%	7.54f/s
YOLOv5	43.7%	26.3%	24.213f/s
TCA-YOLOv5m	58.5%	34.7%	12.96f/s

computation, resulting in a significantly reduced computation speed. When the resolution of the input image is 1536 in the TCA-YOLOv5m, it takes about 34 minutes and 44 seconds to train an epoch in this experimental environment. At the same time, FPS is reduced to 12.96 f/s, about 1.5 times the input image resolution of 608. The real-time performance of the model with the larger network scale is poor because of many parameters and the calculation amount. For example, the number of parameters in the TCA-YOLOv5l model (608) model reaches 230.352 M, which is more than twice as large as the TCA-YOLOv5m (1536) number of parameters. The FPS of the TCA-YOLOv5l model (608) decreases by 6.9 f/s compared with that of the TCA-YOLOv5m (608), reaching 12.71 f/s. These factors may not meet the demands of the real environment. In the practical application scenario, the real-time performance of this model is significantly decreased due to the large network scale of l and x. Therefore, this paper considers that the TCA-YOLOv5m has low numbers of parameters and computation and has certain real-time performance and better detection precision, which can better meet the requirements of the experiments.

B. ANALYSIS OF ABLATION EXPERIMENTS RESULTS

According to Table 2, Fig. 12, and Fig. 13 in the ablation experiments, it can be concluded that compared with the YOLOv5m+CBAM model, the YOLOv5m + CA model has better performance, mAP@50, mAP@75, and mAP@50:95 improves by 2.8%, 2.5%, and 2%, respectively, and the number of network parameters reduces by 0.054M, and the FPS increases by 1.09f/s. Compared with the YOLOv5m,

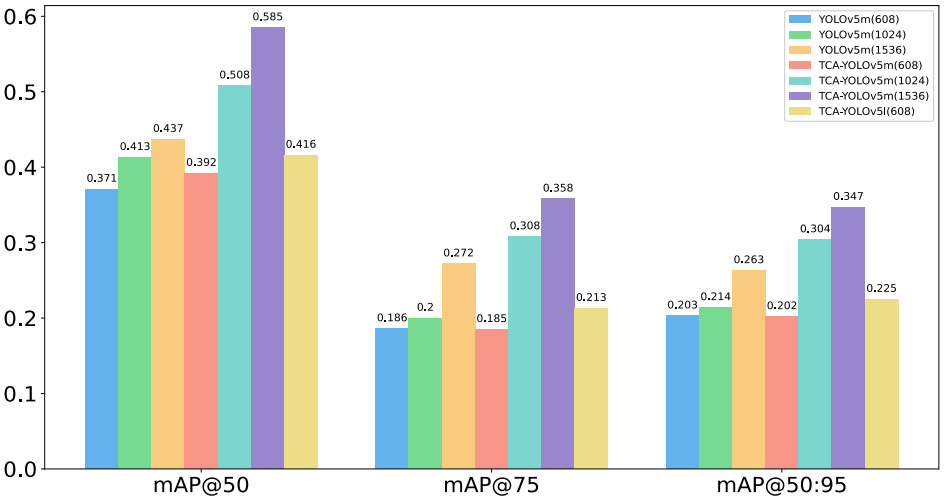

FIGURE 9. Comparison of mAP values of images under different network parameters and different resolutions.

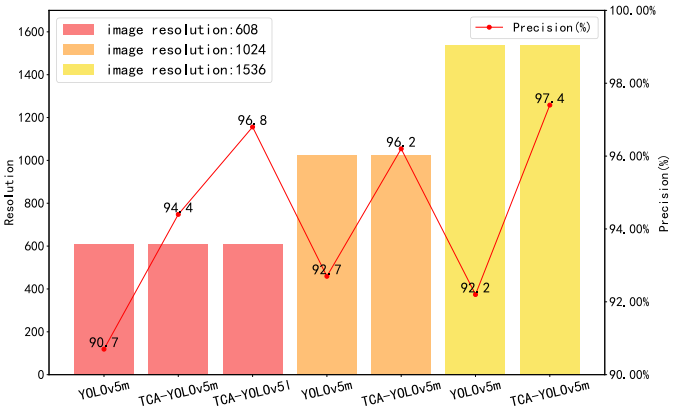

FIGURE 10. Comparison of precision values of images with different resolutions input by different models.

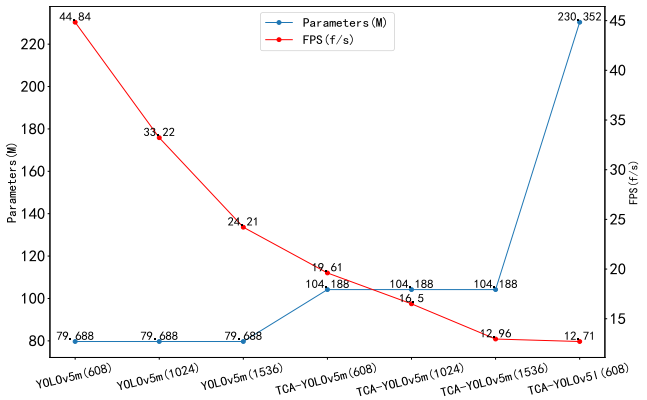

FIGURE 11. Comparison of parameter quantity and FPS of images under different network parameters and different resolutions.

the YOLOv5m+Tr model's mAP@50 is improved by 12.1% and reaches 55.8%, and the number of parameters reduces by 3.476M. TCA-YOLOv5m mAP values increase by 14.8%, 8.6%, and 8.4%, respectively. It can be concluded that the TCA-YOLOv5m is effective in detecting small targets in aerial images.

C. ANALYSIS OF COMPARATIVE EXPERIMENTAL RESULTS

On the basis of the comparative experimental results, it is concluded that the detection precision of the TCA-YOLOv5m is better than that of other algorithms under the condition of ensuring certain real-time performance, as shown in Table 3 and Fig. 14. The mAP@50 of the TCA-YOLOv5m increases by 30.6% compared with YOLOv4 and 24.8% higher than YOLOv3, reaching 58.5%. The mAP@50 of the Faster R-CNN model is only 18.3%. The reasons are that there are a large number of dense small targets in the experimental data, the background is complex, and the extracted target feature information is too little, which leads

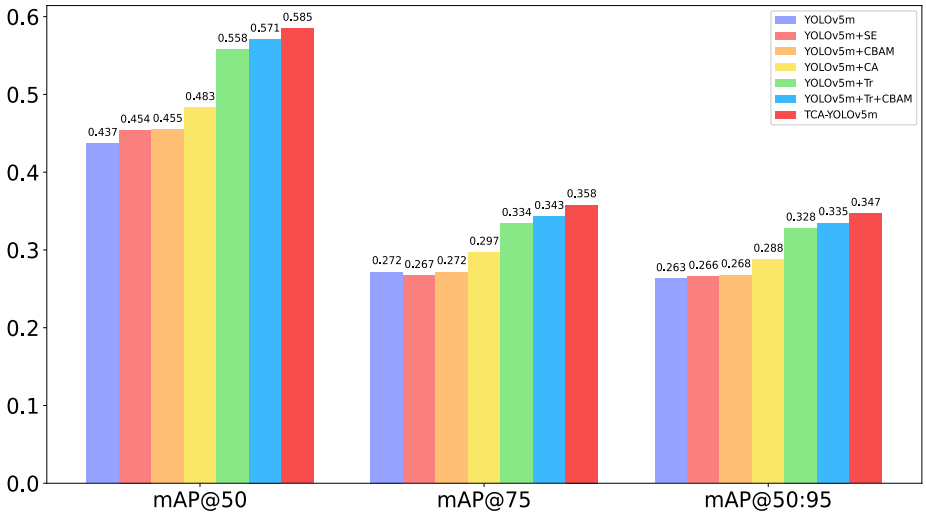

FIGURE 12. Comparison of values of mAP @ 50, mAP @ 75, and mAP @ 50: 95 for different network models.

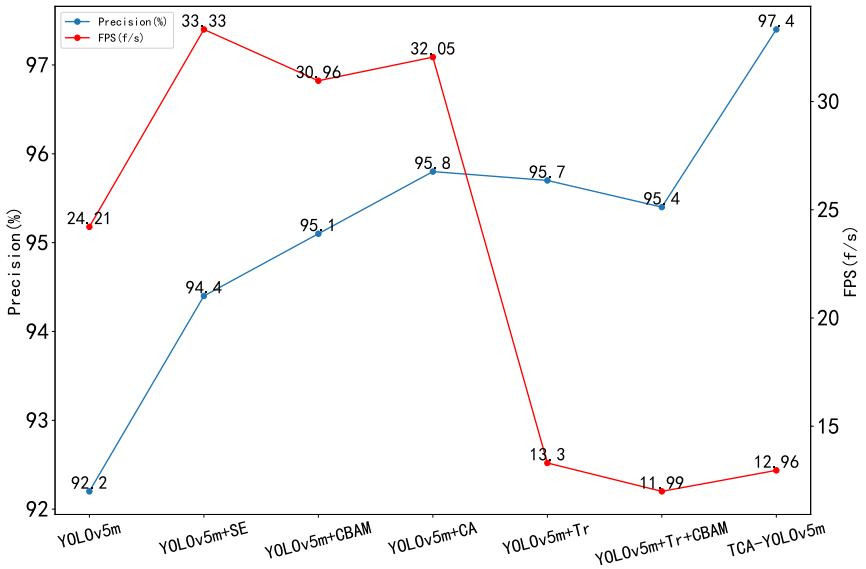

FIGURE 13. Comparison of accuracy and real-time performance of different network models.

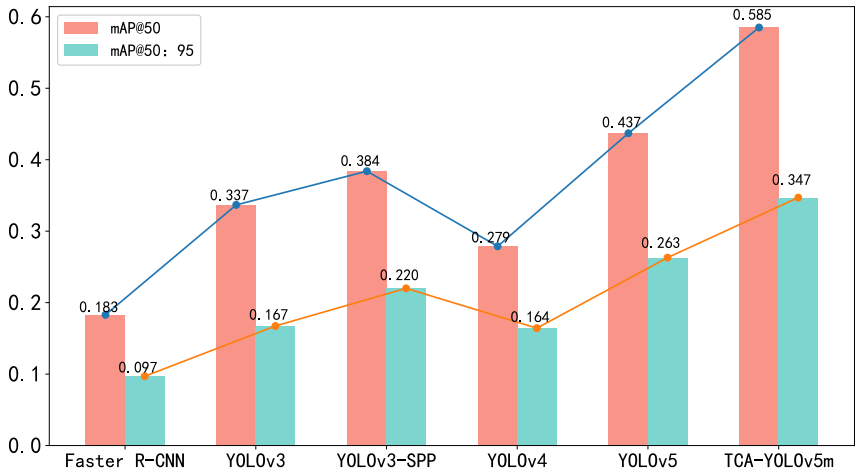

FIGURE 14. Comparison of mAP values between the classical models and TCA-YOLOv5m.

FIGURE 15. Detection Effect Drawing under Light.

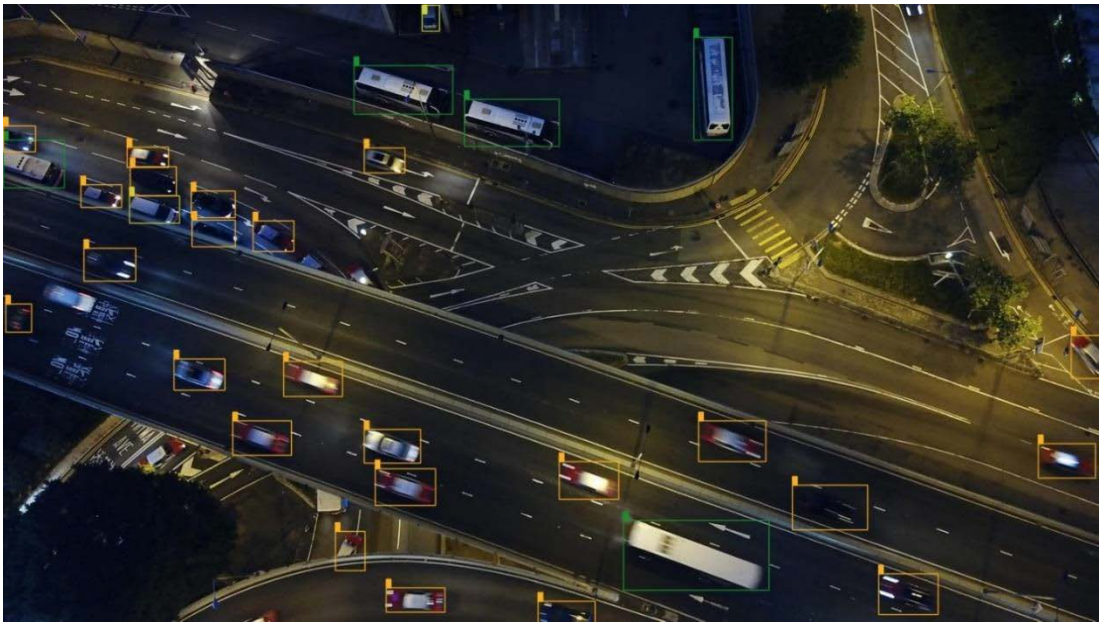

FIGURE 16. Blur distortion detection effect diagram.

to low detection results. This paper proposes to add the transformer algorithm at the end of the backbone of the YOLOv5 model to obtain more image feature information. In the neck layer, the transformer algorithm is integrated into the PANet to enhance the expression ability of the feature pyramid. The CA mechanism is added to obtain feature maps with directional perception and position perception information. This model adds a detection layer to focus more attention on dense small targets and improve the ability of feature extraction of small targets. The TCA-YOLOv5m is more advantageous in small target detection of aerial images.

D. ANALYSIS OF MODEL VISUALIZATION

The TCA-YOLOv5m verifies the effectiveness of small target detection in the actual scene by detecting representative and complex images in the VisDrone2019 dataset. In this paper, the detection results of all categories of small targets are evaluated and visually analyzed.

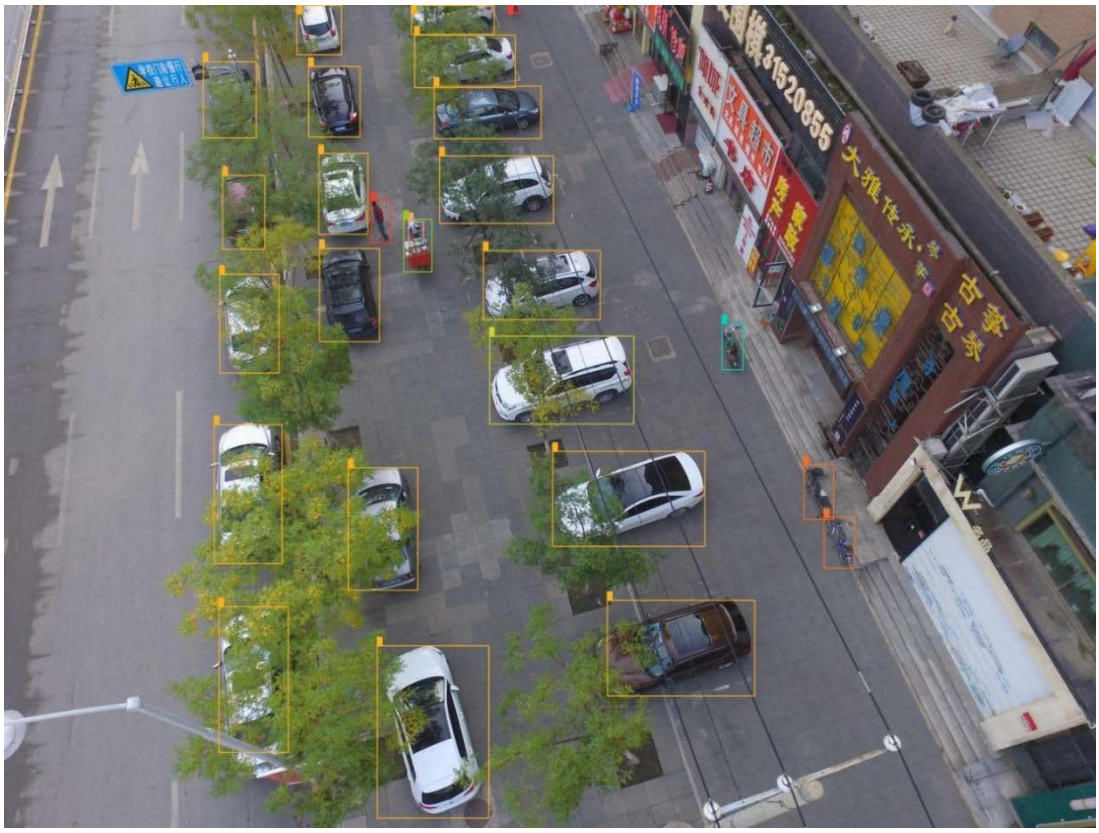

FIGURE 17. Detection effect diagram under target occlusion.

FIGURE 18. Effect diagram of small target detection at high altitude.

In the case of different brightness, the pictures taken include dense targets in dim and sufficient light. The detection results of the TCA-YOLOv5m for such small targets are shown in Fig. 15. When the camera rotates too fast, the pictures taken may be fuzzy. The TCA-YOLOv5m can achieve an accurate detection effect for small targets with

fuzzy, as shown in Fig. 16. The small target features are occluded in the photographed pictures when there are too many shelters. The TCA-YOLOv5m can still accurately detect small targets such as occluded cars, pedestrians, and bicycles, as shown in Fig. 17. In the case of high-altitude shooting, the detected targets have the characteristics of dense distribution and tiny size, and the TCA-YOLOv5m can accurately detect the small targets with such characteristics, as shown in Fig. 18. From the above four kinds of detection results, it can be concluded that this model has outstanding detection ability when dealing with small targets with different characteristics.

VI. CONCLUSION

In this paper, a new model TCA-YOLOv5m for small target detection is proposed, which is intended to improve the detection precision in small and dense scenes to be suitable for more accurate aerial small target detection tasks of UAVs.

This paper first adds the transformer to the end of YOLOv5m's backbone to obtain a feature map with richer global information. Secondly, in the neck layer, the transformer and PANet are integrated to enhance the expression capability of the feature pyramid. It improves the learning ability of the whole feature and the detection precision of the occluded high-density small target. In addition, the CA mechanism is introduced to obtain a multi-scale feature fusion map with directional perception and position information, which improves the accuracy of the model for target detection. Finally, there are a lot of spatial background factors in aerial images taken by UAVs, so a detection layer is added to the original three detection layers of YOLOv5. The TCA-YOLOv5m can enhance small targets' capturing power and reduce the false detection rate caused by too significant size differences of detected targets.

The results show that the value of mAP increases significantly for images with input resolution from low to high. When the resolution of the input image is 1536, the mAP@50 of the TCA-YOLOv5m is 58.5%, which is 14.8% higher than that of the original YOLOv5m. Compared with the Faster R-CNN, the mAP of the TCA-YOLOv5m increases by 40.2%. Therefore, the TCA-YOLOv5m has good robustness and anti-jamming capability and effectively improves the target detection ability. However, the TCA-YOLOv5m still has the phenomenon of missing detection and false detection for some tiny detection targets. Future work will continue to optimize the model to improve the detection results of small targets, and study how to achieve a lightweight network model while ensuring the detection accuracy.

REFERENCES

[1] A. Silva, M. Basso, P. Mendes, D. Rosario, E. Cerqueira, B. J. G. Praciano, J. P. J. Da Costa, and E. P. De Freitas, "A map building and sharing framework for multiple UAV systems," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, Dubrovnik, Croatia, Jun. 2022, pp. 1333–1342, doi: 10.1109/ICUAS54217.2022.9836075.

[2] D. Zou, Y. Shi, W. Su, and G. Xuan, "Steganalysis based on Markov model of thresholded prediction-error image," in *Proc. IEEE Int. Conf. Multimedia Expo*, Toronto, ONT, Canada, Jul. 2006, pp. 1365–1368, doi: 10.1109/ICME.2006.262792.

[3] Y. Zhu, S. Pan, and J. Chai, "Modeling hierarchical and heterogeneous feature representation with conditional random field for visual object detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 2069–2072, doi: 10.1109/ICASSP.2012.6288317.

[4] C. H. Seng, M. G. Amin, F. Ahmad, and A. Bouzerdoum, "Image segmentations for through-the-wall radar target detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 3, pp. 1869–1896, Jul. 2013, doi: 10.1109/TAES.2013.6558025.

[5] K. P. Bennett and E. J. Bredensteiner, "Duality and geometry in SVM classifiers," in *Proc. ICML*, San Mateo, CA, USA, Jun. 2000, pp. 57–64.

[6] Y. Tang, K. Han, J. Guo, C. Xu, Y. Li, C. Xu, and Y. Wang, "An image patch is a wave: Phase-aware vision MLP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 10925–10934, doi: 10.1109/CVPR52688.2022.01066.

[7] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, Feb. 2013, pp. 1139–1147.

[8] F. Zhang, J. Leitner, M. Milford, B. Upcroft, and P. Corke, "Towards vision-based deep reinforcement learning for robotic motion control," 2015, *arXiv:1511.03791*.

[9] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [review article]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018, doi: 10.1109/MCI.2018.2840738.

[10] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Commun.*, vol. 50, no. 10, pp. 782–796, Oct. 2008, doi: 10.1016/j.specom.2008.04.010.

[11] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*.

[12] L. Aziz, M. S. B. H. Salam, U. U. Sheikh, and S. Ayub, "Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review," *IEEE Access*, vol. 8, pp. 170461–170495, 2020, doi: 10.1109/ACCESS.2020.3021508.

[13] S. Zhai, D. Shang, S. Wang, and S. Dong, "DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion," *IEEE Access*, vol. 8, pp. 24344–24357, 2020, doi: 10.1109/ACCESS.2020.2971026.

[14] Y. Ren, C. Zhu, and S. Xiao, "Object detection based on fast/faster RCNN employing fully convolutional architectures," *Math. Problems Eng.*, vol. 2018, pp. 1–7, Jan. 2018, doi: 10.1155/2018/3598316.

[15] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[16] K. Lee, C. Lee, S. A. Kim, and Y. H. Kim, "Fast object detection based on color histograms and local binary patterns," in *Proc. IEEE Reg. Conf. (TENCON)*, Cebu, Philippines, Nov. 2012, pp. 1–4, doi: 10.1109/TENCON.2012.6412323.

[17] S. Zhang and X. Wang, "Human detection and object tracking based on histograms of oriented gradients," in *Proc. 9th Int. Conf. Natural Comput. (ICNC)*, Shenyang, China, Jul. 2013, pp. 1349–1353, doi: 10.1109/ICNC.2013.6818189.

[18] P. Li, L. Dong, H. Xiao, and M. Xu, "A cloud image detection method based on SVM vector machine," *Neurocomputing*, vol. 169, no. 2, pp. 34–42, Dec. 2015, doi: 10.1016/j.neucom.2014.09.102.

[19] T. Aach and A. Kaup, "Bayesian algorithms for adaptive change detection in image sequences using Markov random fields," *Singel Process-Image*, vol. 7, no. 2, pp. 147–160, Aug. 1995, doi: 10.1016/0923-5965(95)00003-F.

[20] Y. Zheng, X. Zhang, B. Hou, and G. Liu, "Using combined difference image and K-means clustering for SAR image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 3, pp. 691–695, Mar. 2014, doi: 10.1109/LGRS.2013.2275738.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105, doi: 10.1145/3065386.

[22] 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.*, Columbus, OH, USA, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[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, Jul. 2015, doi: 10.1109/TPAMI.2015.2389824.
[24] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.
[25] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi, "Spoken language understanding using long short-term memory neural networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, South Lake Tahoe, NV, USA, Dec. 2014, pp. 189–194, doi: 10.1109/SLT.2014.7078572.
[26] 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, doi: 10.1109/TPAMI.2016.2577031.
[27] 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)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.
[28] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.
[29] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[30] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. Berg, "SSD: Single shot multibox detector," in *Computer Vision—ECCV* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2016, pp. 21–37.
[32] L. Q. Zhao and S. Y. Li, "Object detection algorithm based on improved YOLOv3," *Electronics*, vol. 9, no. 3, pp. 1–11, Mar. 2020, doi: 10.3390/electronics9030537.
[33] J. Q. Wang, J. S. Li, X. W. Zhou, and X. Zhang, "Improved SSD algorithm and its performance analysis of small target detection in remote sensing images," *Acta Optica Sinica*, vol. 39, no. 6, pp. 373–382, Jun. 2019, doi: 10.3788/AOS201939.0628005.
[34] M. R. Ju, H. B. Luo, Z. B. Wang, M. He, Z. Chang, and B. Hui, "Improved YOLO V3 algorithm and its application in small target detection," *Acta Optica Sinica*, vol. 39, no. 7, pp. 253–260, Jul. 2019, doi: 10.3788/AOS201939.0715004.
[35] K. Ye, Z. Fang, X. Huang, X. Ma, J. Ji, Q. Wu, and Y. Xie, "Research on small target detection algorithm based on improved YOLOv3," in *Proc. 5th Int. Conf. Mech., Control Comput. Eng. (ICMCCE)*, Harbin, China, Dec. 2020, pp. 1467–1470, doi: 10.1109/ICMCCE51767.2020.00321.

MIN HUANG received the M.Eng. degree in software engineering from the Beijing Institute of Technology, Beijing, China, in 2004. He is currently pursuing the Ph.D. degree in electrical engineering with the National Key Laboratory on Electromagnetic Environment Effects, Shijiazhuang, Hebei, China. He is currently an Associate Professor with the Hebei University of Science and Technology, Shijiazhuang. His research interests include machine learning, natural language processing, and artificial intelligence.

YIYAN ZHANG is currently pursuing the master's degree with the Hebei University of Science and Technology. Her research interests include machine vision and deep learning.

YAZHOU CHEN (Member, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees from the Shijiazhuang Mechanical Engineering College, Shijiazhuang, China, in 1996, 1999, and 2002, respectively. He is currently working as a Professor and the Director of the Electromagnetic Environment Effects Key Laboratory, China. His research interests include EMC and lightning protection.