Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 5384844, 17 pages
https://doi.org/10.1155/2023/5384844

Research Article

Accurate Detection and Tracking of Small-Scale Vehicles in High-Altitude Unmanned Aerial Vehicle Bird-View Imagery

Heshan Zhang, Xin Tan, Mengwei Fan, Cunshu Pan, Zhanji Zheng, Shuang Luo, and Jin Xu

College of Traffic and Transportation, Chongqing Jiaotong University, Chongqing 400074, China

Correspondence should be addressed to Jin Xu; yhnl_996699@163.com

Received 21 August 2023; Revised 3 December 2023; Accepted 13 December 2023; Published 30 December 2023

Academic Editor: Qixiu Cheng

Vehicle detection and tracking from unmanned aerial vehicles (UAVs) aerial images are among the main tasks of intelligent traffic systems. Especially in tasks with long distances, extensive backgrounds, and small objects, it increases the difficulty of localization and regression, which can easily lead to missed detections and false positives. This paper proposes a detection-based small-scale vehicle tracking framework that integrates an improved YOLOX network and the DeepSORT algorithm to address these issues. Based on the original YOLOX network, a shallow feature extraction network, 160×160 pixels, is added to enhance the ability to extract small-scale object features. A convolutional block attention module (CBAM) is inserted in front of the neck network to select crucial information for vehicle detection tasks while suppressing noncritical ones. EIoU_Loss is introduced as the bounding box regression loss function in training to speed up their convergence and improve the localization accuracy of the small objects. Furthermore, an image segmentation method is proposed to effectively reduce missed and false detection events. It divides the original high-definition image into multiple subimages, first detected and then reassembled. Finally, the improved YOLOX network is used as the detector of the DeepSORT to perform small-scale vehicle detection and tracking tasks in various traffic scenarios. Experiments show that the proposed method can significantly improve the detection accuracy of the network and effectively solve the problems of missed detection and false positives in small-scale vehicle tracking tasks in high-resolution aerial images captured by high-altitude UAVs. Significantly, the algorithm proposed in this paper has sufficient robustness for small-scale tracking tasks of aerial videos captured at different altitudes.

1. Introduction

With the rapid development of unmanned aerial vehicle (UAV) technology, computer vision is one of the critical topics in autonomous robots, including UAVs [1, 2]. It plays an essential role in various fields, such as road traffic monitoring [3, 4], military reconnaissance [5], postdisaster search and rescue [6], and environmental monitoring [7, 8]. UAV-based traffic monitoring has significant advantages over traditional ground-based fixed surveillance cameras, as UAVs have higher maneuverability, wider field of view, and no interference with observed traffic [9, 10]. In recent years, the continuous development of new algorithms such as machine vision has supported extracting high-precision vehicle trajectory data from aerial videos, which can support research such as traffic flow feature analysis, traffic management, driving behavior analysis, and road safety evaluation [11, 12]. However, due to the low pixel ratio, high density, shadows, and blurred borders of small-scale vehicles in high-altitude UAV images, it is easy to cause missed detections and false positives, which poses significant challenges to vehicle detection and tracking tasks [13–15].

The detection-free tracking (DFT) framework based on the initial frame can only track the marked target in the first frame. However, it cannot track new targets that appear in the subsequent frames, limiting the DFT's application scenarios [16]. With the development of deep learning algorithms, tracking by detection (TBD) frameworks have become a research hotspot due to their higher flexibility and robustness. The TBD requires the detector to output the

detection results of each image frame and then perform the tracking task. The tracking performance greatly depends on the quality of the detector. Simple online and real-time tracking (SORT) is a typical multiobject tracking algorithm based on the TBD framework. The algorithm predicts and updates the target's current position through the Kalman filter algorithm and then uses the Hungarian algorithm to match the detection and tracking boxes. However, the SORT algorithm ignores the long-term and short-term occlusion of the targets, resulting in frequent ID-switching. Therefore, the DeepSORT algorithm introduces some tricks, such as deep learning models, appearance features, and cascaded matching, to improve the algorithm's accuracy. For the TBD framework, vehicle detection is a prerequisite for multitarget tracking and efficient and accurate object detection algorithms can improve the accuracy of tracking algorithms.

Recently, deep learning technology has made significant progress in vehicle detection, especially with the emergence of the convolutional neural network (CNN), which can extract relevant features automatically. Most deep learning-based vehicle detection research uses CNN architecture as the backbone. Deep learning-based object detection methods are divided into two categories: (i) two-stage object detectors based on proposal regions and (ii) one-stage object detectors based on end-to-end frameworks. The two-stage object detector mainly includes R-CNN [17] and its variants, including spatial pyramid pooling networks (SPPNet) [18], Fast R-CNN [19], Faster R-CNN [20], Mask R-CNN [21], and others. Their main idea is first to generate many proposals and then perform classification and regression on each proposal. In general, two-stage object detection methods based on candidate regions have high detection accuracy. Nevertheless, the high computational complexity requires much time for reasoning, resulting in inefficiency, thus making them unsuitable for real-time object detection, especially on mobile devices. On the other hand, UAVs must detect all objects in their field of view, such as vehicles and pedestrians, with high speed.

From R-CNN to Faster R-CNN and the improved models proposed by subsequent scholars, object detection algorithms have gradually evolved from integrating various independent modules, i.e., region proposal (RP), feature extraction, and bounding box prediction, to being unified and integrated into an end-to-end one-stage detection network. Compared with the two-stage object detector, the one-stage object detector directly predicts the position and category of the target, abandoning the cumbersome RP stage, and has a faster detection speed. Therefore, it has attracted the attention and further research of many scholars. One-stage object detection algorithms mainly include you only look once (YOLO) [22–24], single-shot multibox detector (SSD) [10], and RetinaNet [25]. After years of development, large- and medium-sized object detection networks have matured relatively. However, small object detection network research still has much room for improvement.

UAVs can fly at different altitudes with different perspectives, which pose many challenges compared to vehicle detection in regular perspective images [26]. Vehicle detection based on UAV images is still a challenging problem due to many issues, such as but not limited to, small pixel ratio, high density, scale variation, complex background, similar appearance of vehicles to other object types, partial occlusion of vehicles, and others. Therefore, most detectors cannot be directly used to detect small-scale objects in UAV aerial images. In order to deal with these challenges, scholars have proposed some improved tricks, including training strategies, loss function, attention mechanisms, and others. Xu et al. [27] applied Faster R-CNN to car detection in low-altitude UAVs images. However, due to the relatively rough feature maps, the RPN in Faster R-CNN has poor localization performance for small-scale vehicles. In addition, the classifier after RPN cannot distinguish vehicles from complex backgrounds well. Therefore, Tang et al. [28] proposed an improved detection method based on Faster R-CNN, introducing a hyper region proposal network (HRPN) combined with hierarchical feature maps to extract car-like targets and then replacing the original classifier with a cascaded enhanced classifier. The accuracy and robustness of the algorithm are improved. Han et al. [29] proposed a CNN-based vehicle detector from UAV bird's-eye view image, namely, DRFBNet300. This method has a deeper receptive field block (DRFB), which can enhance the expressive power of feature maps to detect small-scale vehicles in UAV aerial images.

Several challenges limit the application of the R-CNN to object detection in aerial imagery. For example, the size of vehicles in large-scale aerial images is relatively small, and the localization performance of the R-CNN network for small objects is poor. R-CNN is specifically designed to detect bounding boxes of objects rather than extract attributes. YOLO, on the other hand, trains on complete images, directly improving detection performance. Zhang et al. [30] proposed a depthwise-separable attention-guided network (DAGN) based on YOLOv3. First, the attention block was combined with the feature connection, and then the cross-entry loss function was changed to focal loss. Finally, Gaussian non-maximum suppression (NMS) generated new confidence scores for nonoptimal candidates. Luo et al. [31] proposed an improved YOLOv3 network for high-density vehicle detection in parking lots from UAV images, using the K-means++ algorithm to improve the selection of the initial recognition box and increase the AP value of the network. In addition, the application of soft-NMS solved the missed detection of some highly overlapping targets caused by NMS, and the missed detection rate was significantly reduced. Feng et al. [32] proposed a new trajectory extraction framework for mixed traffic flow. The framework integrates the YOLOv3 vehicle detector, image registration method using Shi-Tomasi corner detection, trajectory construction based on correlation compensation, and trajectory denoising of ensemble empirical mode decomposition (EEMD). The framework can extract the trajectories of most road users at urban intersections. Yu et al. [15] proposed a single-stage detector SF-SSD based on a spatial cognition algorithm. The deconvolution operation was introduced into the feature fusion module to enhance the representation of shallow features and effectively improve the detection accuracy of UAV small-scale objects.

In general, these detectors have significantly improved vehicle detection performance in low-altitude UAVs' aerial images. However, there is still a lack of relevant research on small-scale vehicle detection from high-altitude (i.e., over 120 meters) aerial images. Vehicle detection and tracking technology in high-altitude scenarios achieve a "holographic collection" of traffic information in both temporal and spatial dimensions. The extracted high-precision trajectory data can be used to analyze traffic flow characteristics, evaluate road accident risk characteristics, and provide data support for intelligent traffic control, alleviating traffic congestion, and accident analysis. When UAVs fly at a very high altitude (i.e., over 300 meters), the proportion of object pixels in high-resolution aerial images is tiny, approximately 40×19 pixels, with fewer visual features, blurred boundaries, and complex background information. The down-sampling operation of deep CNN networks poses a risk of feature disappearance, making it difficult to extract more representative features and increasing the difficulty of positioning and regression. It is more likely to cause missed detections and false positives, especially in detection tasks with densely distributed vehicles.

This paper proposes a TBD framework for small-scale vehicles in high-altitude UAV aerial images, integrating the improved YOLOX detection network and the DeepSORT tracking algorithms. Given the insufficient performance of the YOLOX network for small object detection, the YOLOX network is optimized from three aspects. First, to enhance the network's ability to extract small-scale object features, a 160×160 pixels feature extraction network is added to the three output layers of the original YOLOX. Second, a convolutional block attention module (CBAM) is embedded before the neck network to suppress the expression of redundant features. Finally, the EIoU_Loss function is used as the bounding box regression loss function in training, focusing on the overlap between the ground truth and the bounding box, the center point distance, and the aspect ratio to speed up the bounding box regression and improving the positioning accuracy of the network. In particular, we also propose a sliding window-based image segmentation method (ISM) to suppress missed detections and false positives. The original sizeable high-definition image is divided into several small-sized subimages, which are detected first and then restored to the original large image. Finally, the improved YOLOX detector is embedded into the DeepSORT tracking algorithm, and tracking tasks are performed in several traffic scenarios to verify the effectiveness of the method proposed in this paper. The main contribution of this paper is to propose an improved YOLOX detector and integrate it with the DeepSORT tracking algorithm to achieve vehicle detection and tracking in high-altitude scenarios, effectively solving the problem of missed and false detections of small-scale vehicles in high-altitude scenarios, and providing data support for intelligent transportation systems such as traffic flow analysis, driving behaviors, road safety evaluation, and traffic management.

2. Proposed Method

This paper proposes a detection-based framework for small-scale vehicle tracking in high-definition images, with an improved YOLOX network as the front-end detector of the DeepSORT tracking algorithm. The YOLOX network can be abstracted into four parts: input, backbone, neck, and head. As shown in Figure 1, the input uses a data augmentation algorithm combining Mosaic and Mixup algorithms to augment the samples. The backbone performs feature extraction, the head performs classification and regression analysis, and the neck performs a multiscale fusion of the feature maps obtained by the backbone network. As shown in Figure 2, the backbone network combines Focus with the CSPDarknet53 network to obtain the input images' multiscale feature maps (i.e., Dark3_out, Dark4_out, Dark5_out). Then, the feature maps are input into the neck network for feature extraction and feature fusion, which integrates the feature pyramid network (FPN) and path aggregation network (PAN) to construct a feature pyramid and output three different feature layers, so the network can effectively predict targets of different scales (i.e., 80×80, 40×40, and 20×20 pixels), as shown in Figure 3. The decoupled detection head performs regression and classification prediction based on the features extracted by the neck network, as shown in Figure 4. Finally, the head implements anchor-free detection based on the SimOTA strategy. In particular, pretraining and multistage freezing strategies are applied in the model training phase.

When the UAV is flying high, the vehicles in the image are incredibly small-scale targets. In deeper networks, the features of small-scale vehicles are severely lost. The original YOLOX topology still has some bottlenecks in detecting tiny objects, especially in high-altitude aerial images. Therefore, we perform some improvements to the YOLOX network topology. It mainly includes four improvements: adding a shallow feature extraction network, introducing an attention module, improving the loss function, and creating an image segmentation method.

2.1. Improvement of Multiscale Feature Extraction Network. There are only three detection layers in the YOLOX network structure. When the input image is 640×640 pixels, detection layers of 80×80, 40×40, and 20×20 pixels can be obtained after extracting features from the feature pyramid. Because small object detection depends on high-level detail information and low-level semantic information, feature fusion is challenging to improve its detection accuracy significantly. In addition, the YOLOX network backbone network will lose more detailed information after several times of down-sampling, and small objects in the input image are easily undetected. Extracting the image's global and local features is beneficial to target reasoning by increasing the multiscale receptive field. Hence, a shallow feature extraction network of 160×160 pixels is added in this

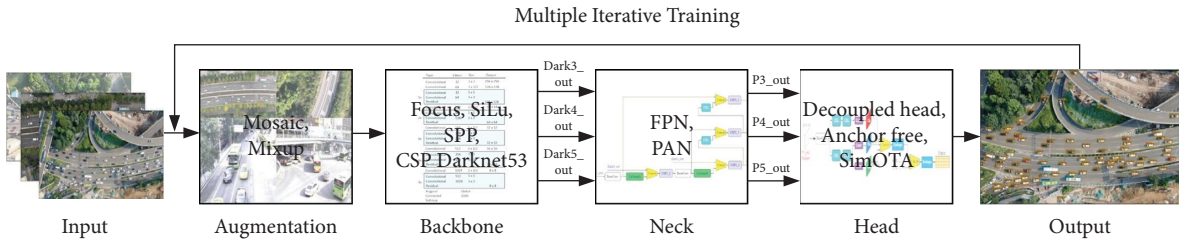


Figure 1: Simplified structure of the YOLOX network.

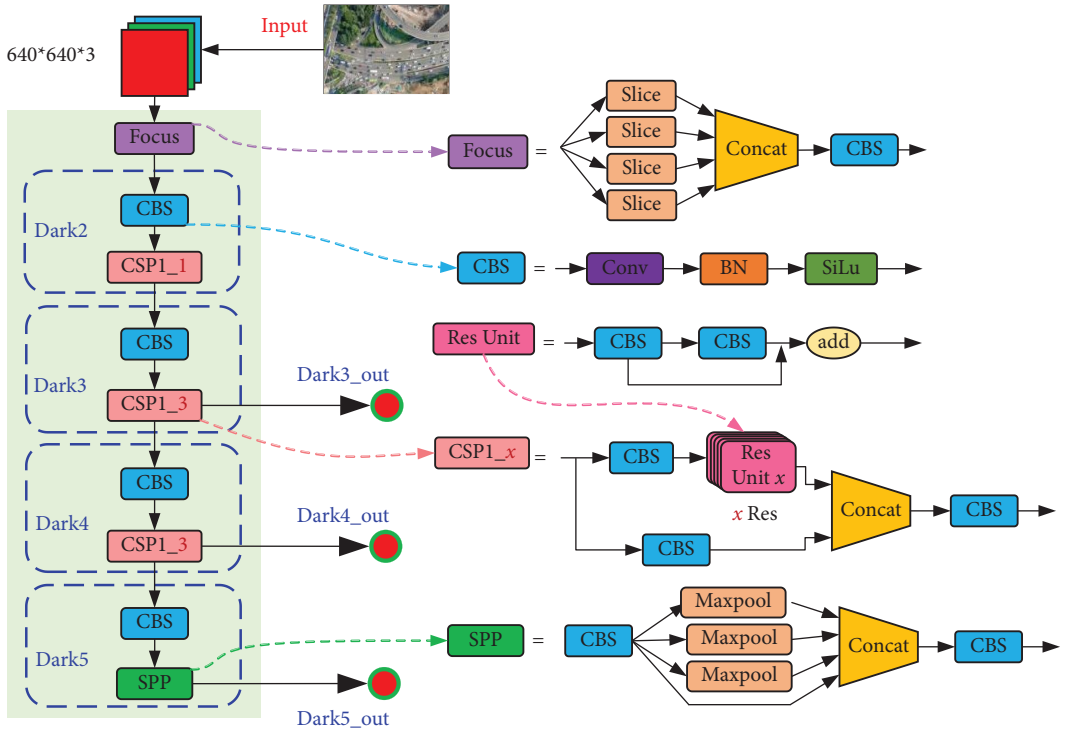


Figure 2: Structure of the backbone network.

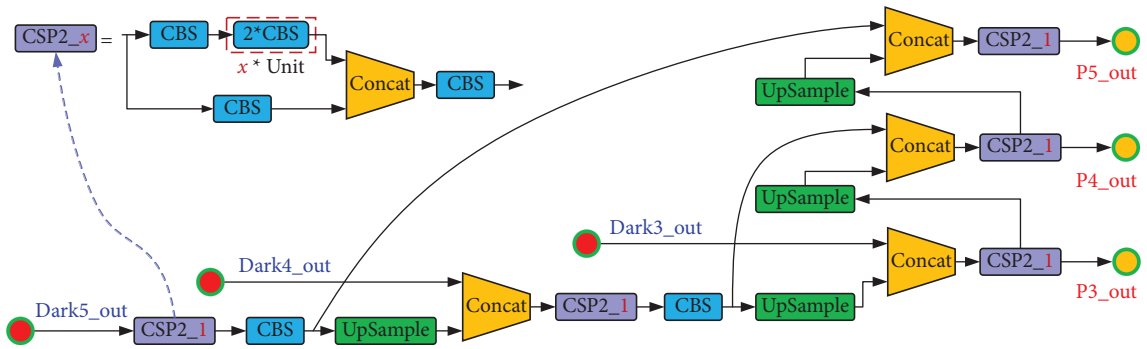


Figure 3: Neck structure of feature fusion network.

paper. Specifically, the Dark2_out detection layer is added to improve the detection performance for small objects. The orange part in Figure 5 is the added shallow feature extraction network. The features are enhanced in the second layer of the backbone network, and the up-sampling operation is added after the 18th layer to expand the feature map. In the detection head, the feature map of 160×160 pixels obtained by up-sampling of the 19 layers is fused with the feature map of the second layer in the backbone to obtain four scale detection layers, namely, 20×20, 40×40, 80×80, 160×160 pixels, and the added 160×160 pixel detection layer can detect small targets down to 4×4 pixels.

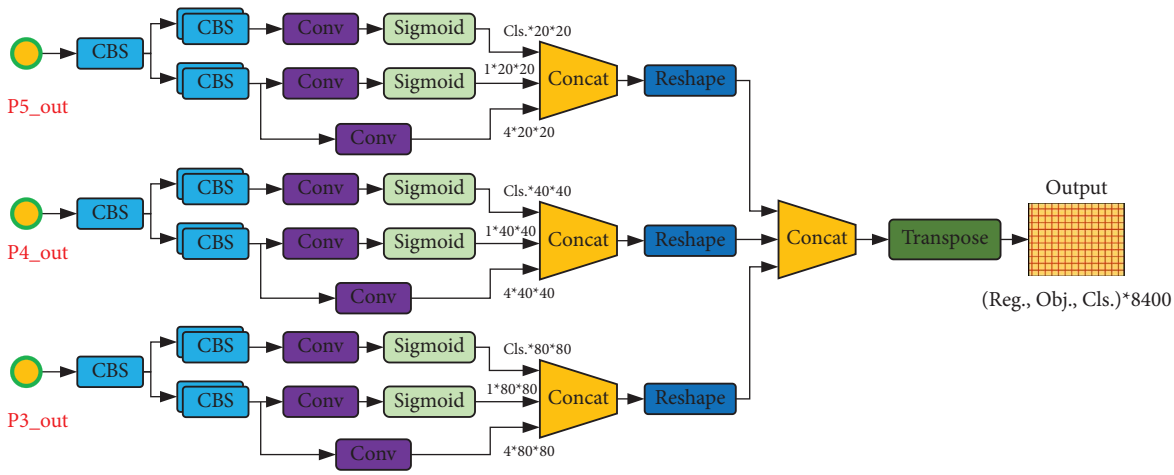


Figure 4: Decouple head network.

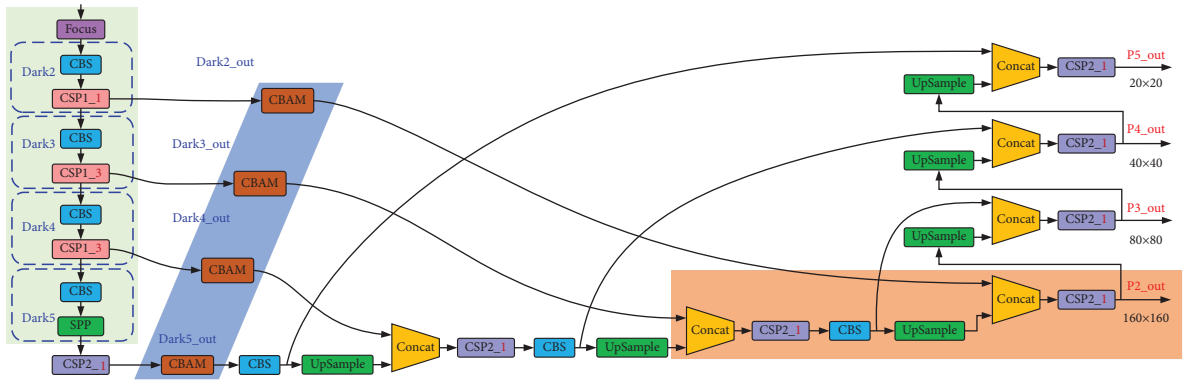


Figure 5: The structure of the improved YOLOX network.

2.2. Attention Mechanism Module. Small target vehicles account for a tiny proportion of UAV images and are susceptible to interference from complex backgrounds. YOLOX, on the other hand, has no attention preference, resulting in poor feature expression ability for small targets, especially in detection and tracking tasks with long distances, extensive backgrounds, and small targets. The attention mechanism suppresses the expression of these inevitable redundant features by increasing the weight of nonredundant features. Therefore, this paper inserts the convolutional block attention module (CBAM) mechanism after the backbone network, as shown in the blue part in Figure 5, to enhance the ability to extract vehicle-like features while suppressing redundant information such as background. Figure 6 illustrates the structure of the CBAM, which is composed of a channel attention (CA) module and a spatial attention (SA) module in series.

Introducing the CA mechanism module makes it possible to effectively detect target contour features and obtain more content for target detection. The calculation method is introduced as follows:

$$M_C(F) = \sigma\left(\mathbf{W_1}\left(\mathbf{W_0}\left(F^c_{\text{avg}}\right)\right) + \mathbf{W_1}\left(\mathbf{W_0}\left(F^c_{\text{max}}\right)\right)\right), \quad (1)$$

where $M_C(F)$ is the output weight of the CA, σ is the sigmoid activation function, F^C_{avg} is the spatial feature map after average pooling, F^C_{max} is the spatial feature map after max pooling, $\mathbf{W_0}$ is the weight matrix of the first fully connected layer, and $\mathbf{W_1}$ is the weight matrix of the second fully connected layer.

By introducing the SA mechanism module, the position of the detected target can be effectively located and the detection rate can be improved. The calculation method is introduced as follows:

$$M_S(F) = \sigma\left(f^{7\times7}\left(F^s_{\text{avg}}; F^s_{\text{max}}\right)\right), \quad (2)$$

where $M_S(F)$ is the output weight of the SA, $f^{7\times7}$ is a convolution operation filter of size 7×7, F^S_{avg} is the feature map after average pooling on the channel, and F^S_{max} is the feature map after max pooling on the channel.

In short, the input feature map F is first multiplied element-by-element by the CA module. Then, the feature result is multiplied by the SA mechanism module, and the

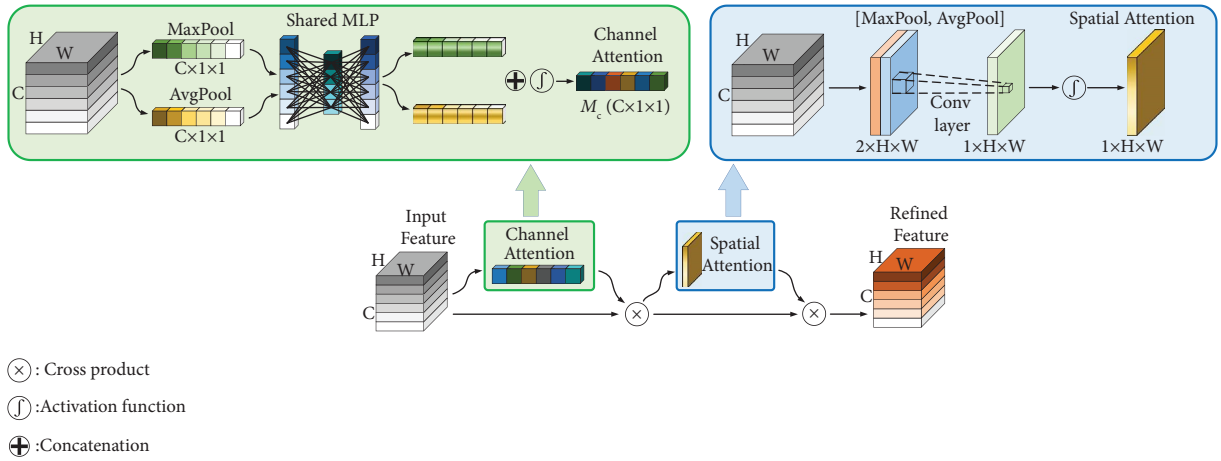


Figure 6: Structure of the CBAM module.

final feature map F'' is obtained after CBAM processing. The calculation method is

$$F' = \mathbf{M}_c(F) \times F, \tag{3}$$

$$F'' = \mathbf{M}_s(F) \times F', \tag{4}$$

where F is the input feature map, F' is the feature map weighted by the CA, and F'' is the feature map weighted by the SA.

2.3. Improvement of the Loss Function. The vehicle scale in UAV images is tiny, and the error of one or two pixels between the bounding box and the ground truth will cause significant positioning loss, easily causing intersection over union (IoU) less than 0.5. Therefore, to improve the positioning accuracy of small targets in dense areas, this paper uses EIoU_Loss (Efficient-IoU) to replace the original IoU loss during training. The EIoU_Loss function can reduce the IoU between the bounding box and the ground truth, reduce the distance between the centers of the two boxes, and reduce the width/height gap, which is more accurate than the original IoU loss function. The EIoU_Loss is divided into IoU loss (L_{IoU}), distance loss (L_{dis}), and width/height loss (L_{asp}), as shown in equation (3) The EIoU_Loss function can independently penalize the w and h of the bounding box instead of the aspect ratio. The schematic diagram of EIoU_Loss is shown in Figure 7.

$$\begin{aligned}\mathscr{L}_{\text{EIoU}} &= \mathscr{L}_{\text{IoU}} + \mathscr{L}_{\text{dis}} + \mathscr{L}_{\text{asp}} \\ &= 1 - \text{IOU} + \frac{\rho^2\left(\mathbf{b}, \mathbf{b}^{\text{gt}}\right)}{c^2} + \frac{\rho^2\left(w, w^{\text{gt}}\right)}{C_w^2} + \frac{\rho^2\left(h, h^{\text{gt}}\right)}{C_h^2},\end{aligned} \tag{5}$$

where $d = \rho^2(\mathbf{b}, \mathbf{b}^{\text{gt}})$ represents the Euclidean distance between the center points of the bounding box and the ground truth. c is the diagonal distance of the smallest circumscribed rectangle covering the two boxes. C_w and C_h represent the width and height of the smallest circumscribed rectangle between the bounding box and ground truth.

2.4. Image Segmentation Method. This paper proposes an image segmentation method (ISM) to improve the detection performance of the network for UAV images with long distances, extensive backgrounds, and small targets. The main idea of the ISM is to segment large-sized images into multiple subimages through the cropping operation. In the subimages, the proportion of targets to be detected is more significant, making it easier to capture adequate feature information. It should be noted that when segmenting the original large-sized image, the vehicle on some boundaries will be truncated into two or more parts, resulting in incomplete object feature information and affecting subsequent detection tasks. This paper adopts the sliding window method (SWM) in response to this issue. The input image is divided horizontally and vertically with the fixed step to obtain several subimages. As shown in Figure 8, the SWM-based image segmentation method ensures that there are overlapping regions in adjacent subimages, and the truncated target will entirely exist in adjacent subimages so that the target feature information can be wholly captured, effectively reducing false detection and missed detection.

Assuming that the size of the original image is $w \times h$ pixels, it is cropped into n subimages of $d \times d$ pixels with the horizontal stride s_w and vertical stride s_h. The calculation formula for the number of the subimages is

$$n = \left\lfloor \frac{(w-d)}{s_w} + 1 \right\rfloor \times \left\lfloor \frac{(h-d)}{s_w} + 1 \right\rfloor. \tag{6}$$

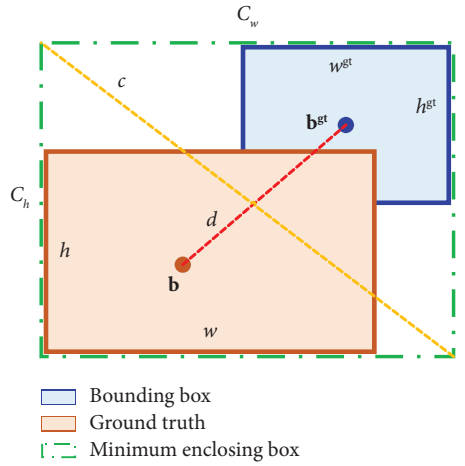


Figure 7: Schematic diagram of the EIoU.

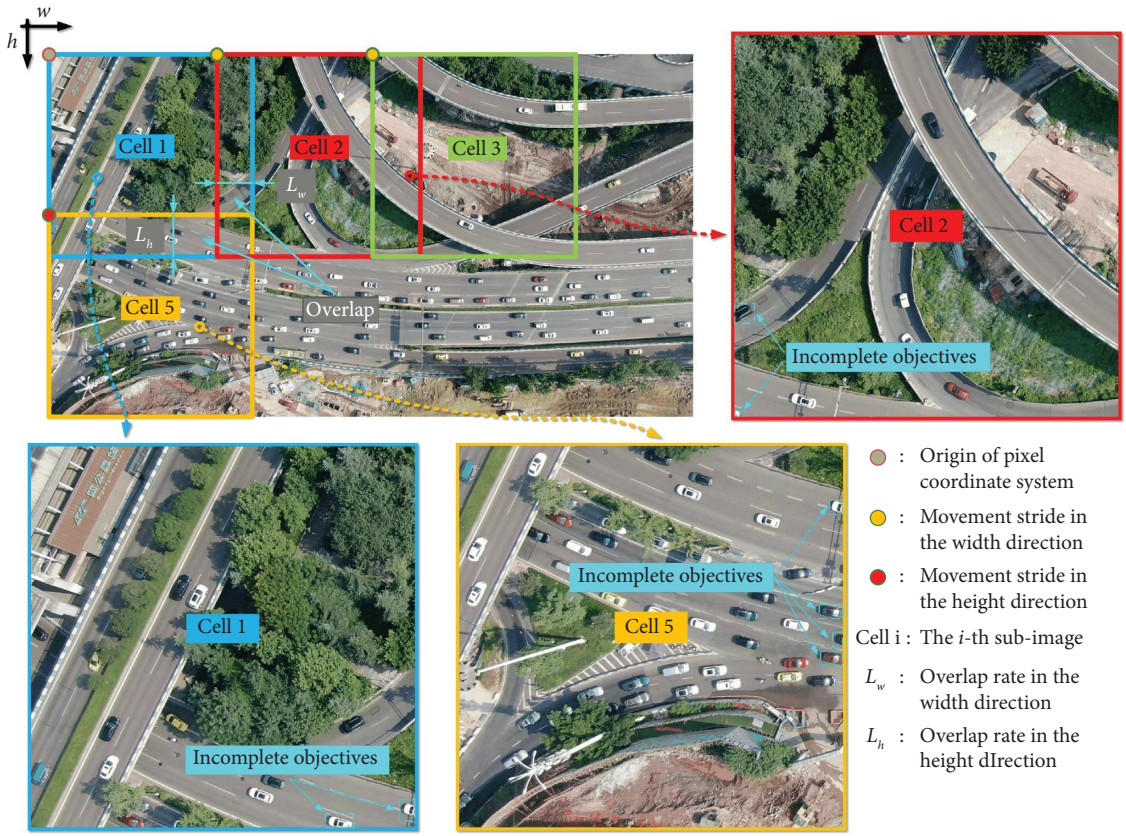


Figure 8: Schematic diagram of image segmentation in a sliding window manner.

The strides moving along the horizontal direction and vertical direction when cutting high-definition large-size images are expressed as

$$h_s = \left\lfloor \frac{(h-d)}{(h_n-1)} \right\rfloor, \quad w_s = \left\lfloor \frac{(w-d)}{(w_n-1)} \right\rfloor. \tag{7}$$

Among them, the parameters w and h are the width and height of the original high-definition image, respectively. d is the side length of the subimage, and w_n and h_n are the number of cuts along the horizontal and vertical directions, respectively.

The segmented subimages are sequentially input into the YOLOX network to obtain detection results of the vehicles, such as category, confidence score, and prediction box. The results of all subimages are then mapped to the original

image coordinate system, attention! instead of directly combining all resulting subimages into the original large-sized image. In addition, the same target in the overlapping area may be repeatedly detected, resulting in multiple bounding boxes for the same target. In order to solve this problem, this paper introduces the nonmaximum suppression (NMS) algorithm to filter duplicate boxes and retain the frame with the highest confidence. The calculation formulas of the NMS algorithm are shown in equations (8) and (9). The overall process of the image segmentation method is demonstrated in Figure 9.

$$\mathrm{IoU}\left(M, b_i\right) = \frac{M \cap b_i}{M \cup b_i}, \tag{8}$$

$$s_i = \begin{cases} s_i, \mathrm{IoU}\left(M, b_i\right) < N_t, \\ 0, \mathrm{IoU}\left(M, b_i\right) \geq N_t, \end{cases} \tag{9}$$

where s_i represents the score of each candidate box, M is the current box with the highest score, b_i is one of the remaining boxes, and N_i is the set threshold, usually taken as 0.5 or 0.7.

2.5. Multitarget Tracking Algorithm. The improved YOLOX detection model proposed in this paper is combined with the DeepSORT tracking algorithm to realize the tracking of small-scale vehicles in UAV aerial images, as shown in Figure 10. First, the UAV aerial video is input into the improved YOLOX network to obtain the detection results. Then, use the DeepSORT algorithm to match the bounding box and the tracking box frame by frame, obtain the vehicle's identity, coordinates, and other information, and finally output the tracking image.

Figure 11 shows the flowchart of the DeepSORT algorithm. The core of the DeepSORT tracking algorithm is to use the Kalman filter and the Hungarian matching algorithm and use the IOU between the tracking and detection result as the cost matrix to track the moving target. In order to track the small target vehicles detected by the YOLOX network, the DeepSORT uses an 8-dimensional variable matrix x to describe the appearance information of the vehicle and the motion information in the image, as shown in the following equation:

$$x = (u, v, \gamma, q, \dot{u}, \dot{v}, \dot{\gamma}, \dot{q}). \tag{10}$$

In the formula, (u, v) represents the center coordinates of the vehicle, γ represents the aspect ratio of the vehicle detection box, q represents the height of the vehicle detection box, and (u, v, γ, q) represents the corresponding speed information of $(\dot{u}, \dot{v}, \dot{\gamma}, \dot{q})$.

The DeepSORT algorithm combines the vehicle's motion and appearance information and uses the Hungarian algorithm to match the prediction and tracking boxes. For the motion information of the vehicle, the Mahalanobis distance is used to describe the degree of correlation between the prediction results of the Kalman filter and the improved YOLOX detection results, as shown in the following equation:

$$d^{(1)}(i, j) = \left(d_j - y_i\right)^T S_i^{-1} \left(d_j - y_i\right), \tag{11}$$

where d_j represents the j-th detection box, y_i represents the state vector of the i-th detection box, and S_i represents the covariance matrix between detection and tracking results.

Appearance models come into play when vehicles are occluded for short or long periods. Currently, the feature extraction network will calculate a 128-dimensional feature vector for each detection box, with a constraint of $\|r_j\| = 1$. At the same time, a 100-frame appearance feature vector with a determined trajectory will be constructed for each detected vehicle. Calculate the minimum cosine distance between these two using the following equation:

$$d^{(2)}(i, j) = \min\left\{1 - r_j^T r_k^{(i)} \mid r_k^{(i)} \in R_k\right\}, \tag{12}$$

where r_j represents the feature vector corresponding to the detection box and r_k represents the feature vector successfully associated with 100 frames.

Mahalanobis distance provides reliable target position information in short-term prediction. After the occluded object reappears, the minimum cosine distance of appearance features can recover the object's ID. In order to complement the advantages of the two measures, the two distances are linearly weighted as the final measure, as shown in the following equation:

$$c_{i,j} = \lambda d^{(1)}(i, j) + (1 - \lambda) d^{(2)}(i, j). \tag{13}$$

In the formula, λ represents the weight coefficient. If $c_{i,j}$ falls within the specified threshold range, it is considered that the correct correlation has been achieved.

3. Experiment and Discussion

3.1. Experimental Environment Platform. The hardware experiment platform is Intel(R) Core(TM) i9-10900K @ 3.00 GHz CPU, 64 G RAM, and an NVIDIA GeForce RTX 3080 Ti (24 G) graphics card. The software experiment platform is CUDA 11.4, cuDNN 8.2.2, the deep learning framework is Darknet and PyTorch 1.9.1, the integrated development environment is PyCharm, the programming language is Python, and the visualization is based on OpenCV. Like YOLO v5, YOLOX networks also use depth and width factors for network splitting, including YOLOX-s, YOLOX-m, YOLOX-l, and YOLOX-x. In this experiment, we use the lightest model, YOLOX-s, based on which various improvements are made and compared with the original network.

3.2. Dataset. Figure 12 shows the UAV used to collect video data in this paper. It supports 8 km long-distance control, 4 K HDR imaging, and $f/2.8 - f/11$ adjustable aperture and can capture ultraclear images. Use the Labeling software to annotate the image and convert it into VOC dataset format. The label categories are divided into "car," "bus," and "truck'." The ratio of the training set to the validation set is 8 : 2. The dataset manufactured in this paper has the

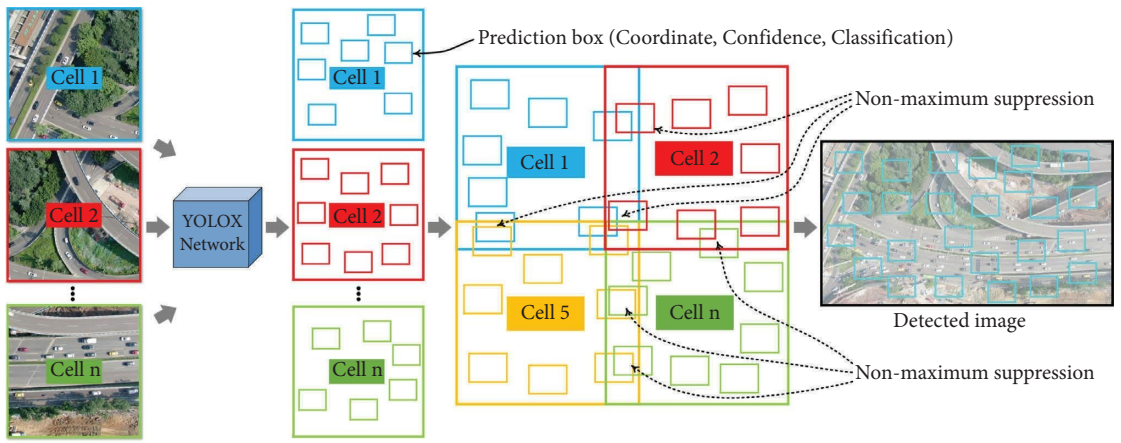


Figure 9: The overall process of image segmentation methods for high-resolution images.

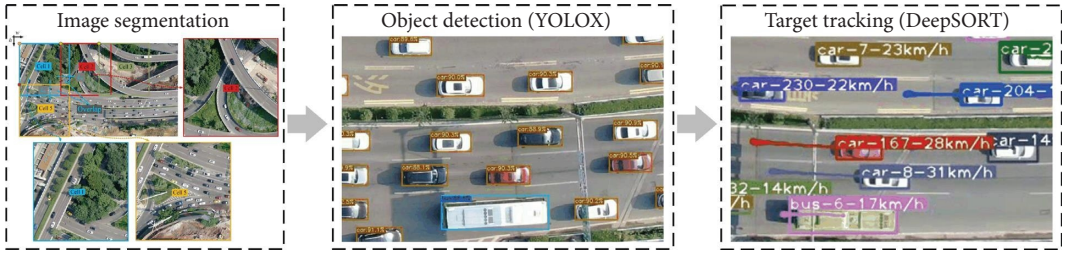


Figure 10: The process of small-scale vehicle detection and tracking tasks.

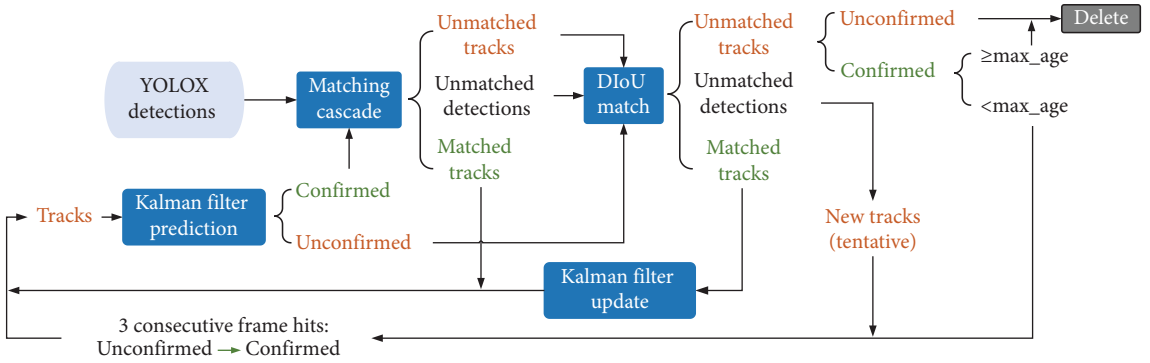


Figure 11: The flowchart of the DeepSORT algorithm.

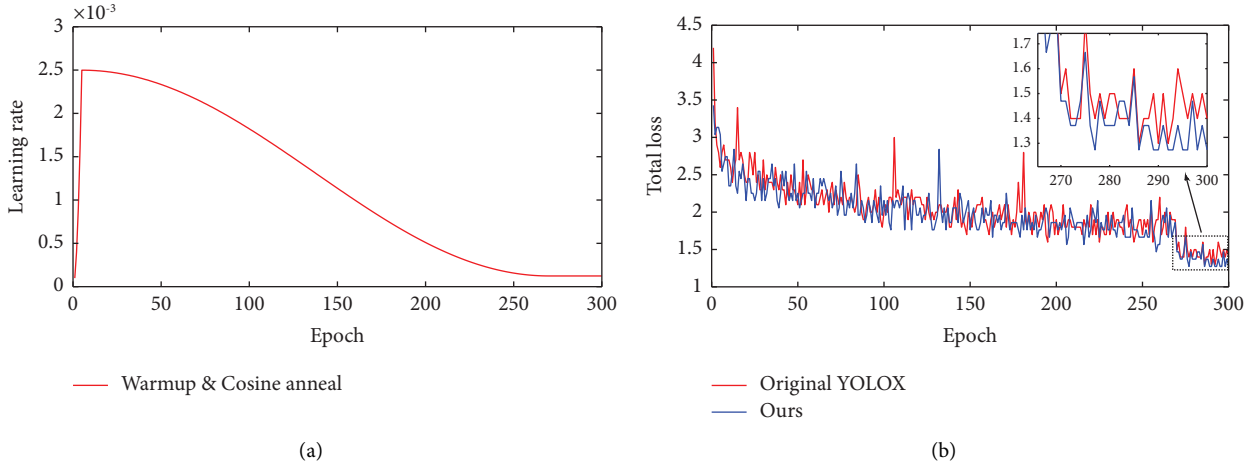
characteristics of long distance, clustered, dense, and tiny, which is suitable for verifying the performance of the proposed algorithm.

3.3. Model Training. A mixed Mosaic and Mixup augmentation strategy expands the vehicle sample. The stochastic gradient descent (SGD) algorithm is used to update and optimize the weight of the network model. The learning rate in the initial stage of training adopts the warm-up strategy to avoid overfitting due to a large initial learning rate. During the warm-up phase, the learning rate increases from 0 to a set value of 0.0025. Once completed, update the learning rate using the cosine annealing algorithm, as shown in equation (14). Figure 13(a) shows the specific change trend of the learning rate. Use the officially recommended pretraining weights for 300 training epochs, set the batch size to 16, the weight decay coefficient to 0.0005, the momentum of SGD to 0.9, and the warm-up epochs to 5. Table 1 lists the main parameters adopted in the training experiment. Too many epochs of data augmentation lead to deviation from the original samples, degrading training performance. Therefore, the Mosaic and Mixup augmentation strategies are turned off in the last 30 epochs.

$$\eta_t = \eta_{\min}^i + \frac{1}{2}\left(\eta_{\max}^i - \eta_{\min}^i\right)\left[1 + \cos\left(\frac{T_{\text{cur}}}{T_{\max}}\pi\right)\right]. \tag{14}$$

Among them, $\eta_{\max}^i$ and $\eta_{\min}^i$ are the maximum and minimum values of the learning rate, respectively, and T_{cur} and $T_{\max}$ are the current and maximum epoch, respectively.

Figure 12: UAV and its operation interface.



(a) (b)

Figure 13: Training process diagram: (a) learning rate update combining warm-up and cosine anneal; (b) total loss changes of the model during training.

Table 1: Main parameters adopted in the training experiment.

Parameters	Value and unit
Input size	768 × 768 pixel
Training epochs	300
Warm-up epochs	5
Number of classes	3
Batch size	16
Optimizer	SGD
Scheduler	Warm-up + cosine annealing
Initial learning rate	0
Minimum learning rate ratio	0.05
Momentum	0.9
Weight decay	$5e-4$
Detection threshold	0.01
IoU threshold for softer NMS	0.65
No augmentation epochs	30

3.4. Evaluation Metrics. In order to comprehensively evaluate the performance of the improved YOLOX-s model proposed in this paper, metrics such as precision (P), recall (R), $F1$-score, average precision (AP), and mean average precision (mAP) are introduced to evaluate the trained model. Index P describes the proportion of correctly predicted true-positive samples among all predicted positive samples, which is defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%. \tag{15}$$

Index R refers to the ratio of the correctly predicted number of positive samples to the total number of positive samples, that is, how many positive samples have been found among all positive samples. Its definition is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%. \tag{16}$$

The $F1$-score is based on the harmonic mean of precision and recall.

$$F1 - \text{score} = \frac{2 \times P \times R}{P + R}. \tag{17}$$

Among them, true positive (TP), false positive (FP), and false negative (FN) refer to correctly classified positive samples, incorrectly classified positive samples, and incorrectly classified negative samples, respectively.

The average precision (AP) is the average of the precision at different recall points: the envelope area of the P-R curve and the coordinate axes.

$$\begin{aligned} \text{AP}_i &= \int_0^1 P_i(R_i) dR_i, \\ \text{AP}_{@0.5:0.95} &= \frac{1}{10}\left(\text{AP}_{@0.5} + \text{AP}_{@0.55} + \cdots + \text{AP}_{@0.95}\right). \end{aligned} \tag{18}$$

The mean average precision (mAP) is the average of the AP of all categories, which is used to measure the quality of the training model in all categories.

$$m\text{AP} = \frac{1}{n}\text{AP}_i. \tag{19}$$

The mAP has two evaluation metrics, $\text{AP}_{@0.5}$ and $\text{AP}_{@0.5:0.95}$. $\text{AP}_{@0.5}$ is the average value of all categories of AP calculated when IoU = 0.5. $\text{AP}_{@0.5:0.95}$ means that the IoU ranges from 0.5 to 0.95 with a step size of 0.05 and calculates the average of all APs under these ten different IoUs, which can more fully reflect the performance of the detection model.

3.5. Results and Discussion. After training for 300 epochs with the parameters given in Table 1, the loss value changes of the original and improved YOLOX networks are shown in Figure 13(b). The abscissa Epoch represents the training rounds, and the ordinate represents the total loss value of the predicted results. It can be seen that the improved network's convergence rate is better than that of the original network. After the 270th round of unfreezing (turn off the Mosaic and Mix Up data augmentation), the convergence speed of the loss value is significantly accelerated. Finally, the loss value of the former is smaller.

Ablation experiments verify the performance changes caused by network structure changes. As mentioned above, we proposed three components: the shallow feature extraction network, CBAM block, and EIoU_Loss function. The experimental results are shown in Figure 14 and Table 2, respectively. Figure 14 shows the training process of different models on the dataset. It can be seen that the performance of different network structures has changed to varying degrees. The proposed improved YOLOX evaluation indicators are significantly higher than those of the original YOLOX network. The specific training results are shown in Table 2, where the symbol "$\surd$" indicates adding the corresponding module. It can be seen that the proposed improved YOLOX network has the highest mAP value. Compared with the original network, the improved YOLOX's $\text{mAP}_{@0.5}$ and $\text{mAP}_{@0.5:0.95}$ have increased by 3.30% and 3.35%, respectively. With the introduction of each module, the network's performance has been improved to varying degrees, proving the effectiveness of the improvement scheme proposed in this paper.

It must be declared that the image segmentation method is only activated in the detection phase, not the training phase. The main parameters of the image segmentation method are shown in Table 3. The UAV aerial images are high-definition images, i.e., 3840×2160 pixels. Of course, the image's pixel can be higher if necessary. We divide the original high-definition image into two rows and four columns, a total of eight subimages with a resolution of 1280×1280 pixels. The moving steps of the sliding window are 853 pixels horizontally and 880 pixels vertically. The overlapping ratios in the horizontal and vertical directions are 31.25% and 33.36%, respectively.

To intuitively demonstrate the detection performance of the improved YOLOX algorithm, UAV images in multiple traffic scenarios, including the interchange weaving area and the urban arterial road weaving area, were selected for comparative experiments. The detection results are shown in Figures 15–17. Figure 15 shows the detection results of the original YOLOX network, and Figure 16 shows the detection results of the improved YOLOX network. Figure 17 shows the detection results obtained using conventional image segmentation techniques without the NMS operations.

The following conclusions can be drawn from the detection performance comparison chart: (1) for the high-resolution UAV aerial images, there will be a large number of missed detections using the original YOLOX network. On the contrary, the improved YOLOX can effectively avoid missed detections, which intuitively proves the positive effect of the improved network. The detection performance of the improved network is substantially improved. (2) Compared with the original YOLOX network, the confidence score of the improved YOLOX network is higher, especially for areas with dense vehicle distribution, proving the improved network's accuracy and robustness. (3) The partially enlarged pictures show that the bounding box of the original YOLOX is larger than the vehicle targets (i.e., the ground truth), especially the upper part of the bounding box (i.e., the area showing the confidence score) is not close to the vehicle contour, which will reduce the positioning accuracy of trajectory tracking. The bounding box of the improved YOLOX network is closer to the actual contour outline of the vehicle, and the vehicle positioning is more accurate. (4) The improved YOLOX network without NMS operation effectively solves the missing detection phenomenon, and the confidence score is also significantly improved. However, some vehicles in the overlapping area have multiple bounding boxes. The image segmentation method proposed in this paper to perform the NMS operations on multiple bounding boxes of the same target in the overlapping area can solve this problem well.

The original YOLOX and the improved YOLOX network are used as the front-end detectors of the DeepSORT tracking algorithm to compare and analyze the performance of the multitarget tracking algorithm. Figure 18 compares vehicle tracking performance in different scenarios, with the original algorithm on the left and the improved algorithm on the right. A large number of vehicles are not tracked correctly by the original algorithm. The reason is that the original YOLOX detector has seriously missed detection, which makes it difficult for the tracking algorithm to perform its role. On the contrary, the improved algorithm can track all small-scale vehicles in high-definition video, and the tracking boxes fit the vehicles' contours more closely. Facts have proved that the improved YOLOX proposed in this paper, as the detector of the DeepSORT algorithm, can solve the problem of missing tracking of small-scale targets in UAV high-altitude aerial video. The vehicle trajectory extracted by the improved algorithm is shown in Figure 19, which reproduces the trajectory characteristics of small-scale vehicles in the UAV aerial videos. Therefore, an accurate tracking framework can locate the vehicle position well and provide a data basis for extracting vehicle trajectory data and conducting research in related fields such as transportation and traffic safety.

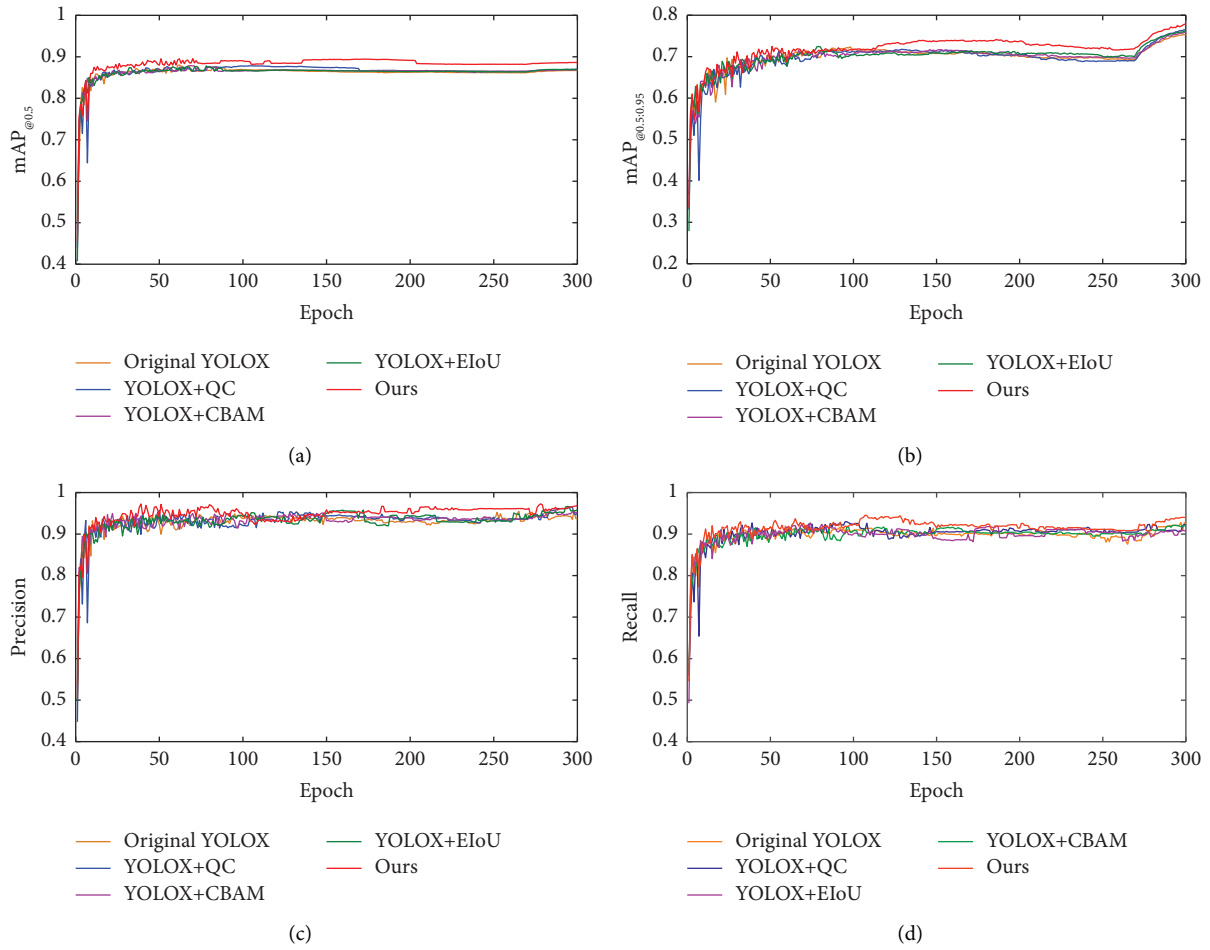


Figure 14: Training process diagram: (a) the change curve of the $mAP_{@0.5}$; (b) the change curve of the $mAP_{@0.5:0.95}$; (c) the change curve of the precision; (d) the change curve of the recall.

Table 2: Comparison results of ablation experiments.

Methods	Block			Metric (%)				
	QC	CBAM	EIoU	$mAP_{@0.5}$	$mAP_{@0.5:0.95}$	Recall	Precision	$F1$-score
YOLOX				86.71	75.32	92.81	95.65	94.21
	√			87.83	76.51	92.99	96.58	94.75
		√		87.87	75.93	92.18	95.52	93.82
			√	87.79	76.14	92.70	96.67	94.64
	√	√		88.31	77.12	92.36	96.45	94.36
	√		√	88.11	77.33	92.07	97.60	94.75
		√	√	88.21	76.75	92.88	96.54	94.67
Ours	√	√	√	89.57	77.84	94.55	97.28	95.39

Table 3: Parameters of image segmentation.

Parameters	Value and unit
Input size	3840 × 2160 pixel
Subimages	1280 × 1280 pixel
Number of rows	2
Number of columns	4
Number of subimages	8
Overlap rate of rows	31.25%
Overlap rate of columns	33.36%
Horizontal movement	853 pixels
Vertical movement	880 pixels

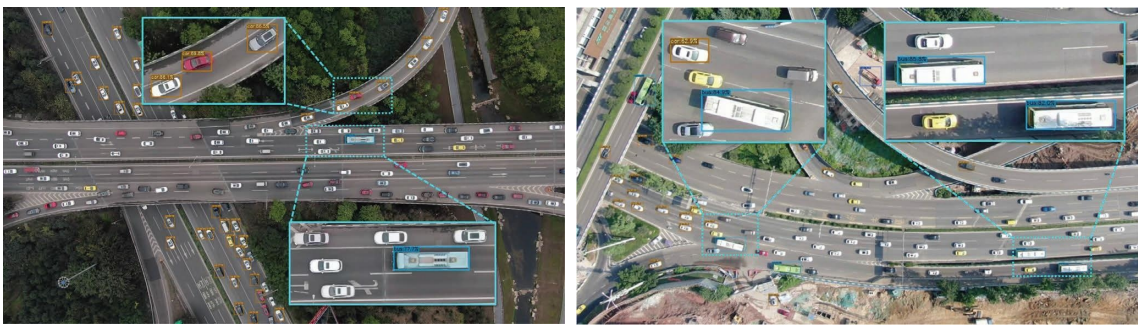


Figure 15: Detection results of the original YOLOX.

Figure 16: Detection results of the improved YOLOX.

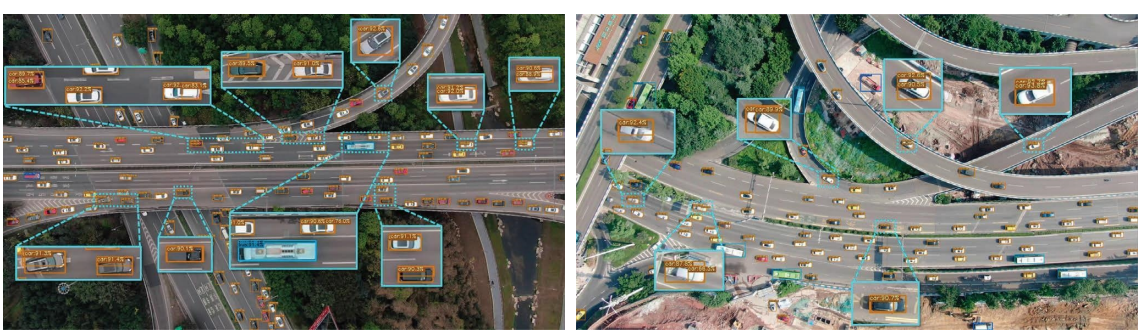
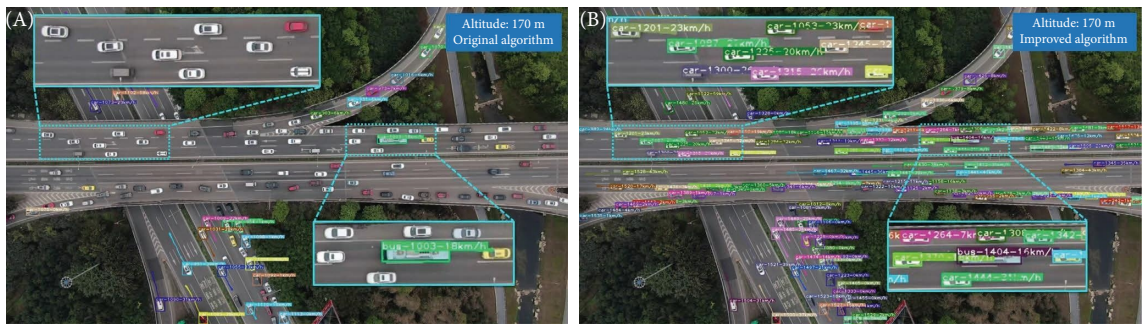


Figure 17: Detection results of the improved YOLOX without the NMS operation.

(a)

Figure 18: Continued.



(b)

Figure 18: Tracking performance comparison between the original model (A) and the improved model (B): (a) weaving area of urban arterial roads; (b) weaving area of urban arterial roads.

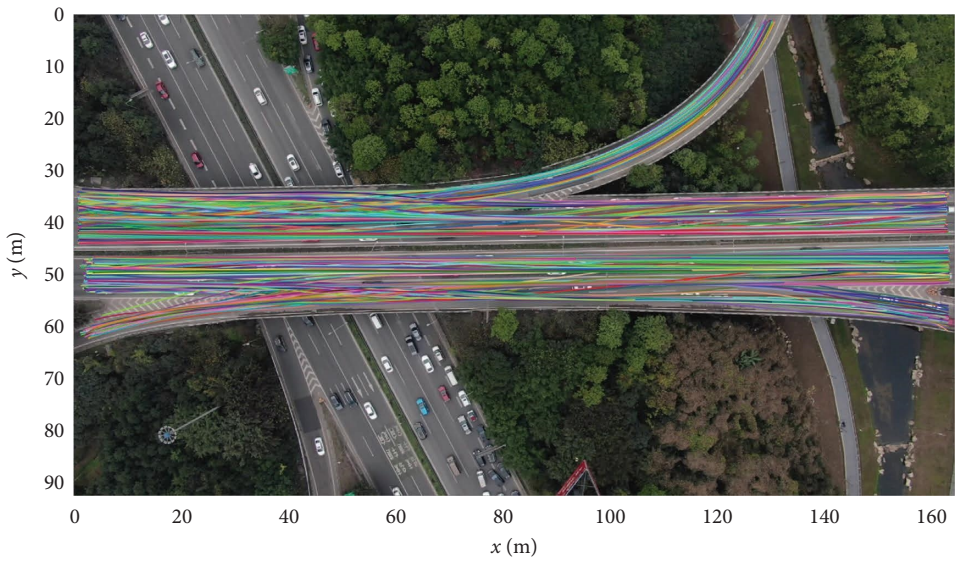


Figure 19: Spatial distribution map of vehicle trajectories.

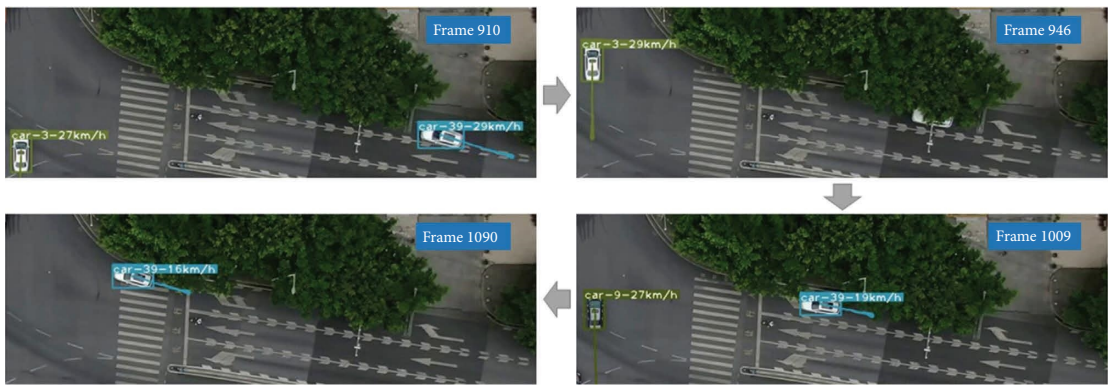


Figure 20: The tracking effect of the vehicle in the case of short-term occlusion.

Figure 20 shows the improved algorithm to deal with the tracking problem when the vehicle is occluded for a long time. When the car turned right and was blocked by trees on the side of the road (Frame 946), there was a short-term tracking failure, and the ID number disappeared for a short time. Once the car leaves the occlusion area, the tracker immediately reidentifies it as the original, the ID number remains the same, and the ID number is not lost or switched.

Figure 21: Tracking results of aerial images taken at different altitudes.

Figure 21 shows the tracking results of videos taken at different aerial altitudes, ranging from 140 meters, 170 meters, 200 meters, 250 meters, and 300 meters. The left side shows the tracking results of the original algorithm, and the right side shows the tracking results of the improved algorithm. As the altitude of UAV aerial photography

increases, the number of missed tracking for small-scale vehicles in the original algorithm gradually increases. In contrast, the improved algorithm can still accurately track small-scale vehicles in high-altitude aerial videos, and there is no missing tracking as the aerial height increases. Experiments have shown that the improved algorithm proposed in this paper has high robustness and can adapt to aerial videos of different heights.

4. Conclusion

Aiming at the problems of missed detection, false detection, and low detection accuracy of small-scale vehicles in high-altitude UAV aerial images, this paper proposes a detection-based small-scale target tracking algorithm. First, some improvements have been made to the original YOLOX network, including adding a shallow feature extraction network, embedding the CBAM module, introducing EIoU_Loss as the regression loss function of the bounding box, and proposing an image segmentation method. Experiments show that the improved YOLOX network significantly reduces the number of missed detections, the bounding box fits the outline of the target more closely, and the confidence score is higher.

Then, the improved YOLOX is used as the detector of the DeepSORT tracking algorithm. Experiments show that the improved algorithm can track all small-scale vehicles in high-resolution UAV images. In addition, the bounding box of the improved algorithm is also smaller and closer to the vehicle's outline. It solves massive missing tracking events in the tracking task and proves the effectiveness of the improved TBD algorithm proposed in this paper.

High vehicle detection accuracy helps improve the stability of trajectory tracking and provides more solid data support for vehicle trajectory tracking. Vehicle trajectory tracking technology is widely used in intelligent transportation, such as traffic conflict discrimination, traffic safety early warning, traffic intelligence control, and others.

Data Availability

The data used to support the findings of this study can be obtained upon request from the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

Jin Xu is grateful for the support provided by the National Natural Science Foundation of China (52172340); Heshan Zhang is grateful for the support provided by the Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN202200710), the Open Fund Project of Chongqing Key Laboratory of Traffic and Transportation (2018TE01), and the Joint Training Base Construction Project for Graduate Students in Chongqing (JDLHPYJD2021006); Zhanji Zheng is grateful for the support provided by the Chongqing Natural Science Foundation Project of China (CSTB2022NSCQ-BHX0731); Shuang Luo is grateful for the support provided by the Chongqing Natural Science Foundation Project of China (cstc2021jcyj-msxmX0794).

References

[1] J. Lin, J. Peng, and J. Chai, "Real-time UAV correlation filter based on response-weighted background residual and spatio-temporal regularization," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[2] J. Li, D. H. Ye, M. Kolsch, J. P. Wachs, and C. A. Bouman, "Fast and robust UAV to UAV detection and tracking from video," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 3, pp. 1519–1531, 2022.

[3] R. Ke, Z. Li, J. Tang, Z. Pan, and Y. Wang, "Real-time traffic flow parameter estimation from UAV video based on ensemble classifier and optical flow," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 54–64, 2019.

[4] S. Wang, Z. Qu, C. Li, and L. Gao, "BANet: small and multi-object detection with a bidirectional attention network for traffic scenes," *Engineering Applications of Artificial Intelligence*, vol. 117, Article ID 105504, 2023.

[5] Y. Dai, Z. Hu, S. Zhang, and L. Liu, "A survey of detection-based video multi-object tracking," *Displays*, vol. 75, Article ID 102317, 2022.

[6] J. Hu, H. Niu, J. Carrasco, B. Lennox, and F. Arvin, "Fault-tolerant cooperative navigation of networked UAV swarms for forest fire monitoring," *Aerospace Science and Technology*, vol. 123, Article ID 107494, 2022.

[7] R. Eskandari, M. Mahdianpari, F. Mohammadimanesh, B. Salehi, B. Brisco, and S. Homayouni, "Meta-analysis of unmanned aerial vehicle (UAV) imagery for agro-environmental monitoring using machine learning and statistical models," *Remote Sensing*, vol. 12, no. 21, p. 3511, 2020.

[8] Y. Zhang, W. Zhang, J. Yu, L. He, J. Chen, and Y. He, "Complete and accurate holly fruits counting using YOLOX object detection," *Computers and Electronics in Agriculture*, vol. 198, Article ID 107062, 2022.

[9] W. Tian, M. Lauer, and L. Chen, "Online multi-object tracking using joint domain information in traffic scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 374–384, 2020.

[10] J. Zhu, K. Sun, S. Jia et al., "Urban traffic density estimation based on ultrahigh-resolution UAV video and deep neural network," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 12, pp. 4968–4981, 2018.

[11] X. Chen, S. Wu, C. Shi et al., "Sensing data supported traffic flow prediction via denoising schemes and ANN: a comparison," *IEEE Sensors Journal*, vol. 20, no. 23, pp. 14317–14328, 2020.

[12] X. Chen, Z. Wang, Q. Hua, W. L. Shang, Q. Luo, and K. Yu, "AI-empowered speed extraction via port-like videos for vehicular trajectory analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 4541–4552, 2023.

[13] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude UAV datasets: a survey," *Image and Vision Computing*, vol. 104, Article ID 104046, 2020.

[14] A. Bouguettaya, H. Zarzour, A. Kechida, and A. M. Taberkit, "Vehicle detection from UAV imagery with deep learning:

a Review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6047–6067, 2022.

[15] J. Yu, H. Gao, J. Sun, D. Zhou, and Z. Ju, "Spatial cognition-driven deep learning for car detection in unmanned aerial vehicle imagery," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 4, pp. 1574–1583, 2022.

[16] M. Kim, I. Kim, J. Yong, and H. Kim, "Scheduling framework for accelerating multiple detection-free object trackers," *Sensors*, vol. 23, no. 7, p. 3432, 2023.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[19] R. Girshick, "Fast R-CNN," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Washington, DC, USA, July 2015.

[20] S. Ren, K. He, R. Girshick, J. Sun, and R.-C. N. N. Faster, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[21] K. He, G. Gkioxari, P. Dollár, R. Girshick, and R.-C. N. N. Mask, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[22] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[23] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: exceeding yolo series in 2021," 2021, https://arxiv.org/abs/2107.08430.

[24] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2778–2788, Montreal, Canada, October 2021.

[25] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[26] H. Zhou, L. Wei, C. P. Lim, D. Creighton, and S. Nahavandi, "Robust vehicle detection in aerial images using bag-of-words and orientation aware scanning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7074–7085, 2018.

[27] Y. Xu, G. Yu, Y. Wang, X. Wu, and Y. Ma, "Car detection from low-altitude UAV imagery with the faster R-CNN," *Journal of Advanced Transportation*, vol. 2017, Article ID 2823617, 10 pages, 2017.

[28] T. Tang, S. Zhou, Z. Deng, H. Zou, and L. Lei, "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining," *Sensors*, vol. 17, no. 2, p. 336, 2017.

[29] S. Han, J. Yoo, and S. Kwon, "Real-time vehicle-detection method in bird-view unmanned-aerial-vehicle imagery," *Sensors*, vol. 19, no. 18, p. 3958, 2019.

[30] Z. Zhang, Y. Liu, T. Liu, Z. Lin, and S. Wang, "DAGN: a real-time UAV remote sensing image vehicle detection framework," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1884–1888, 2020.

[31] X. Luo, X. Tian, H. Zhang et al., "Fast automatic vehicle detection in UAV images using convolutional neural networks," *Remote Sensing*, vol. 12, p. 1994, 2020.

[32] R. Feng, C. Fan, Z. Li, and X. Chen, "Mixed road user trajectory extraction from moving aerial videos based on convolution neural network detection," *IEEE Access*, vol. 8, pp. 43508–43519, 2020.