High-altitude Multi-object Detection and Tracking based on Drone Videos

Qiang Zhao[1], Limei Peng[1,*]
[1]School of Computer Science and Engineering, Kyungpook National University, Daegu, 41566, Republic of Korea
* Corresponding author
Email: {zhaoqiang, auroraplm}@knu.ac.kr

Drone videos have more extensive shooting ranges, more angles, and no geographical limitations. Thus the object detection algorithm based on drone videos is increasingly playing a role in various fields, such as military surveillance, space remote sensing, smart city, disaster monitoring scenes, etc. Compared to low-altitude object detection and tracking (LA-ODT), high-altitude object detection and tracking (HA-ODT) are receiving increasing attention, especially in modern cities with massive high buildings, because of their higher flying h eight, w ider v iewing a ngle, a nd t he a bility t o t rack m ultiple f ast-moving o bjects s imultaneously. However, high-altitude aerial videos (HA-AVs) are constrained by small objects that can be measured, fewer feature points, occlusions, and light changes. Therefore, HA-AVs suffer from blurry images with fewer feature points of objects and missed detection due to occlusion, degrading the ODT accuracy. Since the accessible HA datasets are very limited, not to mention featured datasets considering angles, weather, etc., this paper directly uses drones to collect HA pictures and videos of different angles, different illuminations, and different heights for self-labeling training. Regarding this, we adopt super-resolution reconstruction to increase the data diversity and add artificial o cclusions t o e nhance t he c ollected d ata t o i mprove t he a ccuracy o f HA-ODT.

***Index Terms*— drone videos, neural network, Multi-object detection and tracking**

I. Introduction

Object detection and tracking (ODT) [1][2] is widely used in autonomous driving, surveillance security, traffic monitoring, robot vision, etc. [3]. Since drone aerial photography is not restricted by time and location, it is advantageous to obtain larger ranges and multiple angles, making it attractive to military reconnaissance, space remote sensing, smart cities, disaster monitoring, etc. Combining the ODT functions with drone aerial photography can control vehicle traffic flow [4], search and rescue missing persons, track sports scenes, etc.

The object tracking technology has developed rapidly and achieved promising tracking performance, thanks to the improved processing power of computers. *P. Viola et al.* realized for the first time the real-time face detection without any constraints (such as skin color segmentation) [5][6] in 2001, creating a precedent for object detection algorithms. Authors in [7] processed the drone videos to detect moving targets and reproduced the target moving trajectory in 2005. With the mature of manual feature selection technology, object detection has reached a stable development period after 2010. *R. Girshick et al.* proposed a region with CNN features (RCNN) for object detection [8] in 2014, opening the era of deep learning. *R. Joseph et al.* proposed the YOLO detection algorithm [9] in 2015, which is the first single-stage detection algorithm based on deep learning. *S. Ren et al.* proposed the faster RCNN detection algorithm [10], which is the first end-to-end and the first near-real-time deep learning detection algorithm. *A. Wahab et al.* proposed to detect moving vehicles based on a video image which is suitable for airborne and fixed cameras [11]. The proposed technique extracts the image feature points, removes the background by measuring the histogram changes of surrounding pixels of each feature point, obtains the foreground feature points, and then divides the feature points into different vehicle objects based on the motion characteristics.

On the other hand, the above work is mostly based on low altitude, where the images for detected objects are relatively large with high resolution. Object detection and tracking (ODT) based on high-altitude aerial videos (HA-AV) is still in its infancy and has a lot of issues to address. First, HA-ODT needs a large amount of data for training; nonetheless, there are very few accessible data sets since most of the available data sets are for LA platforms. Second, the long distance between the drones and the ground results in small objects with low resolution and single feature due to the vertically shooting of drones above the sky. Besides, pictures of small objects obtained from HA-AV are easy to miss feature points, interfered with by objects with similar shapes, and more vulnerable to occlusions, which degrade the ODT accuracy.

The paper is to address the above issues of HA-ODT. We use drones to collect massive data with different features, such as different angles, different illuminations, different heights, pictures, videos, etc., for self-label training. To address the issue of fewer object feature points, we use super-resolution reconstruction [12] to enhance data and add additional artificial occlusions such as trees, lights, bridges, etc., to increase the data diversity for improving the training accuracy as well as the multi-object tracking accuracy. The rest of this paper is organized as follows. Section II introduces the traditional object detection and tracking algorithms. Section III introduces the proposed methods of super-resolution and adding artificial occlusions. Section IV shows the experiment settings and

Manuscript received October 10, 2021; revised February 24, 2022. Corre-sponding author: Limei Peng (email: auroraplm@knu.edu.kr).

results. Section V summarizes the paper.

II. Preliminary

A. Object detection algorithms

The convolutional neural networks (CNN)-based multi-object detection algorithms [1] are divided into one-stage and two-stage ones. The CNN-based object detection processing can be divided into two steps, i.e., extracting the picture features and the areas where objects may exist and then performing content classification and object frame regression from the extracted areas.

The one-stage algorithms omit the regional suggestion network and directly predict the object category classification and the regression of the location box from the feature map. Under the same calculation capacity, the one-stage algorithms are faster than the two-stage ones but have lower detection accuracy. R-CNN is a representative two-stage algorithm [13]. Faster R-CNN [8] is an upgraded version, which can be adapted to different scenarios, different scales, different appearances, and other complex situations, but with severe background interference and is applicable to small-scale object detection.

Since drones are restricted by the available electrical power to support the flight duration, they solicit compact algorithms with less memory and shorter inference time. Therefore, the object detection algorithms for drone HA-AVs mostly adopt the one-stage detection method. YOLO [9] is the most commonly used one-stage detection model, including its improved versions of Tiny-YOLO[14], SlimYOLO[15], YOLOV3-Tiny[16], etc. It can improve the detection accuracy and speed fast by cutting the network model to reduce the network parameters.

This paper adopts another more accurate and fast one-stage algorithm, i.e., Efficientdet-D2 [17], considering the object to be detected is small and the change in each frame is large. The specific network structure of Efficientdet-D2 is shown in Fig. 1. Note that we also tried to use other models such as Efficientdet-D3-D7 declared with higher accuracy, but based on the experimental results, the Efficientdet-D2 model is more suitable for this study. Efficientdet-D2 is improved from Efficientdet which uses the Image-Net pre-trained EfficientNets [18] as the skeleton network, as shown in Fig. 1. As a feature network, BiFPN uses layers 3-7 as input to repeatedly use top-down and bottom-up bidirectional structures for feature fusion. These fused features are output to the class and box network to predict the coordinates of the object category and bbox.

B. Object tracking algorithms

Multi-object tracking [5] refers to multiple detection, extraction, recognition and tracking of objects in an image sequence. People can better understand and describe the behavior of the object by obtaining the motion parameters such as the position, speed, and trajectory of the object. The more popular Multi-object detection algorithms at this stage first use visual features for association matching, then use Kalman filtering [23] to remove abnormal motion trajectories, and finally use IoU tracking as a supplement. But after testing, we found that the actual effect is not ideal. The main reason is that the drone's high-altitude aerial video flight height is relatively high, resulting in smaller target objects, unobvious visual features, and more similar objects. Since there is basically no occlusion between various objects and the movement law is obvious, this paper chooses to use the IoU of the detected objects in the adjacent video frames as the main basis for tracking. Besides, we use the Kalman filter to repair the trajectory disconnection caused by the recognition failure or the viaduct occlusion for multi-object tracking.

III. Super-resolution reconstruction & Data enhancement

A. Super-resolution reconstruction algorithms

Due to the limited performance of the equipment, the long shooting distance, the poor resolution of the output videos and pictures, and the blurred objects to be detected, we use super-resolution (SR) methods to enhance the feature values of the objects to be detected, so as to improve the detection and tracking accuracy.

Image SR reconstruction technology is divided into two types [5], one is to synthesize a high-resolution image from multiple low-resolution images, and the other is to obtain a high-resolution image from a single low-resolution image. Based on CNN model, super-resolution convolutional neural network (SRCNN)[19] first introduced CNN into single-image super-resolution (SISR)[20], it achieved advanced and promising results by only using a three-layer network.

SISR Models based on deep learning are roughly divided into the following two major directions. One aims to recover the image details using structural similarity (SSIM), peak signal-to-noise ratio (PSNR), and other evaluation standard algorithms, among which the SR-CNN [19] model is representative. The other does not care about the details and aims to reduce the perceptual loss and looks at the big pictures. The representative algorithm is the super-resolution generative adversarial network (SRGAN) [21].

In this paper, we use the RealSR model [22] in the SRGAN network to improve the feature values of the image to be detected. RealSR is advantageous in the following aspects, compared to the existing super-resolution methods. First, RealSR uses a self-designed new image degradation method to simulate the degradation process of real pictures by analyzing blur and noise in real pictures. Second, paired training data is not required, and thus unlabeled data can be used for training. Third, it can deal with the problems of blur and noise in low-resolution images and therefore gets more precise and cleaner high-resolution results. The specific network structure is shown in Fig. 2.

B. Data enhancement by adding artificial occlusions

Drone high-altitude aerial videos have the features of small areas and a large number of objects due to high altitude. Regarding this, we first enlarge the images in the reasoning test, and then compare the detection results. On some roads,

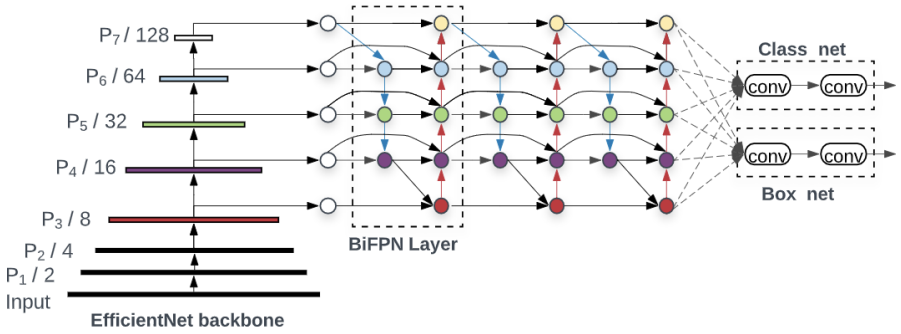

Fig. 1. EfficientdetD2 Algorithm structure

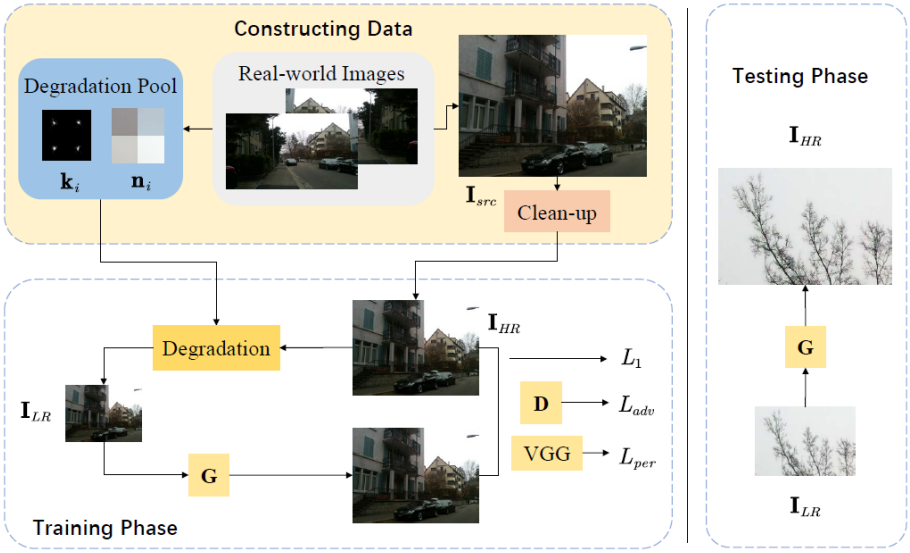

Fig. 2. RealSR Algorithm structure

due to the occlusion of obstacles such as trees, bridges, and street lights, the detection frame is lost during object detection or is not detected for tracking. To this end, we use image synthesis technology to superimpose and synthesize occluded objects artificially, label some of the occluded objects to conduct special training, and increase the occlusion training samples. In the training process, various data enhancement methods such as splicing, rotation, and flipping are also used, and the effect is shown in Fig. 3. The picture shows the relevant data set made with the street lights on the viaduct as the primary obstruction. In addition, we also make some data sets with the overpass and trees as the obscuration.

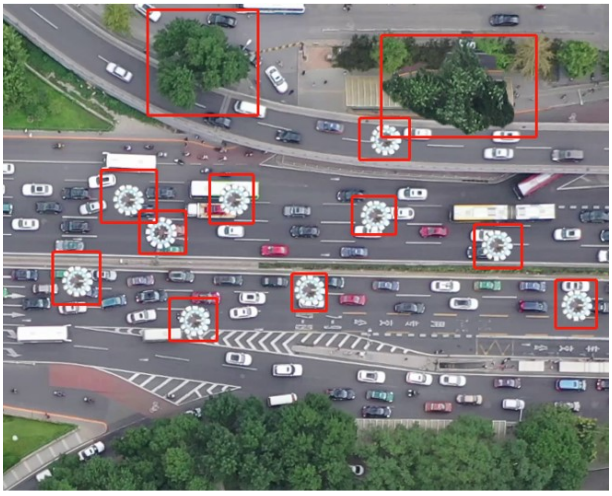

Fig. 3. Increase the occluded image

IV. Simulation and Result Analysis

A. Data sets and Pre-labeled

Even though we can find a few public drone-aerial data sets collected in low altitudes, we can hardly find any collected in high altitudes. Note that a shooting height of no more than 100 meters is called low altitude and high altitude above that. To conduct our experiment on HA-ODT, we are authorized to shoot relevant videos suitable for this experiment in areas

such as intercity highways, national highways, bridges, intersections, ramps, and label them.

The shooting environment covers from day to night, including the occlusions and scale changes of the real environment. We collect a total of 20,000+ pictures and 200+ videos. In order to increase the data diversity, we also download from the Internet some relevant pictures that can meet the requirements of this experiment for labeling and training.

In order to speed up the labeling process, we used relevant labeling software written in C++, specifically for the trajectory labeling of vehicles and pedestrians in aerial video. We first use the detection algorithms to get a rough detection frame, and then perform fine labeling. This greatly improves the efficiency of labeling. The marking software supports marking the vehicle position, length and width, attributes, category, and heading direction. We can export various data set formats such as Darknet, COCO, Pascal VOC, etc.

The results are shown in Fig. 4-7.Fig 4 is the UI interface of the labeling software. Our data set annotation and production are all done here. It can produce data sets of different formats and sizes according to different needs. Fig. 5 is the part of the picture being marked.

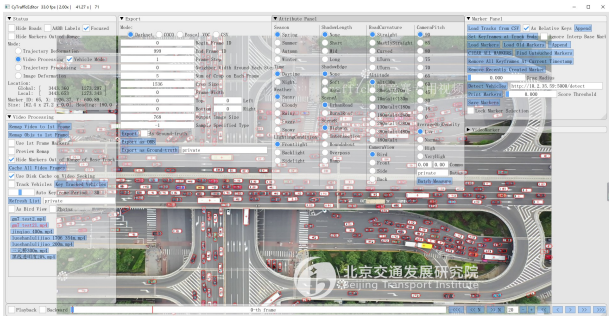

Fig. 4. Dataset annotation tool

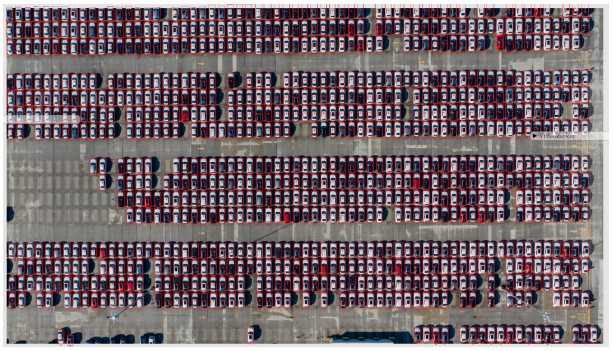

Fig. 5. Complete the labeled image data set

Fig. 6 is a part of the data we marked according to different light and different environments, aiming to investigate how these parameters would affect the detection accuracy under different illumination. To this end, we collect as much as possible the relevant videos and images of different weather and different time periods. The collected data includes different

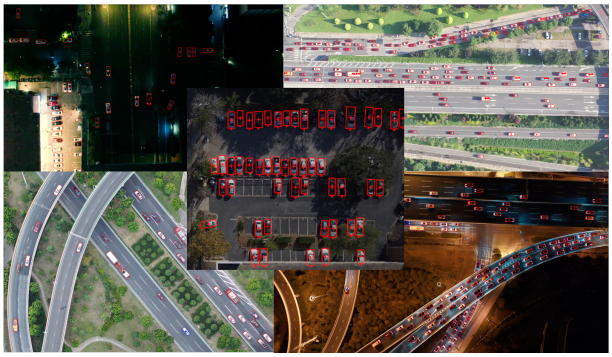

Fig. 6. under different lighting data set

light angles from day to night, different weather, different flying heights, and different vehicle types. Fig. 7 shows some of the video files in different scenes we shot.

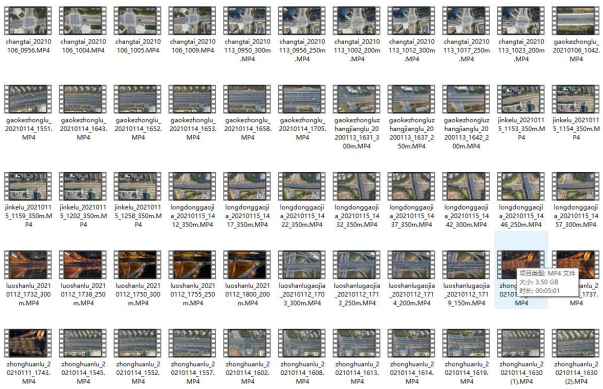

Fig. 7. Vehicle video data set

We compare the results of several popular super-resolution algorithms and finally choose the RealSR algorithm with the best results. The comparison chart before and after applying super-resolution is shown in Fig. 8.

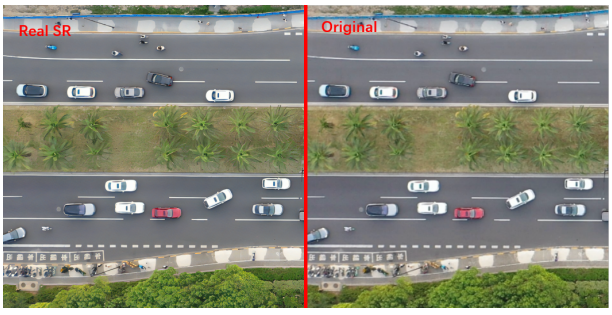

Fig. 8. Images before and after super resolution

The left picture has a resolution of 3840x2160 after super-resolution, and the right picture is an original one with an unprocessed 1920x1080 resolution. The resolution of the super-resolution image is higher, which can solve the problem of low detection accuracy result caused by the blurred image of the captured image to a certain extent.

B. Evaluation metrics

For the one-stage Efficientdet-D2 algorithm[16] used in this paper, we select 500 images under different scene lighting as the testset. As shown in Fig. 9, the Efficientdet-D2 model has an average precision rate of 82.82% in the test set when the intersection over union (IOU) is 50%. Note that IOU refers to the ratio of the intersection and union of the target prediction bounding box and the groundtruth bounding box. It is calculated by dividing the overlapping part of the two regions by the collective part of the two regions. Generally speaking, IOU larger than 50% can be considered a good result, which is the reason we use 50% in our experiment.

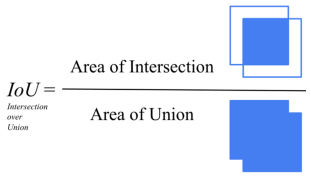

Fig. 9. IOU

The metric to evaluate the detection accuracy is the mean average precision (mAP), which is the average value of each category of average precision. AP is the evaluation index of the mainstream target detection model, which refers to the area under the Precise-Recall rate (P-R) curve. Precision is defined as P = TP/(TP+FP) and recall rate is defined as R = TP/(TP+FN), where TP, TN, FP, and FN represent the numbers of positive types predicted as positive, negative types predicted as negative, negative types predicted as positive, and positive types predicted as negative.

For multi-object tracking, we use the Multiple Object Tracking Accuracy (MOTA) indicator to measure the tracking accuracy, which is defined in Eq. (1). Note that higher value of MOTA indicates higher tracking accuracy.

$$MOTA = 1 - \Sigma_t(m_t + fp_t + mme_t)/\Sigma_t g_t \qquad (1)$$

where m_t, fp_t, mme_t, g_t, and t represent the missed number, number of false positives, and number of mismatches, number of groundtruth, and number of corresponding frames.

C. Results on multi-object detection

The experimental results are shown in the Fig. 11. We compare the methods of whether using super resolution (SR) and data enhancement (DE) based on adding artificial occlusions, , i.e., EfficientdetD2, SSD, Faster-rcnn, and YOLOV4. The following observations can be obtained. We can see that Efficientdet-D2 performs best in multi-target detection compared with models such as YoloV4, SSD, and Faster-RCNN. Take the Efficientdet-D2 algorithm as an example; after super-resolution (SR) is used, the detection accuracy is improved by 2.9% compared to before using it without super-resolution; other algorithms have large and small accuracy after using super-resolution. Upgrade, which shows that it is feasible to use super-resolution to solve the image blur problem and improve the detection accuracy. The use of super-resolution and high-altitude data collection (HADC) and additional occlude (OC) simultaneously helps increase the detection accuracy by 5.5%. In this experiment, to reflect fairness, our algorithms only took the same part of the data for experimentation. There is still much room for improvement in the accuracy of the algorithm. In addition, the quality of the super-resolution model will also affect the results of the experiment.

The detection result is shown in Fig. 11 obtained by using the single-stage Efficientdet-D2 [16]. We select 100 images under different scene lighting as the test set. In order to test the effectiveness of data enhancement and super-resolution in improving the object detection accuracy, we also evaluate several other common and efficient object detection algorithms for comparison, including faster-rcnn, single short multibox detector (SSD), and YoloV4. For each algorithm, we compare the three schemes differing in whether super resolution (SR) and data enhancement(DE), i.e., without SR + DE, with SR, with SR + DE.

We can get the following observations from Fig. 10. First, the selected EfficientdetD2 has the highest accuracy when compared to the other counter parts, for all the three schemes, nonetheless, with a relatively longer time. For all the schemes, the proposed SR + DE scheme performs the best.

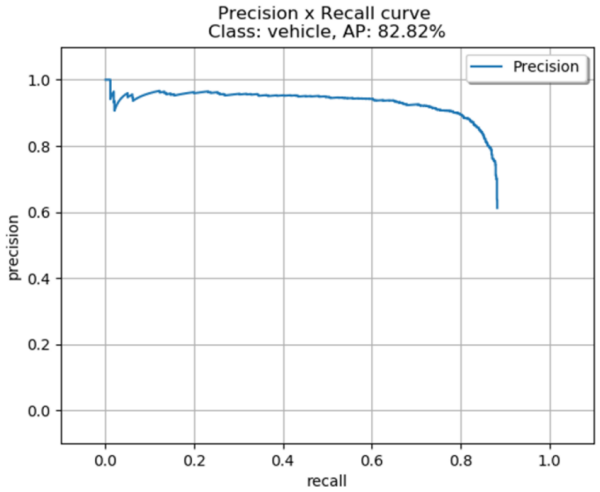

Fig. 10. OUR Dataset Efficientdet IOU50 mAP

HAdatset detection mAP

Algorithm / Method	EfficientdetD2	SSD	Faster-rcnn	YOLOV4
Without SR+ data enhancement	77.3%	73.5%	75.1%	75.4%
With SR	80.2%	77%	78.1%	78.3%
With SR+ data enhancement	82.8%	79.5%	80.4%	81.1%

Fig. 11. Comparison of algorithm results

D. Results on multi-object tracking

After we get the detection result, we use the IOU of the object detected in the adjacent video frame as the primary basis for tracking and use the Kalman filter to repair the missing trajectory caused by the recognition failure or the viaduct occlusion. We use the detection results of each detector for object tracking, and the result largely depends on the accuracy of the detector. The tracking results are shown in Table. 2 as an example, according to the data in the Table. 2. We can see that Efficientdet-D2 performs best in multi-object tracking compared with models such as YoloV4, SSD and Faster-RCNN. But the SSD algorithm has the lowest accuracy in multi-object tracking. Take the Efficientdet-D2 algorithm as an example. After super-resolution (SR) is used, the tracking accuracy is increased by 3.8% compared to before it is used, other algorithms use super-resolution, and the accuracy is greatly improved. Simultaneous use of super-resolution and high-altitude data acquisition (HADC) and additional occlude (OC) can help improve tracking accuracy by 6.9We use the results of the detection algorithm to track, so the tracking results largely depend on the accuracy of the detection algorithm.

After we get the detection result, we use the IoU of the object detected in the adjacent video frame as the main basis for tracking, and use the Kalman filter to repair the missing trajectory caused by the recognition failure or the viaduct occlusion. We use the detection results of each detector for target tracking, and the final result largely depends on the accuracy of the detector. The tracking results are shown in fig 11 as an example.In fig 12, we show the relevant result picture of the tracking video. Because the targets tracked at the same time are too dense, we use different colored detection frames instead of the vehicle ID to display them. The green arrow is the area where the target object predicted by the Kalman filter may appear in the future.In the fig, we can see that there are still some false detections. In the follow-up, relevant training of the model is also needed to improve the accuracy.

HAdatset tracking MOTA

Algorithm / Method	EfficientdetD2 + SROT	SSD + SORT	Faster-rcnn+ SORT	YOLOV4 + SORT
Without SR+ data enhancement	57.5%	48.3%	52.6%	53.2%
With SR	61.3%	51.1%	56.4%	58.1%
With SR+ data enhancement	64.4%	54.8%	60.1%	63.2%

Fig. 12. Tracking algorithm results

V. Conclusion

In this paper, we use drones to collect a large amount of relevant data and perform training after annotation to solve the problem of fewer public data sets related to high-altitude drone videos. And the use of super-resolution and the increase

Fig. 13. Efficientdet D2 + sort Tracking image

of occluders for special training and other methods have solved the problems of low resolution of image blur caused by flying height problems and missed detection caused by occlusion of occluders. According to the test results, the accuracy of the model that has been specially trained for super-resolution and occlusion is higher than that of the original model without training. However, there are still some problems in this article that need to be improved. 1. Difficulty in data collection, resulting in small data diversity, few large vehicle samples, and over-fitting in multiple training sessions. 2. The quality of the super-resolution model will affect the accuracy of the detection and tracking algorithm to a certain extent.3. In order to improve the accuracy of the target detection algorithm, the detection speed is sacrificed to a certain extent, and the detection and tracking speed needs to be further improved. At the same time, we are also looking for a detector that can balance detection accuracy and speed for future research. Based on these problems, we will continue to study and solve these problems in future research.

References

[1] Zou Z, Shi Z, Guo Y, et al. Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055, 2019.
[2] Yilmaz A, Javed O, Shah M. Object tracking: A survey. Acm computing surveys (CSUR), 2006, 38(4): 13-es.
[3] Gordon R, Herman G. Three-dimensional reconstruction from projections: A review of algorithms. International review of cytology, 1974, 38: 111-151.
[4] Abderrahmane A, Bin Y, and Tarik T. Generalized Traffic Flow Model for Multi-Services Oriented UAV System. Journal of Networking and Network Applications,Vol.1, Iss.1, pp.1–8. https://doi.org/10.33969/J-NaNA.2021.010101
[5] Viola P,Jones M. Rapid object detection using a boosted cascade of simple features. Computer Vision and Pattern Recognition(CVPR), 2001. vol.1. IEEE, 2001, pp.1-1.
[6] Viola P, Jones M. Robust real-time face detection. International journal of computer vision, vol.57, no.2, pp.137–154, 2004.
[7] Ali S, Shah M. COCOA: tracking in aerial imagery.Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications III. International Society for Optics and Photonics, 2006, 6209: 62090D.
[8] Girshick R, Donahue J, Darrell T. Regionbased convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence, vol.38, no.1, pp.142– 158, 2016.
[9] Redmon J, Divvala S, Girshick R.You only look once: Unified, real-time object detection. IEEE conference on computer vision and pattern recognition, 2016, pp.779–788.

[10] Ren S, He K, Girshick R, and Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. in Advances in neural information processing systems, 2015, pp. 91–99.
[11] Abdelwahab M A, Abdelwahab M M. A novel algorithm for vehicle detection and tracking in airborne videos. IEEE International Symposium on Multimedia (ISM). IEEE, 2015: 65-68.
[12] Park S C, Park M K, Kang M G. Super-resolution image reconstruction: a technical overview. IEEE signal processing magazine, 2003, 20(3): 21-36.
[13] Girshick R. Fast r-cnn. IEEE international conference on computer vision. 2015: 1440-1448.
[14] Khokhlov I, Davydenko E, Osokin I, et al. Tiny-YOLO object detection supplemented with geometrical data. 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring). IEEE, 2020: 1-5.
[15] Zhang P, Zhong Y, Li X. SlimYOLOv3: Narrower, faster and better for real-time UAV applications. International Conference on Computer Vision Workshops. 2019: 0-0.
[16] Redmon J,Farhadi A. YOLOv3: an incremental improvement (2018). arXiv preprint arXiv:1804.02767 (1804). https://pjreddie.com/darknet/yolo/.
[17] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection. IEEE CVF conference on computer vision and pattern recognition. 2020: 10781-10790.
[18] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning. PMLR, 2019: 6105-6114.
[19] Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(2): 295-307.
[20] Glasner D, Bagon S, Irani M. Super-resolution from a single image. 2009 IEEE 12th international conference on computer vision. IEEE, 2009: 349-356.
[21] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.
[22] Ji X, Cao Y, Tai Y, et al. Real-world super-resolution via kernel estimation and noise injection. Conference on Computer Vision and Pattern Recognition Workshops. 2020: 466-467.
[23] Grewal M S, Andrews A P. Kalman filtering: Theory and Practice with MATLAB. John Wiley & Sons, 2014.