AFTD: A FOREGROUND TARGET DETECTION DATASET FOR AIRPORT SCENARIO

Anonymous authors
Paper under double-blind review

ABSTRACT

The detection of foreground targets on airport surface is the foundation of airport surveillance applications. However, effective algorithms and specialized benchmarks are still lacking in this area. Based on this fact, we propose an Airport Foreground Target Detection dataset (AFTD), which contains the three most important foreground targets moving on the airport surface: airplane, vehicle, and person. Through self collection and collection of web images, we have obtained a total of over 200000 images and filtered out 10050 images based on diversity principles to form the AFTD dataset, which includes a total of 26968 airplane instances, 24759 vehicle instances, and 5064 person instances. AFTD includes a variety of changes of these targets, such as super multi-scale, multi-level occlusion and viewangle changes, etc. In addition, we further illustrate the challenges posed by AFTD to existing algorithms through statistical analysis and detailed experiments, and discuss how to solve these challenges in the airport surveillance scenario.The AFTD dataset can be downloaded from http://www.agvs-caac.com/aftd/aftd.html.And our moduel is available at https://github.com/cpc1111-lab/Additional-discussion-for-ATFD.

1 INTRODUCTION

In recent years, with the rapid growth of global population and economy, airports have experienced a significant increase in passenger and cargo traffic. Meanwhile, there has been an upward trend in the number of security incidents related to movable targets on airport surface. For example, on January 2, 2024, two planes collided at Haneda Airport in Japan, resulting in five deaths, and on January 16, two planes collided again at New Chitose Airport in Hokkaido, Japan. Therefore, it is particularly important to develop intelligent applications to enhance the surveillance of foreground targets on the airport surface.

Object detection is widely used in airport surveillance applications. However, we found that the performance of object detection algorithms has decreased to varying degrees in airport scenarios, and subsequent experiments have also proven this. This is because fundamental research emphasizes the generalization of algorithms, and thus classical datasets usually contain rich object categories, but without considering all states of each class. Practical applications, such as airport surveillance, do not focus on algorithm generalization, but rather on robustness to variations of specific targets. In other words, the study of airport surveillance requires a dataset with rich samples for airport foreground targets. Obviously, classical datasets do not meet this requirement.

In this paper, we propose the first Airport Foreground Target Detection dataset, AFTD, for airport surveillance. AFTD only includes three types of movable targets on the airport surface, but strives to cover variations of each type of target. We obtained permission to collect data at multiple airports through cooperation with the institute of civil aviation, and fully utilized network resources, ultimately collecting over 200000 raw images. We then rigorously filtered the data based on the diversity principle. Due to the high similarity of airport surface structures, only 10050 images were retained, totaling 26968 airplane instances, 24759 vehicle instances and 5064 person instances. In addition, we designed more refined label format for airplane. For example, for the airplane category, we designed three scale formats, three occlusion formats and eight viewangle formats. All images are manually annotated following strict rules. Finally, we fully tested the AFTD dataset and analyzed the experimental results.

In summary, our main contributions to the field of object detection are:

- We have established a dataset AFTD for airport surveillance, which covers various challenges of movable targets on the airport surface.
- We illustrated the challenges posed by AFTD to existing algorithms through statistical analysis and detailed experimentst.
- We discussed how to solve these challenges in the airport surveillance scenario.

2 RELATED WORK

2.1 OBJECT DETECTION DATASETS

In the past more than ten years, many classical object detection datasets have been proposed, such as VOC2007Everingham et al., VOC2012Everingham & Winn (2012), ImageNetDeng et al. (2009), MS-COCOLin et al. (2014), Objects365Shao et al. (2019), etc.

The VOC2007 dataset consists of 5k images and 12k annotated objects. The VOC2012 dataset consists of 11k images and 27k annotated objects. These two datasets annotate 20 common objects in daily life. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) has promoted the technological level of universal object detection, with the ILSVRC detection dataset containing 200 categories of visual objects. The MS-COCO dataset has fewer object categories than ILSVRC, but it has more object instances and contains more small objects and more densely located objects. At present, the MS-COCO dataset has become the de facto standard in the field of object detection.

None of the above datasets specifically consider airport scenarios. The FGVC airplane datasetMaji et al. (2013) is an airplane classification dataset that contains 10200 airplane images, with 100 images of each of the 102 types of Boeing airplanes. The majority of the airplanes in the FGVC-airplane dataset are presented in sideways view, with a very single view viewangle and no scale variation.

2.2 OBJECT DETECTION ALGORITHMS

Since the successful application of AlexNetKrizhevsky et al. (2012) in 2012, Convolutional Neural Networks (CNN) have gradually become mainstream. Object detection has been revolutionized by CNN, experiencing a shift from traditional methods to deep learning methods.

R-CNNGirshick et al. (2014) generates candidate frames and further refines classification and localization through a region proposition network, which marked the rise of two-phase detection. SPPNetHe et al. (2015) avoids repeated calculation of convolution features by introducing a spatial pyramid pooling layer that generates a fixed-length representation of a candidate region frame of arbitrary size. Fast RCNNGirshick (2015) enables neural networks to train both classification and regression tasks simultaneously, further reducing redundant calculations. Faster RCNNRen et al. (2015) is the first near real-time object detector to replace traditional candidate region extraction methods by introducing a region proposal network.

YOLO seriesRedmon et al. (2016)Redmon & Farhadi (2017) adopts an end-to-end approach to predict both bounding box and category probabilities, pioneering the paradigm of one-stage object detection algorithms. SSDLiu et al. (2016) improves the detection accuracy of one-stage detectors by detecting objects at different scales on different layers of the network. RetinaNetLin et al. (2017) reshapes the standard cross-entropy loss by introducing a new loss function, focal loss, which allows the detector to focus more on hard, misclassified samples during training. CornerNetLaw & Deng (2018) detects each bounding box as a pair of keypoints. CenterNetDuan et al. (2019) treats an object as a single point and regresses all its attributes.

DETRCarion et al. (2020) first applies the TransformerVaswani et al. (2017) to the field of object detection, directly predicting bounding boxes and categories through encoder-decoder structure, eliminating anchor, NMS, and other steps. Deformable DETRZhu et al. (2020) reduces computational complexity by introducing a deformable attention module, and improves performance for small objects by using multi-scale feature maps.

3 AFTD BENCHMARK

The AFTD dataset focuses on three types of movable targets on the airport surface: airplane, vehicle, and person, the movements of which are the primary causes of various airport security incidents. We will introduce how we build the AFTD dataset from three aspects: dataset design, data collection, data annotation.In addition,we illustrate the challenges raised by AFTD through statistical analysis.

3.1 DATASET DESIGN

The principle of dataset design is to try to cover as many variations or challenges of the target of interest as possible. We will introduce the challenges for airplane, vehicle and person in turn.

1) airplane are our primary target of interest. We have summarized six main challenges to airplane detection in airport scenarios, each of which is described below.

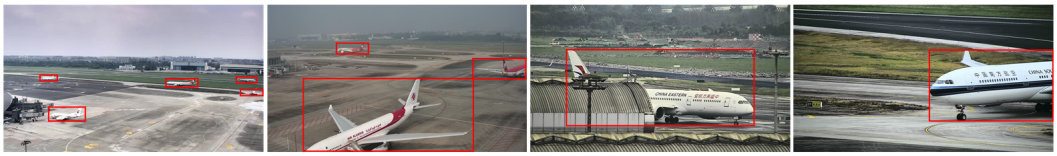


Figure 1: (a) and (b) show airplane presented at different scales on the screen. (c) and (d) show the situation where airplane is obstructed by the airport terminal and part of the fuselage is not on the monitoring screen.

Super multi-scale. Compared with the scale change in ordinary monitoring scenes, the scale of airplane in airports varies greatly on the imaging plane, as shown in Figure 1(a) and 1(b). This is due to the airport has a vast area, with monitoring distances spanning the range of ten meters to kilometers. When the scale of the airplane is too large, details may be magnified, resulting in the algorithm being unable to effectively capture the global features of the object. When the scale of the airplane is too small, it is difficult to effectively capture the local features of the object. In addition, there is a multi-scale coexistence in the airport, as shown in Figure 1(b), where very large and small objects occur simultaneously. Because the solutions to objects at different scales may be incompatible, the simultaneous appearance of multiple scales poses greater challenges.

Multi-level occlusion. In airport, there is a frequent occurrence of airplane being obstructed by other objects, such as bridges, adjacent airplane, etc. and it is also frequent that part of the airplane's body is outside the screen, as shown in Figure 1(c) and (d). Occlusion often leads to incomplete feature information of the object, or even results in a serious lack of key identification features.

Viewangle changes. The appearance of the airplane varies tremendously from one viewangle to another because of the wingspan shape. The shape of is completely different in the side and front and rear viewangle, as shown in Figure 1. The characteristics of the airplane may be more varied when the viewangle occurs in conjunction with other changes, thus increasing the difficulty in detection.

Figure 2: (a) and (b) show the evening and night scenes. (c) and (d) show the foggy and snowy day.

Illumination changes. Airports, as a typical outdoor scene, make all the data in AFTD subject to illumination variations, as shown in Figure 2(a) and (b). Weak illumination at night significantly reduces the contrast of the image, as shown in Figure 2(b), resulting in the silhouette of the airplane being difficult to recognize, which may lead to missed detection.

Weather changes. There are various weather changes in airports and different weather conditions have a significant impact on the visibility. In foggy days, as shown in Figure 2(c), the image contrast is sharply attenuated, and the silhouette of the object become blurred or even indistinguishable. During snowy days, as shown in Figure 2(d), the high degree of similarity in hue between the color of the airplane fuselage and the snow-covered ground greatly increases the difficulty in distinguishing the airplane from the background, resulting in a decrease in the accuracy of the boundary localization.

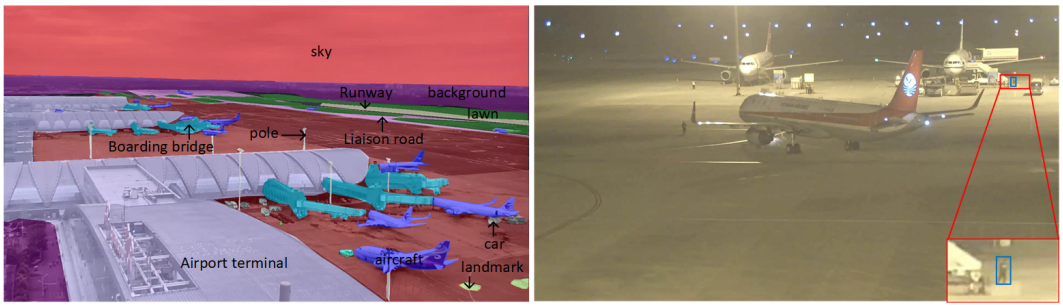


Figure 3: (a) shows the complex background of the airport surface. (b) shows that due to the long imaging distance, person's silhouette is difficult to identify.

Complex background. In addition to movable targets, there are a large number of buildings and greenery in and around the airport, which constitute a complex visual background, as shown in Figure 3(a). Some objects on the airport surface, such as high pole lights, signs, etc., are not our focus and most of them are stationary. Therefore, we will also include them in the background part.

Figure 4: From (a) to (d) shows some vehicles we observed while collecting data.

2) Vehicle and person are also the main movable targets on airport surface. Some of the functions they provide, such as cargo handling and runway inspection, are important part of airport production activities. Although vehicle and person are common object categories, their attributes in airport surface differ greatly from those in other scenarios. Firstly, due to the wide range of airport surveillance beyond conventional traffic scenarios, vehicles and people are far away from surveillance cameras, resulting in a significant reduction in the size of the imaging. Therefore, small objects such as person is often poorly outlined or even difficult to recognize, as shown in Figure 3(b). Secondly, vehicles in airports contain various types such as tractor-trailers, fuel trucks, etc, as shown in Figure 4. However, since their movement behaviors are similar, we collectively refer to such movable targets as "vehicles" at present. If necessary in future research, we will differentiate different types of vehicles based on specific needs.

3.2 Dataset Acquisition

Through cooperation with a institute of civil aviation, we have been granted permission to collect data at several airports. The collection equipment consists of multiple fixed cameras and PTZ cameras, with image resolutions of 1920 * 1080, 1280 * 720, etc.

1) Raw data collection. The central principle of data collection is to ensure the diversity of movable target patterns and backgrounds, covering all previously mentioned challenges, each with rich examples. For example, for airplane categories, by continuously adjusting the viewangle and focal length of the PTZ camera, try to cover various imaging distances, viewangles and activities within the camera monitoring range. In order to ensure diversity against certain challenges, such as weather and

Table 1: Definition of the occlusion degree and scale in the airplane annotation process.

Occlusion			Scale		
Percentage of occlusion	Degree of occlusion	Abbreviations	Percentage of airplane size	Scale	Abbreviations
none	No occlusion	0	<8%	Small scale	s
0% - 30%	Mildly occlusion	1	8% - 25%	Medium scale	m
30% - 70%	Medium occlusion	2	>25 %	Large Scale	x
70% - 100%	Heavy occlusion	3			

illumination changes, the collection cycle has been continuously extended. We collect as much raw image data as possible for subsequent screening. In over half a year, we have collected approximately 100000 raw image data in total.

Figure 5: (a) and (b) shows images taken from two separate airports, but with some similarities. (c) and (d) are web images.

2) Data screening. The purpose of data filtering is to minimize data redundancy while ensuring diversity. The data collection and the data filtering of the AFTD dataset are extremely challenging. Although we conducted data collection at several different airports, since the airports are very similar in terms of layout, facility configuration, etc., when we analyzed the collected data, we found that the images collected from different airports were highly similar, and even unable to distinguish which airport they came from, as shown in Figure 5(a) and (b). In addition, due to the high standardization of airport production activities, the activities of movable targets are also very similar, so the redundancy of data is very high. After screening, we only retained 5050 out of over 100000 raw images. Opening more data collection points can increase the number of effective samples. However, in addition to the semi-military restrictions of airports, the number of airports is also relatively limited. There are only 255 civil airports in the whole of China.

3) Data expansion. Due to the difficulty in opening airport data sources, we decided to enrich the AFTD dataset by collecting public images of domestic and foreign civil airports from web resources. We collected approximately 100000 images from the internet and followed similar principles as in the previous section to filter the data. In addition, we especially considered the imaging angle. Web images were removed if their imaging angles, especially the pitch angle, differed significantly from the self-collection data. For example, satellite images viewed directly downwards were all deleted. After filtering, only 5000 out of over 100000 web images were retained. The web images did enrich the diversity of the dataset to some extent, as shown in Figure 5(c) and (d).

In summary, the AFTD dataset currently contains 10050 image data. Although it has been possible to conduct experiments on this dataset and draw the conclusions, the amount of data is still slightly insufficient. We plan to enrich the data in two ways in the future. Firstly, we plan to open up more data collection points. However, due to the high similarity of airport structures and the necessity of obtaining permits, this work may be laborious. Secondly, we plan to collect data at the simulated validation airport, such as the one to be built in Chengdu in 2025.

3.3 Data Annotation

We use the annotation tool, AnyLabeling, to annotate the AFTD dataset. The label format of groundtruth is <class, $x_{\max}$, $x_{\min}$, $y_{\max}$, $x_{\min}$, w, h>. The values of the class are 0, 1, and 2, respectively representing airplane, vehicle, and person. ($x_{\max}$, $x_{\min}$, $y_{\max}$, $x_{\min}$) represent represents the four vertices of the bounding box. (w, h) respectively represent the resolution of the image which contain the bounding box. Basic label information can be obtained by drawing rectangular boxes for objects, as shown in Figure 6(a).

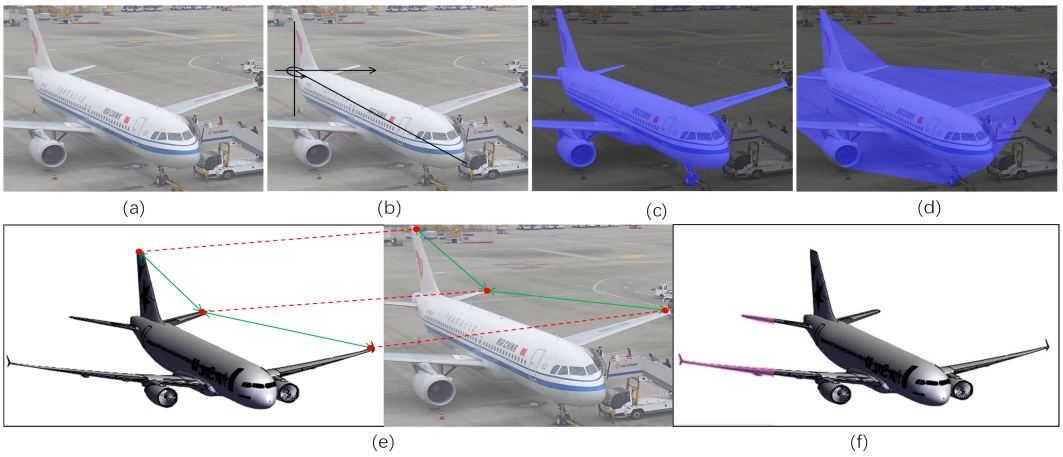
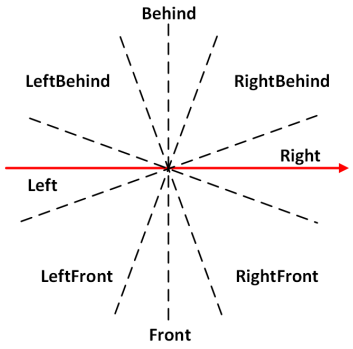


Figure 6: Annotation: (a) shows the airplane to be annotated; (b) shows the segmentation masks of the airplane; (c) shows the outline of key points of the airplane; (d) shows the airplane viewangle; (e) shows finding an approximate 3D model and projecting the 3D model to a 2D plane; (f) shows roughly calculating the proportion of obscured.



Orientation Angle	Viewangle	Abbreviations
-20° ~ 20°	Right	r
20° ~ 70°	RightBehind	rb
70° ~ 110°	Behind	b
110° ~ 160°	LeftBehind	lb
160° ~ 200°	Left	l
200° ~ 250°	LeftFront	lf
250° ~ 290°	Front	f
290° ~ 340° (-20°)	RightFront	rf

Figure 7: Orientation angle and viewangle division.

In addition, AFTD is a application-oriented dataset with the goal of covering the variations of limited movable targets, especially airplane, so we added three additional labels for airplane for more granular description, with the format of <viewangle, occlusion, scale>, which will contribute to design of detection algorithms for foreground targets on airport surface.

The definition and labeling of these three additional labels are described next.

Viewangle. The viewangle refers to the angle of orientation of the airplane in the image, which is defined by measuring the angle between an imaginary line from the tail to the head of the airplane and the positive X-axis direction. The division of viewangle is shown in Figure 7. As shown in Figure 6(b), the viewangle of the airplane is "RightFront", denoted as "rf".

Occlusion. The occlusion refers to the proportion of the unpresented parts to the complete area of the airplane from the same viewangle. The division of occlusion degree is shown in Table 1. In order to calculate the degree of occlusion, we need to obtain the area of the airplane's segmentation mask and the approximate full area at the same viewangle. We will label twice to obtain accurate segmentation masks and key location silhouette respectively, as shown in Figure 6(c) and (d). With keypoint mapping, we can adjust and scale a 3D airplane model and project it onto a 2D plane to obtain the pseudo-real mask area, as shown in Figure 6(e). Then we can calculate the degree of occlusion based on Equation1, where $Degree$ denotes the degree of occlusion, S_F denotes the full area and S_S denotes the segmented area. The obscured part is shown in Figure 6(f) is mildly obscured,

denoted as "1".

$$Degree = (S_{Full} - S_{segmentation})/S_{Full}. \quad (1)$$

Scale. Since the image sizes in the AFTD dataset are not uniform, we use relative scale to define the scale of the airplane. By calculating the proportion of the complete area of the airplane to the image area, rather than directly dividing it based on the bounding box area obtained from annotations. This is to ensure the effectiveness of scale comparison between images of different sizes. The division is shown in Table 1. The airplane in Figure 6(a) belongs to the large scale, denoted as "x".

Other issues. In addition, for all class of objects, there is a common question: how small a object does not need to be labeled? Our principle is to label objects that can roughly distinguish categories based on silhouettes. If the object is too small to distinguish its class, it is not labeled. To ensure the quality of labeling, we have established data review standards, with three people who were not involved in the annotation work responsible for data verification. The annotated results are visualized and independently reviewed by two reviewers. If both reviewers agree the data annotation is correct, the data will be sent to the last reviewer for review. If one or both of them think the annotation has problems, the data will be discussed and the conclusion will be fed back to the annotator, and the annotation will be reviewed again. It is only qualified if it is approved by the final reviewer.

3.4 Statistical analysis of dataset

We conducted statistical analyses on the variations in scale, angle, and occlusion levels in our dataset to showcase the inherent challenges faced by this dataset.

Scale distribution. In order to calculate the scale distribution of all objects in AFTD, we introduce the area ratio to measure the scale of the objects. The calculation formula of the area ratio is as follows:

$$Arearatio = \frac{(x_{\max} - x_{\min})(y_{\max} - y_{\min})}{wh} \quad (2)$$

The statistical results are shown in Figure 8. It shows that the scales of all types of objects are concentrated in the small scale. The scales of people and cars are concentrated in extremely small scales, while the distribution of airplane scales is broader but still primarily concentrated in smaller scales. This poses challenges for current object detectors in detecting small objects, handling intra-class diversity, and detecting multi-scale objects.

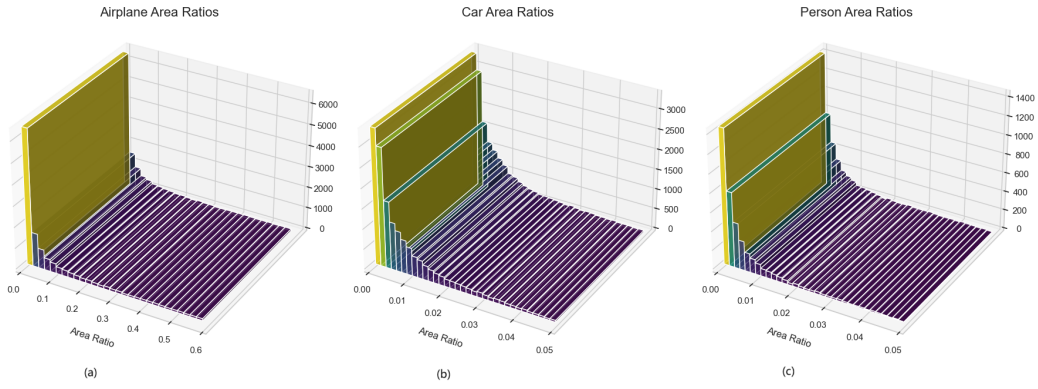


Figure 8: Scale distribution:(a) shows area ratio aistribution for airplane,(b) shows area ratio aistribution for car,(c) shows area ratio aistribution for person.

Viewangle distribution. Figure 9(a) shows the distribution of viewangle of the airplane in the AFTD. According to the distribution, the airplane in the airport are widely distributed in all angles, and the airplane in the right and left direction account for the majority of all airplane. According to the 3D model of the airplane, it is easy to obtain that the wing target exists on the non-left-right orientation of the airplane parent mark. The visual distribution of airplane shows that the detection of the wing is a problem that can not be ignored in the realistic foreground target surveillance scene of the airport,

but the existing object detection algorithms are not good in the performance of slender objects such as the wing. This is also a challenge posed by AFTD to existing object detection algorithms.

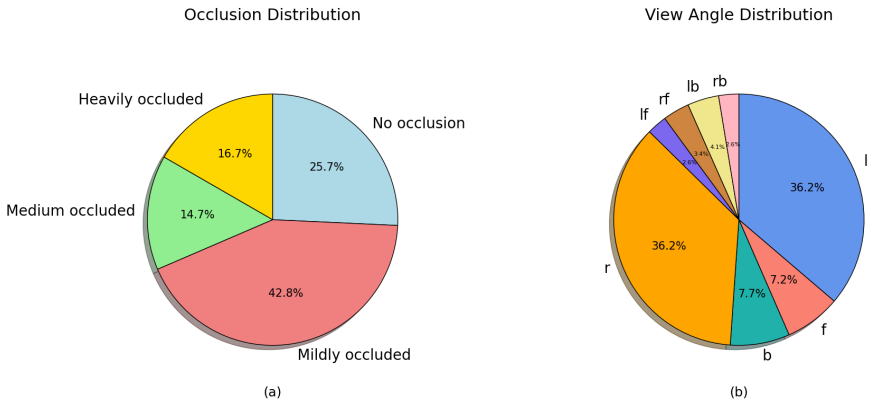


Figure 9: Viewangle distribution and occlusion distribution.

Occlusion distribution. Figure 9(b) shows the distribution of airplane occlusion degree in the AFTD dataset. It shows that the problem of airplane occlusion exists widely in the airport surveillance scene, and a large number of medium occluded and heavily occluded aircraft seriously affect the detection results.

4 EXPERIMENTS

In this section, we select 14 object detection algorithms that perform well on MS-COCO to test our dataset and analyze the results. These algorithms include two-stage, one-stage, and end-to-end object detection algorithms. Moreover, we discuss algorithm design approaches for future research.

4.1 EXPERIMENTAL SETTINGS

For the each algorithm we chosed, we used the source code provided in the original paper. For model configuration, unless specified otherwise, we adopted the default training settings. It should be noted that the three additional labels we designed for the airplane are for the need of subsequent algorithm design, so they are not used in the experiments in this section.

4.2 EVALUATION METRICS

Average Precision (AP) is an average of the precision at different recall points, and the larger the AP value indicates that the model is more effective. mAP is an average of the AP values of all the categories, AP can reflect the accuracy of the prediction of each category, and mAP is used to reflect the accuracy of the whole model. Since the AFTD dataset contains multiple categories, we use mAP as the evaluation metric. Here we calculate AP50 and AP75 for each algorithm on the AFTD dataset separately. AP50 denotes the value of mAP for the IOU threshold of 0.5, and AP75 is the same.

4.3 EXPERIMENTAL RESULT AND ANALYSIS

The experimental results of the 14 algorithms on the AFTD dataset are shown in Table 2, and the highest scores under each indicator are bolded. As a comparison, the experimental results of these algorithms on the MS-COCO dataset are also shown in Table 2, which are taken from the official website. It can be seen that the boxAP of each algorithm on the AFTD dataset decreased by almost 15 percentage points compared to the boxAP on the MS-COCO dataset. On the one hand, this indicates the fact that the AFTD dataset is more difficult than MS-COCO. On the other hand, it also reflects that general object detection algorithms can still be further optimized for specific domains and tasks. Among the 14 algorithms we tested, the Sparse RCNN and TOOD performed comparably well with

Table 2: The performance comparison of some object detection algorithms on MS-COCO and on the AFTD. Best scores are marked in bold.

Algorithms	Backbone	COCO	AFTD					
		boxAP	boxAP	AP50	AP75	APS	APM	APL
Mask RCNNHe et al. (2017)	ResNet-50	38.2	25.6	47.1	24.7	16.4	37.3	55.4
Cascade RCNNCai & Vasconcelos (2018)	ResNet-50	40.4	26.7	46.2	26.5	16.3	37.5	56.6
CenterNetDuan et al. (2019)	ResNet-50	40.2	25.8	45.1	25.8	20.0	41.3	59.8
Libra R-CNNPang et al. (2019)	ResNet-50	38.3	25.9	48.0	24.9	17.8	39.1	55.0
TridentNetLi et al. (2019)	ResNet-50	37.7	24.8	44.0	24.4	12.6	35.3	56.3
FOCSTian et al. (1904)	ResNet-50	38.5	22.7	42.4	21.6	16.2	37.2	55.3
Double-Head RCNNWu et al. (2020)	ResNet-50	40.0	26.4	46.6	26.0	17.1	38.5	56.8
Dynamic RCNNZhang et al. (2020)	ResNet-50	38.9	25.6	45.5	25.2	15.0	36.6	55.1
DETRCarion et al. (2020)	DETR	39.9	21.9	41.8	19.3	11.7	30.5	56.0
Sparce RCNNSun et al. (2021)	ResNet-50	**42.8**	26.8	46.8	26.2	**21.0**	40.1	59.4
YOLOFChen et al. (2021)	ResNet-50-C5	37.5	22.3	41.4	20.7	10.8	35.1	57.4
YOLOX-SGe et al. (2021)	-	40.5	23.1	41.5	22.3	13.4	34.3	54.4
TOODFeng et al. (2021)	ResNet-50	42.4	**28.1**	**47.7**	**28.1**	20.5	**43.5**	**62.4**
Conditional DETRMeng et al. (2021)	ResNet-50	41.1	24.1	45.7	21.4	14.0	34.7	58.5

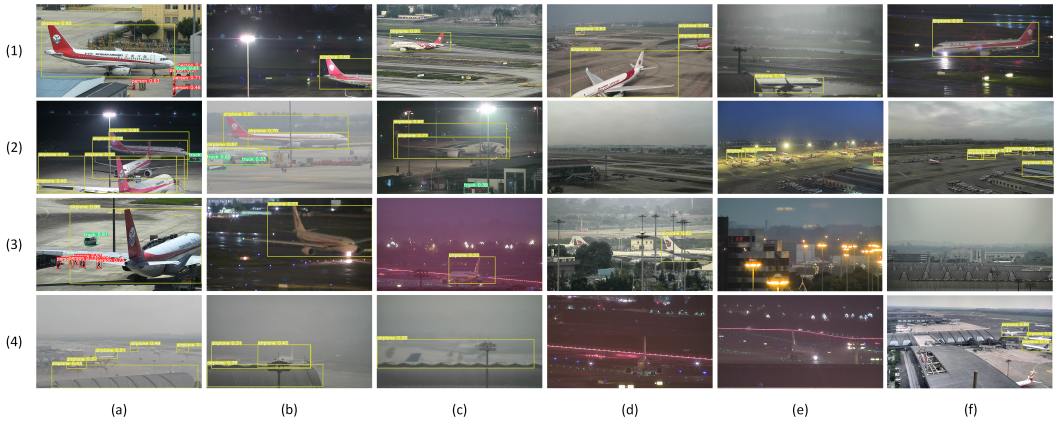


Figure 10: Result: The first row of images shows the exact detection results. The next three rows show some problems with the detection algorithm, with phenomena such as inaccurate boundary localization, missed detection, and false detection.

relatively high boxAP. This may be attributed to the fact that both algorithms show relatively good performance in small object detection.

We next illustrate the reasons for the poor performance of existing methods in terms of challenges. The first row in Figure 10 shows some results with relatively accurate localization.

Super multi-scale. The challenges brought by super multi-scale are mainly manifested in two aspects: objects with too large or small in scale. When the airplane is close to the camera, the imaging scale is large and the algorithm may give more than one detection result for the same airplane, as shown in Figure 10(2a) to (2c). When movable targets are displayed at a small size in the image, it may result in the algorithm being unable to accurately recognize these objects, as shown in Figure 10(2d) to (2f).

Viewangle changes. There are cases of inaccurate boundary localization for movable targets, as shown in Figure 10(3a) to (3c).This situation mainly occurs when the airplane wings form the left or right boundaries of the airplane, mainly because the slender characteristics of the wings are not easy to accurately locate the boundaries.

Multi-level occlusion. Occlusion can lead to the loss of key features for movable targets, resulting in the algorithm missing these objects, as shown in Figure 10(3e) to (3f).

Illumination changes and weather changes. In conditions of low visibility and poor Illumination, such as foggy, rainy days and nights, it may be difficult to distinguish movable targets from the background, which in turn causes the algorithm to miss them, as shown in Figure 10(4a) to (4e).

Complex background. In complex backgrounds, movable targets are easily masked by similar texture features of the surrounding environment, leading to a reduced differentiation between object and background, as shown in Figure 10(4f).

4.4 Additional Discussion

The above experiments show that object detection does face challenges in airport surveillance. Here, we take one challenge as examples and propose some possible algorithm design ideas.

Super multi-scale problem. Since objects at different scales differ in texture details, relationship with the context, etc., detection at different scales may not be fully compatible or have different optimal solutions. Therefore, a single detection algorithm might not be effective for all scales at the same time. We could use different attention mechanisms for objects at different scales. This method first requires predicting the approximate locations where different scale objects will appear. Observations have shown that there is a special "target/background" co-occurrence in airport scenarios, so that the scale of the object can be initially predicted by dividing the background region, for example, airplane appearing in the sky region generally belong to the small-scale objects.

Based on the above ideas, we designed a module that uses different attention mechanisms for different scale features. Figure 11 shows the structure of the module.In order to prove the effectiveness of our module, we conducted experiments on yolov8 with small target detection head, and the experimental results are shown in Table 3.

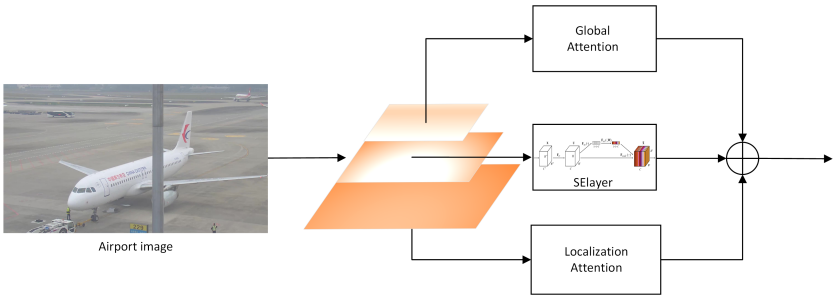


Figure 11: The structure of our multiscale attention module.

For small-scale features, we use localization attention to enhance their attention to local information, for middle-scale features, we use SElayerHu et al. (2017) to provide attention to channel information, and for large-scale features, we use global attention to enhance their attention to global information. Finally, these features are weighted and fused for detection.

Table 3: Comparison of performance metrics with and without multi-scale attention.

Class	Instances	RT-DETR without attention			RT-DETR with attention		
		BoxAP	mAP50	mAP50-95	BoxAP	mAP50	mAP50-95
airplanes	1766	0.617	0.607	0.383	0.609(-0.08)	0.625(+0.018)	0.389(+0.006)
airplanem	804	0.452	0.485	0.398	0.485(+0.032)	0.514(+0.029)	0.411(+0.013
airplanex	683	0.666	0.798	0.71	0.724(+0.058)	0.819(+0.021)	0.73(+0.02)

The experimental results show that using different attention mechanisms for different scales can effectively improve the performance of the object detection algorithm on the AFTD, which indicates that our previous analysis on the scale of our dataset is correct.

5 Conclusion

In this paper, we proposed an airport foreground target detection dataset, AFTD, which contains a total of 10050 image data. The AFTD dataset covers a variety of changes of foreground targets on the airport surface. Our experiments showed that some algorithms that performed well on classical

datasets exhibited varying degrees of performance degradation on the AFTD dataset. We analyzed the reasons for this phenomenon and proposed some algorithmic design ideas for future research. Further expanding the AFTD dataset and improving the detection performance of airport foreground targets will be the focus of our future research.

REFERENCES

Z.W. Cai and N. Vasconcelos. "cascade r-cnn: Delving into high quality object detection,". In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 6154–6162, 2018.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. "end-to-end object detection with transformers,". In *Eur. Conf. Comput. Vis.*, pp. 213–229. Springer, 2020.

Q. Chen, Y.M. Wang, T. Yang, X.Y. Zhang, J. Cheng, and J. Sun. "you only look one-level feature,". In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 13039–13048, 2021.

J. Deng, W. Dong, R. Socher, L.J. Li, L. Kai, and F.F. Li. "imagenet: A large-scale hierarchical image database,". In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

K.W. Duan, S. Bai, L.X. Xie, H.G. Qi, Q.M. Huang, and Q. Tian. "centernet: Keypoint triplets for object detection,". In *IEEE Int. Conf. Comput. Vis.*, pp. 6569–6578, 2019.

M. Everingham and J. Winn. "the pascal visual object classes challenge 2012 (voc2012) development kit,". *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007(1-45):5, 2012.

M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, and A. Zisserman. "the PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascalnetwork.org/challenges/VOC/voc2007/workshop/index.html.

C.J. Feng, Y.J. Zhong, Y. Gao, M.R. Scott, and W.L. Huang. "tood: Task-aligned one-stage object detection". In *IEEE Int. Conf. Comput. Vis.*, pp. 3490–3499. IEEE Computer Society, 2021.

Z. Ge, S.T. Liu, F. Wang, Z.M. Li, and J. Sun. "yolox: Exceeding yolo series in 2021,". *arXiv preprint arXiv:2107.08430*, 2021.

R. Girshick. "fast r-cnn,". In *IEEE Int. Conf. Comput. Vis.*, pp. 1440–1448, 2015.

R. Girshick, J. Donahue, T. Darrell, and J. Malik. "rich feature hierarchies for accurate object detection and semantic segmentation,". In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 580–587, 2014.

K.M. He, X.Y. Zhang, S.Q. Ren, and J. Sun. "spatial pyramid pooling in deep convolutional networks for visual recognition,". *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9):1904–1916, 2015.

K.M. He, G. Gkioxari, P. Dollár, and R. Girshick. "mask r-cnn,". In *IEEE Int. Conf. Comput. Vis.*, pp. 2961–2969, 2017.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. *arXiv e-prints*, art. arXiv:1709.01507, September 2017. doi: 10.48550/arXiv.1709.01507.

A. Krizhevsky, I. Sutskever, and G.E. Hinton. "imagenet classification with deep convolutional neural networks,". *Adv. neural Inf. Process. Syst.*, 25, 2012.

H. Law and J. Deng. "cornernet: Detecting objects as paired keypoints,". In *Eur. Conf. Comput. Vis.*, pp. 734–750, 2018.

Y.H. Li, Y.T. Chen, N.Y. Wang, and Z.X. Zhang. "scale-aware trident networks for object detection". In *IEEE Int. Conf. Comput. Vis.*, pp. 6054–6063, 2019.

T.Y. Lin, Mi. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick. "microsoft coco: Common objects in context,". In *Proc. Eur. Conf. Comput. Vis.*, pp. 740–755. Springer, 2014.

T.Y. Lin, P. Goyal, R. Girshick, K.M. He, and P. Dollár. "focal loss for dense object detection,". In *IEEE Int. Conf. Comput. Vis.*, pp. 2980–2988, 2017.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg. "ssd: Single shot multibox detector,". In *Eur. Conf. Comput. Vis.*, pp. 21–37. Springer, 2016.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. "fine-grained visual classification of aircraft,". Technical report, 2013.

D.P. Meng, X.K. Chen, Z.J. Fan, G. Zeng, H.Q. Li, Y.H. Yuan, L. Sun, and J.D. Wang. "conditional detr for fast training convergence". In *IEEE Int. Conf. Comput. Vis.*, pp. 3651–3660, 2021.

J.M. Pang, K. Chen, J.P. Shi, H.J. Feng, W.L. Ouyang, and D.H. Lin. "libra r-cnn: Towards balanced learning for object detection,". In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 821–830, 2019.

J. Redmon and A. Farhadi. "yolo9000: better, faster, stronger,". In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 7263–7271, 2017.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "you only look once: Unified, real-time object detection,". In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 779–788, 2016.

S.Q. Ren, K.M. He, R. Girshick, and J. Sun. "faster r-cnn: Towards real-time object detection with region proposal networks,". *Adv. Neural Inf. Process. Syst.*, 28, 2015.

S. Shao, Z.M. Li, T.Y. Zhang, C. Peng, G. Yu, X.Y. Zhang, J. Li, and J. Sun. "objects365: A large-scale, high-quality dataset for object detection,". In *IEEE Int. Conf. Comput. Vis.*, pp. 8430–8439, 2019.

P. Sun, R.F. Zhang, Y. Jiang, T. Kong, C.F. Xu, W. Zhan, M. Tomizuka, L. Li, Z.H. Yuan, C.H. Wang, et al. "sparse r-cnn: End-to-end object detection with learnable proposals,". In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 14454–14463, 2021.

Z. Tian, C. Shen, H. Chen, and T. He. "fcos: Fully convolutional one-stage object detection. arxiv 2019,". *arXiv preprint arXiv:1904.01355*, 1904.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N Gomez, Ł. Kaiser, and I. Polosukhin. "attention is all you need,". *Adv. Neural Inf. process. Syst.*, 30, 2017.

Y. Wu, Y.P. Chen, L. Yuan, Z.C. Liu, L.J. Wang, H.Z. Li, and Y. Fu. "rethinking classification and localization for object detection,". In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 10186–10195, 2020.

H.K. Zhang, H. Chang, B.P. Ma, N.Y. Wang, and X.L. Chen. "dynamic r-cnn: Towards high quality object detection via dynamic training,". In *Eur. Conf. Comput. Vis.*, pp. 260–275. Springer, 2020.

X.Z. Zhu, W.J. Su, L.W. Lu, B. Li, X.G. Wang, and J.F. Dai. "deformable detr: Deformable transformers for end-to-end object detection,". *arXiv preprint arXiv:2010.04159*, 2020.