A NEW DATASET AND RESEARCH APPROACH FOR MOVABLE TARGET DETECTION IN AIRPORTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Movable target detection on airport surfaces is fundamental to intelligent airport surveillance, yet the field lacks efficient detection algorithms and dedicated benchmark datasets. To address this gap, this paper introduces AFTD, a foreground target detection dataset for three types of movable airport targets: airplanes, vehicles, and persons. Through autonomous data acquisition and web data collection, over 200000 images were obtained, from which 10050 images were selected based on diversity to form the AFTD dataset. It contains 26968 airplane instances, 24759 vehicle instances, and 5064 person instances. AFTD covers diverse real-world challenges and provides fine-grained annotations, including three-level scale and occlusion annotations, as well as eight-level viewpoint annotations. Based on AFTD, we conducted extensive comparative experiments and confirmed that foundational detection algorithms are not robust to real-world scenarios. Subsequently, we propose a scene-specific prior-based approach for airport detection, along with its preliminary implementation. Furthermore, we demonstrate the cross-scenario generalizability of this research approach. AFTD can be downloaded from http://www.agvs-caac.com/aftd/aftd.html, and NSAD from https://github.com/AFTD-NSAD/AFTD-NSAD.

1 INTRODUCTION

In recent years, the rapid growth of the global population and economy has led to a significant increase in passenger and cargo traffic at airports. Meanwhile, there has been a rising trend in security incidents involving movable targets on airport surfaces. For example, on March 22, 2026, a collision between an airplane and a fire truck occurred at LaGuardia Airport in New York, USA, resulting in two deaths and dozens of injuries. Therefore, it is particularly important to develop intelligent applications to enhance the surveillance of movable targets on airport surfaces.

Target detection is widely used in airport surveillance because it serves as the fundamental building block for these applications. For example, in route planning, airplanes must be detected in real time to ensure they follow the planned path at the airport. However, we find that the performance of target detection algorithms degrades to varying degrees in airport scenarios, and subsequent experiments support this observation. This stems from the fact that foundational research focuses on generalization. For example, foundational research datasets typically contain a rich variety of object categories but do not emphasize covering all states of a single category. In practical applications such as airport surveillance, however, the primary need is not generalization, but robustness to various alterations within specific scenes. In other words, airport surveillance requires a dataset that reflects the various changes of airport targets. Clearly, classical datasets do not meet this requirement.

In this paper, we propose the first Airport Foreground Target Detection dataset, AFTD. Foreground targets are defined as movable objects, which in airports are airplane, vehicle, and person. AFTD focuses on foreground targets, because they are the main body of activity on the airport surface and the root cause of various airport accidents. We obtain permission to collect data at multiple airports through cooperation with a civil aviation research institute of China, and fully utilize network resources, ultimately collecting over 200000 raw images with various target changes such as super multiscale, object occlusion, as well as scene changes such as weather and illumination variations. We then rigorously filtered the data based on the diversity principle. Due to the high similarity in airport structure and the shape of movable targets, only 10050 images are retained, totaling 26968

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Example samples in AFTD dataset.

airplane instances, 24759 vehicle instances and 5064 person instances. In addition, we present a refined label format for annotation. For example, for the airplane category, we design three scale formats, three occlusion formats and eight viewangle formats. Fig. 1 displays some example images from the AFTD dataset. Through robustness experiments on AFTD, we confirmed that baseline algorithms are unsuitable for real-world scenarios. We then propose a scene-specific prior based approach that leverages various priors specific to a given scene to improve detection performance. Finally, we present NSAD, a preliminary implementation of this research approach, and demonstrate its cross-scenario generalizability.

Our main contributions to the field of target detection are:

- We present AFTD benchmark featuring diverse challenges and fine-grained annotations. It supports both airport detection research and robustness evaluation of foundational algorithms in real-world applications.
- We propose a scene-specific prior based approach for airport target detection, exemplified by the NSAD algorithm. As each scene possesses unique priors, this paradigm generalizes to other scenarios as well.

2 Related Work

2.1 Object Detection Datasets

Over the past decade, many object detection datasets have been released. Some contain multiple categories, such as VOC2007 Everingham et al. (2007) (5k images, 12k instances), VOC2012 Everingham & Winn (2012) (11k images, 27k instances), ImageNet Deng et al. (2009), MS-COCO Lin et al. (2014), and Objects365 Shao et al. (2019). Others focus on airplanes, e.g., FGVC Maji et al. (2013) and DOTA Xia et al. (2018) (see Tab. 1). VOC2007 and VOC2012 cover 20 common daily categories. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Russakovsky et al. (2015) provides 200 visual categories. MS-COCO has fewer categories but more instances and small/dense objects, becoming the de facto standard. OID Krasin et al. (2017) contains 1.7M images across 600 classes, though annotations are sparser per image. While these datasets contain airplane samples, they prioritize category coverage over scene-specific research, making them unsuitable for airport foreground surveillance.

Dedicated datasets exist for faces, such as WiderFace Yang et al. (2016); for pedestrians, such as Stanford Crowd Zhang et al. (2016); and for UAV-based road surveillance, such as UAVDT Du et al. (2018) and AU-AIR Bozcan & Kayacan (2020). Although they share categories like person and vehicle with our AFTD, object morphology differs due to distinct scene contexts. Airplane-specific datasets include FGVC Maji et al. (2013) (10,200 images, 102 Boeing types, mostly side views, limited viewpoint/scale, only for classification) and ASSD Zhou et al. (2024) (only 2,000 images, high redundancy). Remote sensing datasets like DOTA Xia et al. (2018) capture birds-eye views where airplanes are very small and other objects barely recognizable. Consequently, none of these are suitable for foreground target detection on airport surfaces.

2.2 Object Detection Algorithms

Modern object detection architectures are broadly categorized into one-stage detectors, two-stage detectors, and end-to-end (E2E) transformer-based frameworks.

Table 1: Comparison of AFTD with other target detection datasets. In this table, multi-level scale/occlusion/viewangle indicate that these attributes received finer-grained annotations.

Datasets	Year	Specific	Target Categories			Multi-level	Multi-level	Multi-level	Airplane	Airplane
		Scenario	Airplane	Person	Vehicle	Scale	Occlusion	Viewangle	Image no.	Instance no.
VOC2012 Everingham & Winn (2012)	2012	×	✓	✓	✓	×	×	✓	716	1002
Object Discovery Rubinstein et al. (2013)	2013	×	✓	✓	✓	×	×	×	72	88
ImageNet Deng et al. (2009)	2017	×	✓	✓	✓	×	×	×	603	12115
MS-COCO-2017 Lin et al. (2014)	2017	×	✓	✓	✓	✓	×	×	3083	5278
Objects365-2019 Shao et al. (2019)	2019	×	✓	✓	✓	×	×	×	6246	11067
OID-2020 Krasin et al. (2017)	2020	×	✓	✓	✓	×	×	×	–	–
UAVDT Du et al. (2018)	2018	✓	×	×	✓	×	✓	✓	–	–
Stanford Crowd Zhang et al. (2016)	2016	×	×	✓	×	×	×	×	–	–
WiderFace Yang et al. (2016)	2016	×	×	✓	×	×	✓	×	–	–
AU-AIR Bozcan & Kayacan (2020)	2020	✓	×	✓	✓	×	×	×	–	–
AC dataset Steininger et al. (2021)	2021	✓	✓	×	×	×	×	✓	1909	1935
MAD dataset Kaggle Community (2023)	2023	✓	✓	×	×	×	×	×	3520	6513
AFTD (ours)	2025	✓	✓	✓	✓	✓	✓	✓	10050	26968

YOLO-based single-stage detectors dominate real-time applications. YOLOX Ge et al. (2021) revives anchor-free design with data augmentation and a decoupled head. YOLOv12 Tian et al. (2025) integrates attention mechanisms into the backbone, achieving state-of-the-art accuracy while preserving speed. Anchor-free detectors have also gained prominence. FCOS Tian et al. (2020) uses per-pixel regression; CenterNet Duan et al. (2019) treats objects as keypoints; TOOD Feng et al. (2021) introduces task alignment; and RTMDet-M Lyu et al. (2022) achieves strong results with carefully designed training pipelines. The two-stage R-CNN family remains a cornerstone. Mask R-CNN He et al. (2017) unifies detection and segmentation. Key improvements include Cascade R-CNN Cai & Vasconcelos (2018) for box refinement, Libra R-CNN Pang et al. (2019) for imbalance handling, TridentNet Li et al. (2019) for scale-aware detection, Double-Head R-CNN Wu et al. (2020) for decoupling classification and regression, and Dynamic R-CNN Zhang et al. (2020) for adaptive training. FPN Lin et al. (2017) provides multi-scale feature representations for two-stage detectors.

DETR Carion et al. (2020) introduces an end-to-end detection framework, eliminating anchors and NMS Neubeck & Van Gool (2006). Its slow convergence spurred follow-ups: Conditional DETR Meng et al. (2021) accelerates training by conditioning queries on reference points; Sparse R-CNN Sun et al. (2021) adopts learnable proposals; DINO Zhang et al. (2023) integrates denoising training, contrastive learning, and improved Hungarian matching, becoming a strong baseline.Later variants focus on localization precision and efficiency. RT-DETR Zhao et al. (2024) optimizes for real-time inference. D-FINE Peng et al. (2024) redefines regression by modeling box edges as fine-grained distributions, iteratively refined across decoder layers—enabling precise localization and efficient knowledge transfer, yielding substantial gains over DINO. Recently, RF-DETR Robinson et al. (2025) advances detection via a representation-first philosophy: leveraging a strong single-scale feature from a pre-trained vision transformer paired with a lightweight deformable transformer decoder.

3 AFTD Benchmark

3.1 Dataset Design and Challenges

AFTD design focuses on two aspects: selecting target categories and maximizing variation coverage. Among the many categories in airport scenes, AFTD targets three movable objects—airplanes, vehicles, and persons—which are the core of airport activities and the primary contributors to airport incidents. As an application-oriented dataset, AFTD must maximize coverage of both object appearance variations and scene variations. Take the airplane—the most important movable target—as an example. Observations show that its appearance varies primarily along three dimensions: scale, occlusion, and viewangle. At airports, imaging distances range from tens to thousands of meters, causing substantial scale variation. Airplanes are large, non-compact rigid objects whose appearance changes significantly with occlusion and viewangle. Therefore, AFTD samples must cover as many variations in scale, occlusion, and viewangle as possible. In addition, scene variations—such as weather, illumination, and background patterns—must also be addressed. Based on these considerations, we categorize AFTD samples into six subsets: super multiscale, occlusion change, viewangle change, illumination change, weather change, and complex background. However, due to the non-cooperative nature of airport scenes, we cannot isolate single variation types; we can only prioritize one factor per sample. In reality, all samples exhibit multiple variations simultaneously. The typical challenges in AFTD are described as follows.



Figure 2: Illustration of super multiscale phenomenon.

Super multiscale. Multiscale is a common challenge in object detection, but the scale variation in airports is extreme compared to other scenes. As shown in Fig. 2, movable targets can range from a few pixels (Fig. 2a) to a bounding box occupying nearly half the image (Fig. 2b). The super multiscale phenomenon stems from two factors. First, airports cover vast areas with surveillance distances ranging from tens of meters to several kilometers. This large variation in imaging distance causes objects—including those of the same type—to appear at vastly different scales, often coexisting from very small to very large (Fig. 2c). Second, different object types vary significantly in size. For instance, vehicles and persons are much smaller than airplanes, sometimes appearing as just a few blurred pixels. The combination of these two factors results in super multiscale phenomena. To quantify the super multiscale phenomenon in AFTD, we define the area ratio metric A_{ratio} for the scale distribution of the three movable target types as $A_{ratio} = (w \times h)/(W \times H)$, where w and h are the object's width and height, and W and H those of the image. For each object in AFTD, we compute A_{ratio}; Fig. 3 presents the per-type statistics, with the horizontal axis showing A_{ratio} and the vertical axis its frequency. The results indicate that persons and vehicles are concentrated at very small scales, whereas airplanes have a broader but still predominantly small-scale distribution. Thus, super multiscale includes traditional multiscale, multiscale coexistence, and extremely small scales, becoming a core challenge for airport foreground detection.

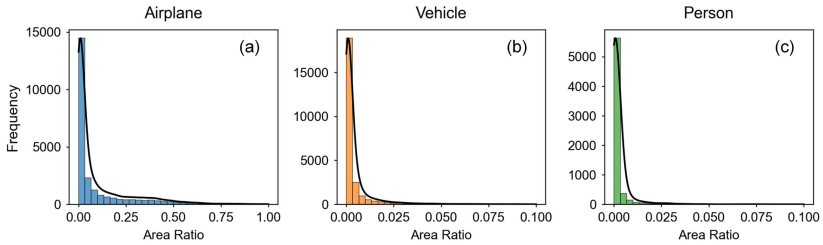


Figure 3: The statistic results of A_{ratio} for each kind of targets.

Occlusion change. Occlusion is particularly prominent in AFTD due to the unique shape of airplanes (Fig. 4). Airplanes have non-compact, jagged wingspans that cause drastic appearance changes under occlusion. As shown in Fig. 4, occlusion varies in severity; in extreme cases, the overall shape cannot be discerned. We categorize airplane occlusion into four levels—none, mild, medium, and heavy—and present their frequencies in AFTD (Fig. 4d). Over three-quarters of airplanes experience occlusion, each level with high probability. Occlusion-induced appearance change is the main cause of low localization accuracy. Definitions of the four levels are given later.



Figure 4: Illustration of object occlusion.

Viewangle change. For wingspan targets like airplanes, apart from occlusion, different viewangles can also lead to drastic changes in appearance. For instance, the appearance of an airplane viewed from the front (Fig. 5g) is completely different from that viewed from the side (Fig. 5a). We quantify the viewangles into eight cases, and airplanes from different cases are shown in Fig. 5. The viewangle quantification will be explained in detail later.

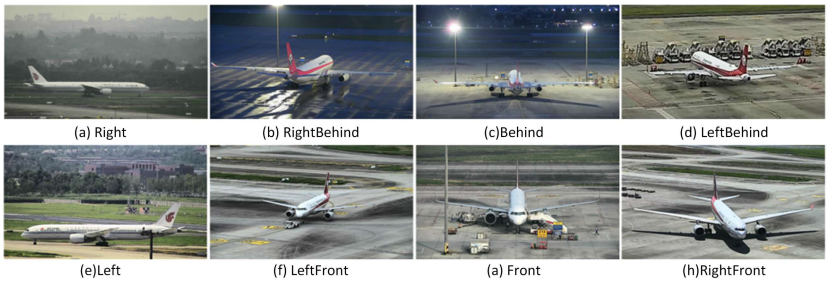


Figure 5: Illustration of airplane from different viewangles.

Additional challenges in the AFTD dataset, including class imbalance, illumination and weather changes, are described in detail in Section A.1 of the supplementary material.

3.2 Data Acquisition

Raw data collection and screening. Through cooperation with an institute of the Civil Aviation Administration of China, we collected data at five airports using fixed and PTZ cameras (resolutions: 1920×1080, 1280×720, etc.). The core principle was maximizing coverage of variations in movable targets and scenes. For airplanes, we adjusted PTZ camera settings to capture diverse distances, view angles, and activities. The collection cycle was extended to cover weather and illumination variations. Over twenty months, we collected approximately 100000 raw images. Data screening aimed to reduce redundancy while maintaining diversity. However, this task is challenging since airport layouts and aircraft shapes are highly similar, leading to significant redundancy. Thus, only 5050 valid images were retained from over 100000 raw images.

Data expansion. Since self-capturing data is challenging, we further collected public airport images from the web to enrich AFTD. Non-natural images, manually edited images, and remote sensing satellite images (overhead views) were excluded. In total, we gathered approximately 100000 online images. After screening with the same principles described in the previous section, 5000 images were retained. Thus, AFTD currently contains 10050 images. In the future, we plan to expand AFTD using a simulated airport, where aircraft will become cooperative rather than uncontrollable targets. For instance, a 4E-level simulated airport—comparable in scale to a real airport—is currently under construction in Chengdu, China, exclusively for scientific research. Once completed, we will seek authorization to use the facility and conduct cooperative data collection.

3.3 Data Annotation

The ground-truth format of AFTD is $\langle class, x, y, w, h \rangle$. The values of $\langle class \rangle$ are 0, 1, and 2, representing airplane, vehicle, and person, respectively. The pair $\langle x, y \rangle$ denotes the bounding box center, and $\langle w, h \rangle$ its width and height. During annotation, $\langle class \rangle$ is determined by visual inspection, while $\langle x, y \rangle$ and $\langle w, h \rangle$ are computed by the software from the manually outlined bounding box. Since airplanes are the most important movable targets, three additional attribute tags $\langle sl, od, av \rangle$ are added for finer-grained annotation, capturing scale level, occlusion degree, and approximate viewangle, respectively. The annotation process for $\langle av \rangle$ and $\langle od \rangle$ is illustrated in Fig. 6.

Approximate viewangle. Based on the orientation angle between the airplane's principal axis and the image's X-axis, we approximate the viewangle into eight cases: Right, Left, Behind, Front, RightBehind, LeftBehind, LeftFront, and RightFront, as shown in Fig. 6a. The computation of the orientation angle is illustrated in Fig. 6b. Taking the tail as the origin, the principal axis connects the tail and nose. The orientation angle is measured counterclockwise from the positive X-axis to this axis. The approximate view angle is then obtained by comparing the orientation angle with Fig. 6a. If the tail or nose is occluded but the fuselage is visible, the principal axis can still be determined from the fuselage direction. If the fuselage is also invisible, $\langle av \rangle$ is not annotated.

Occlusion degree. We quantify occlusion into four levels: none, mild, medium, and heavy. The division is based on the mask ratio M_{ratio}, defined between the occluded target and the complete target. However, outlining the complete target mask is time-consuming, which is more challenging when the target is occluded. To address these issues, we design an approximate computation method based on 3D models. First, manually mark four feature points on the airplane: the endpoints of the two main

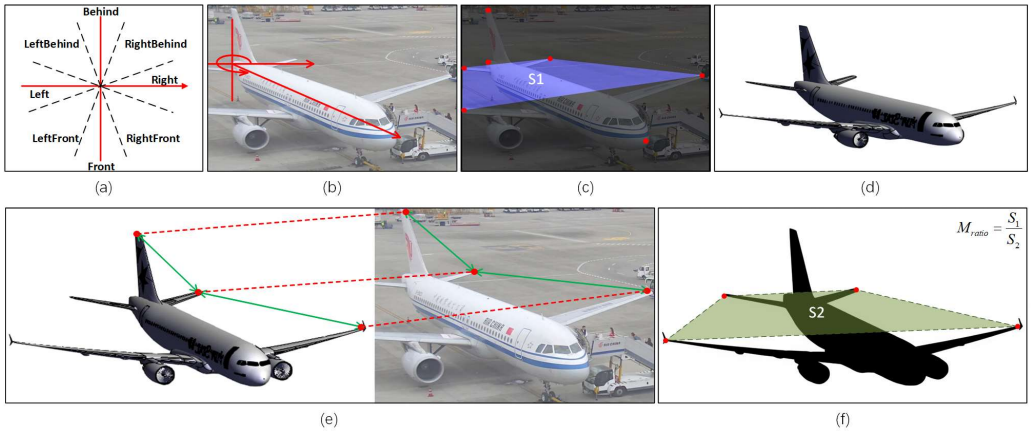


Figure 6: As in (a) and (b), $\langle av \rangle$ is annotated based on the orientation angle. As from (c) to (f), $\langle od \rangle$ is annotated by comparing the approximate target mask and the approximate complete-target mask.

wings and the two horizontal tail endpoints. If a wing is occluded, mark the point closest to its end. These four points are then connected to form a quadrilateral, which approximates the target mask (Fig. 6c). Next, we compute an approximate complete-target mask using a readily available online 3D airplane model. First, manually adjust the 3D model to a pose close to the target (Fig. 6d). Then, using a feature point (e.g., the vertical tail endpoint) as an anchor, carefully adjust the model's scale and pose to align at least two additional feature points (Fig. 6e). The 3D model thus approximates the complete target. Finally, the 2D projection of the quadrilateral formed by connecting the main wing and horizontal tail endpoints of the 3D model serves as the approximate complete-target mask (Fig. 6f). The ratio between the masks in Fig. 6c and Fig. 6f is the mask ratio M_{ratio}.

Scale level. We quantify airplane scale into three levels: small, medium, and large. The division is based on the area ratio A_{ratio}. Note that w and h in A_{ratio} refer to the width and height of the target's bounding box. However, when the airplane is occluded—especially heavily occluded—the bounding box does not accurately reflect the true scale. Fortunately, while annotating $\langle od \rangle$ in the previous section, we already obtained an approximate complete-target mask for the occluded target. Therefore, we compute A_{ratio} based on this mask instead.

Detailed definitions of each scale level, occlusion degree, and approximate viewangle, as well as the complete data annotation process, can be found in Section A.2 of the supplementary material.

4 Experiments and Discussions

This section has two tasks. First, evaluate the robustness of foundational research algorithms in specific scenarios based on the AFTD dataset. Second, establish a research approach for airport detection, and then verify the effectiveness of this approach as well as its cross-scenario generalizability. Regarding the experimental setting, we select COCO val2017, a widely used foundational research benchmark, for comparative evaluation with AFTD. A total of 22 algorithms are tested, including the YOLO family, the DETR family, the two-stage R-CNN family, and other notable algorithms. Multiple metrics are adopted for comprehensive evaluation, e.g., *mAP*, mAP_{50}, $mAP_{50:95}$, mAP_S, mAP_M, and mAP_L. See Section B.1 of the supplementary material for more experimental setting details.

4.1 Robustness Evaluation of Foundational Algorithms

Tab. 2 presents the results of 22 algorithms on AFTD and COCO val2017. It can be seen that, except for mAP_{50}, the performance of foundational research algorithms on AFTD is markedly worse than that on COCO val2017 across all other metrics, sometimes by nearly 20 percentage points. This confirms that such algorithms lack robustness for real-world scenarios, for the reason noted in Section 1: foundational research prioritizes generalizability, whereas real-world applications demand robustness to diverse scene challenges. Motivated by this finding, we establish a dedicated

Table 2: Performance comparison of 22 algorithms on AFTD and COCO val2017.

Model	AFTD					COCO val2017				
	mAP_{50}	$mAP_{50:95}$	mAP_S	mAP_M	mAP_L	mAP_{50}	$mAP_{50:95}$	mAP_S	mAP_M	mAP_L
YOLOX-S Ge et al. (2021)	41.5	22.3	13.4	34.3	54.4	59.4	36.8	38.4	60.2	73.9
YOLOX-M Ge et al. (2021)	64.6	39.5	15.2	38.1	49.4	63.2	42.3	**41.5**	64.1	75.0
YOLOF Chen et al. (2021)	41.4	20.7	10.8	35.1	57.4	51.5	30.3	32.4	55.1	67.3
YOLOv8-M Jocher et al. (2023)	66.5	42.9	14.8	42.5	55.2	63.2	42.9	40.3	64.8	75.2
YOLO12-M Tian et al. (2025)	60.2	37.8	12.5	36.0	50.9	69.6	52.5	35.7	58.2	68.8
DETR Carion et al. (2020)	41.8	19.3	11.7	30.5	56.0	58.9	34.6	36.3	61.1	74.7
Conditional DETR Meng et al. (2021)	45.7	21.4	14.0	34.7	58.5	60.1	36.5	38.0	62.4	75.2
Sparse RCNN Sun et al. (2021)	61.8	39.4	17.1	36.6	48.4	59.2	35.5	37.0	61.3	74.5
DINO Zhang et al. (2023)	65.8	41.8	15.5	41.3	53.4	63.3	42.3	41.2	**65.1**	**76.0**
RT-DETR Zhao et al. (2024)	66.3	42.6	18.1	40.0	52.1	71.2	52.8	34.4	57.4	70.8
D-FINE Peng et al. (2024)	**72.0**	**48.4**	**23.0**	**48.1**	63.1	71.6	**54.0**	36.5	58.0	71.9
RF-DETR Robinson et al. (2025)	71.3	45.6	17.6	43.4	**64.8**	**72.1**	53.2	34.1	58.3	71.3
Mask RCNN He et al. (2017)	47.1	24.7	16.4	37.3	55.4	58.7	35.3	37.5	60.8	74.2
Cascade RCNN Cai & Vasconcelos (2018)	46.2	26.5	16.3	37.5	56.6	60.3	37.8	39.2	62.5	75.1
Libra R-CNN Pang et al. (2019)	48.0	24.9	17.8	39.1	55.0	59.7	36.9	38.7	61.9	74.8
TridentNet Li et al. (2019)	65.3	41.9	15.4	41.5	53.2	60.3	37.6	39.0	62.4	75.0
Double-Head RCNN Wu et al. (2020)	46.6	26.0	17.1	38.5	56.8	59.8	37.1	38.9	62.0	74.9
Dynamic RCNN Zhang et al. (2020)	45.5	25.2	15.0	36.6	55.1	59.5	36.7	38.5	61.7	74.6
RTMDet-M Lyu et al. (2022)	65.5	43.8	14.3	41.8	57.3	63.0	42.5	40.7	64.9	75.6
FCOS Tian et al. (2020)	62.1	34.5	13.5	35.5	44.4	59.8	35.7	37.8	61.5	74.0
CenterNet Duan et al. (2019)	45.1	25.8	20.0	41.3	59.8	52.2	29.3	32.0	55.6	66.1
TOOD Feng et al. (2021)	62.7	37.4	14.7	37.1	47.4	61.3	39.1	39.5	63.2	74.3

airport detection dataset and design algorithms tailored to it. For additional experiments on AFTD, particularly qualitative analyses, see Section B.2 of the supplementary material.

4.2 Research Approach for Airport Detection and Its Generalizability

Our observations reveal that multiple types of prior information exist in airport scenarios, including structure-motion, scale, co-occurrence, shape and multimodal priors, as shown in Fig. 7. This prior information provides knowledge beyond image data and can complement the image data to enhance detection performance. However, this approach has a limitation: because it relies on airport-specific prior information, it is applicable only to airport scenes. Nevertheless, since every scene inherently possesses its own prior knowledge, the approach of target detection based on scene-specific prior is generalizable—a point that is also validated in this section.

Structure-motion prior. Airports follow a standardized layout design. For example, all airports consist of architectural components such as runways, taxiways, connecting ways, and aprons, with prominent markings on the apron surface to guide aircraft during approach. We refer to this as structure prior knowledge, as illustrated in the left panel of Fig. 7a. This structure prior makes the motion patterns of movable objects within the airport predictable. Specifically, an aircraft can enter acceleration mode on the runway (but slow-speed mode elsewhere), adopts straight-line motion on the runway or taxiway (turning mode on the connecting way), and when moving on the apron, its direction can be predicted from ground markings (right panel of Fig. 7a). Research on airport detection based on structure-motion prior consists of two aspects: prior modeling and prior-guided detector design. For example, a Markov state model can be employed to model structure-motion prior: first, the motion patterns (states) of aircraft in different regions are summarized, and then the state transition probabilities between these motion patterns are computed, thereby constructing a standard Markov state model. This model can be used to predict possible aircraft locations. In the detector design stage, the model can be embedded into anchor or query based optimization mechanisms to guide the detector to focus on critical regions, thereby improving detection performance.

Scale prior. If the samples from a given scene are sufficiently abundant, the scale prior of important target categories in the scene can be statistically obtained. For example, by directly computing the scales of aircraft in the AFTD dataset, the prior scale distribution of aircraft can be derived, as shown in Fig. 3a. This scale prior is bound to improve the performance of airport target detection. For instance, during multi-scale feature extraction, one can selectively focus on scales where the probability of target occurrence is high, as indicated by the scale prior, thereby obtaining feature representations that are more conducive to detection. Furthermore, there also exists a contrastive scale prior in the airport scenario. For example, aircraft may appear either in the sky or on the apron, but aircraft in the sky are typically smaller than those on the apron, as illustrated in Fig. 7c.

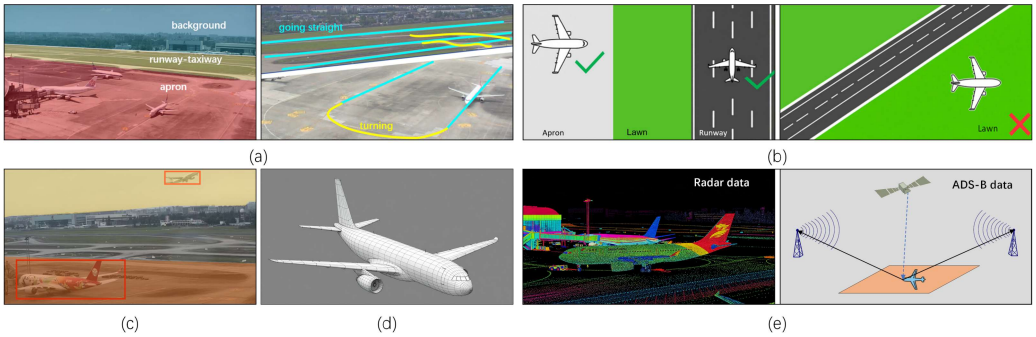


Figure 7: Illustration of five kinds of scene-specific priors in airport scenarios.

Co-occurrence prior. Movable airport targets exhibit either compatible or mutually exclusive co-occurrence relationships with other airport semantics. For instance, aircraft are compatible with aprons and the sky but mutually exclusive with grass (see Fig. 7b), where entry into grass typically indicates a serious accident. Such co-occurrence relationships have a dual impact on detection. On the one hand, compatibility can be detrimental: since compatible semantics often co-appear in training samples, classifiers trained on them struggle to differentiate these classes effectively. On the other hand, co-occurrence priors can be leveraged to enhance detection performance. For example, for compatible semantics, generating synthetic samples with mutually exclusive characteristics can help weaken the original compatibility bias in the training set.

Shape prior. Shape priors of movable airport targets (e.g., the 3D aircraft model in Fig. 7d) are readily available and can be used to enhance detection performance. For instance, the non-compact, wingspan shape of an aircraft often hinders accurate localization of wing boundaries by bounding boxes. In contrast, enhancing wing features through strong, 3D model-based shape constraints facilitates more precise boundary localization.

Multimodal prior. Observations indicate that airports also contain some multimodal data, such as radar data, multilateration data, and ADS-B data, as shown in Fig. 7e. Multimodal data can provide information not found in image and video data. For example, radar data includes 3D depth information, while ADS-B data includes satellite positioning information.

Preliminary implementation. We take the scale prior as an example to demonstrate how to design an airport target detection algorithm. This task consists of two stages: scale prior modeling and scale-prior-guided detector design. To this end, we propose a Nonparametric Scale-based Airport Detection (NSAD) algorithm, which models the scale prior in a nonparametric manner and leverages this prior to constrain the anchor or query optimization mechanism within the detector.

The first module of NSAD is *nonparametric scale modeling*. For each foreground category c, we directly fit its scale distribution $p(s|c)$ based on nonparametric kernel density estimation, where s denotes the geometric scale. The results of nonparametric fitting can be seen in Section B.3 of the supplementary materials. The second module is *scale-guided anchor optimization*, which has two steps, dynamic anchor initialization and scale-normalized regression. The initial prediction is usually generated by fixed or uniformly sampled anchors, which often do not match the real scale distribution, especially when there are rare or extreme scales. In dynamic anchor initialization, for each initial prediction we dynamically sample a scale from the prior distribution $p(s|c)$, and then reconstruct the initial prediction box with the sampled scale. Since $p(s|c)$ reflects the true scale distribution, anchors generated in this way can be well aligned with the real object scales. Standard bounding box regression loss treats all scales equally. When the scales of different objects vary greatly, especially in scenarios like airports where scales change drastically, it may lead to unstable optimization. In scale-normalized regression, we introduce a scale-aware regression loss that penalizes errors according to the normalized scale rather than the absolute scale. Please refer to Section B.3 of the supplementary materials for a comprehensive introduction to NSAD. The source code of NSAD can be downloaded from https://github.com/AFTD-NSAD/AFTD-NSAD.

The comparative experimental results of the NSAD algorithm are shown in Tab. 3 and Fig. 8. For the NSAD algorithm, Tab. 3 presents the statistical results based on random seed analysis, used to further evaluate the robustness of the algorithm under different random initialization conditions. Random seeds were used to control the stochastic processes in training, including data shuffling, stochastic data

Table 3: Comparison of NSAD with other algorithms on AFTD and COCO val2017. Results of NSAD are reported as mean±standard deviation over multiple independent runs with different random seeds.

Method	AFTD					COCO val2017				
	AP_{50}	$AP_{50:95}$	AP_S	AP_M	AP_L	AP_{50}	$AP_{50:95}$	AP_S	AP_M	AP_L
YOLOv12-X Tian et al. (2025)	65.6	41.6	18.1	39.6	51.1	72.0	55.2	35.7	58.2	68.8
RT-DETR Zhao et al. (2024)	66.3	42.6	18.1	40.0	52.1	71.2	52.8	34.4	57.4	70.8
D-FINE Peng et al. (2024)	72.0	48.4	23.0	48.1	**63.1**	71.6	54.0	**36.5**	58.0	71.9
NSAD (ours)	**73.8±0.12**	**49.3±0.08**	**23.9±0.15**	**49.4±0.10**	62.3±0.09	**73.5±0.11**	**55.3±0.07**	36.1±0.16	**60.0±0.09**	**73.0±0.06**

augmentation, random initialization of training components, and other random operations involved in optimization. We conducted five independent experiments using a fixed random seed set {0,1,2,3,4}, with all other training configurations remaining completely consistent. The mean and standard deviation obtained from the five experiments are shown in Tab. 3.

As shown in Tab. 3 and Fig. 8, the proposed NSAD is simple yet effective. When compared with D-FINE Peng et al. (2024), a current state-of-the-art detection method, NSAD delivers improved overall performance and notably lowers the missed detection rate. This experiment clearly confirms that target detection based on scene-specific prior is effective: by leveraging the mutual complementarity between prior information and image information, the performance of target detection in a given scene can be improved. Due to the use of airport prior information, NSAD is only applicable to airport detection. However, by modifying its first module—the nonparametric scale modeling module—NSAD can be extended to other scenes or datasets. We adapted NSAD to COCO val2017 by refitting the scale distribution of all classes in the first module while keeping the second module unchanged. Results on the full dataset are also shown in Tab. 3, and results on the person class are shown in Fig. 8. This experiment clearly demonstrates that the research approach of target detection based on scene-specific prior information is generalizable across different scenes.



Figure 8: Top to bottom: original images, results of D-FINE Peng et al. (2024) and NSAD. left two columns from AFTD, right two from COCO val2017. Red boxes indicate missed detections.

5 Conclusion

In this paper, we first presented AFTD, an airport movable target detection dataset that contained tens of thousands of instances of airplane, vehicle, and person, covering various challenges encountered in airport detection. AFTD was designed to serve both as a benchmark for airport detection research and as a testbed for evaluating the robustness of foundational algorithms in real-world scenarios. Second, we proposed a scene-specific prior knowledge based approach for airport target detection, along with its initial implementation, NSAD. We also demonstrated that this research approach was generalizable across different scenes. Our future work includes leveraging other priors beyond scale and fine-grained annotations in AFTD to further improve airport detection, and replicating the research approach proposed in this paper in more scenarios.

REFERENCES

I. Bozcan and E. Kayacan. AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 8504–8510, 2020. doi: 10.1109/ICRA40945.2020.9196713.

Z. Cai and N. Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 6154–6162, 2018. doi: 10.1109/CVPR.2018.00644.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *Proc. Eur. Conf. Comput. Vis.*, pp. 213–229, 2020. doi: 10.1007/978-3-030-58452-8_13.

Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun. You only look one-level feature. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 13039–13048, 2021. doi: 10.1109/CVPR46437.2021.01285.

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

D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. UAVDT: A benchmark dataset for UAV object detection. In *Proc. ACM Int. Conf. Multimedia*, pp. 261–269, 2018. doi: 10.1145/3240508.3240631.

K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. Centernet: Keypoint triplets for object detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 6569–6578, 2019. doi: 10.1109/ICCV.2019.00667.

M. Everingham and J. Winn. The PASCAL visual object classes challenge 2012 voc2012 development kit. Tech. Rep. 1-45, Pattern Anal. Stat. Model. Comput. Learn., 2012.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes challenge 2007 (VOC2007) results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang. TOOD: Task-aligned one-stage object detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 3490–3499, 2021. doi: 10.1109/ICCV48922.2021.00348.

Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. YOLOX: Exceeding YOLO series in 2021. *arXiv*, 2021.

K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2961–2969, 2017. doi: 10.1109/ICCV.2017.322.

G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics YOLO (version 8.0.0). https://github.com/ultralytics/ultralytics, 2023.

Kaggle Community. Military aircraft detection dataset. https://www.kaggle.com/datasets/a2015003713/militaryaircraftdetectiondataset, 2023.

I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. https://github.com/openimages, 2017.

Y. Li, Y. Chen, N. Wang, and Z. Zhang. Scale-aware trident networks for object detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 6054–6063, 2019. doi: 10.1109/ICCV.2019.00615.

T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proc. Eur. Conf. Comput. Vis.*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2117–2125, 2017. doi: 10.1109/CVPR.2017.106.

C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen. RTMDet: An empirical study of designing real-time object detectors. *arXiv*, 2022.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv*, 2013.

M. Meng, X. Chen, H. Fan, Z. Li, S. Sun, P. Luo, and J. Wang. Conditional DETR for fast training convergence. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 3651–3660, 2021. doi: 10.1109/ICCV48922.2021.00363.

A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *Proc. Int. Conf. Pattern Recognit.*, volume 3, pp. 850–855, 2006. doi: 10.1109/ICPR.2006.479.

J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin. Libra R-CNN: Towards balanced learning for object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 821–830, 2019. doi: 10.1109/CVPR.2019.00091.

Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu. D-FINE: Redefine regression task in DETRs as fine-grained distribution refinement. *arXiv*, 2024.

I. Robinson, P. Robicheaux, M. Popov, D. Ramanan, and N. Peri. RF-DETR. https://github.com/roboflow/rf-detr, 2025.

M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1939–1946, 2013. doi: 10.1109/CVPR.2013.251.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 8430–8439, 2019. doi: 10.1109/ICCV.2019.00852.

D. Steininger, V. Widhalm, J. Simon, A. Kriegler, and C. Sulzbachner. The aircraft context dataset: Understanding and optimizing data variability in aerial domains. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 3823–3832, 2021. doi: 10.1109/ICCV48922.2021.00381.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, Ping Luo, and Jifeng Dai. Sparse R-CNN: End-to-end object detection with learnable proposals. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 14454–14463, 2021. doi: 10.1109/CVPR46437.2021.01423.

Y. Tian, Q. Ye, and D. Doermann. YOLOv12: Attention-centric real-time object detectors. *arXiv*, 2025.

Z. Tian, C. Shen, H. Chen, and T. He. FCOS: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(4):1922–1933, 2020. doi: 10.1109/TPAMI.2020.3032166.

Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, and Y. Fu. Rethinking classification and localization for object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 10186–10195, 2020. doi: 10.1109/CVPR42600.2020.01020.

G.S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3974–3983, 2018. doi: 10.1109/CVPR.2018.00418.

S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 5525–5533, 2016. doi: 10.1109/CVPR.2016.596.

Cong Zhang, Kai Kang, Hongsheng Li, Xiaogang Wang, Rong Xie, and Xiaokang Yang. Data-driven crowd understanding: A baseline for a large-scale crowd dataset. *IEEE Transactions on Multimedia*, 18(6):1048–1061, 2016. doi: 10.1109/TMM.2016.2542585.

H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen. Dynamic R-CNN: Towards high quality object detection via dynamic training. In *Proc. Eur. Conf. Comput. Vis.*, pp. 260–275, 2020. doi: 10.1007/978-3-030-58555-6_16.

H. Zhang, F. Li, S. Liu, L. Zhang, and Y. Qiao. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *Proc. Int. Conf. Learn. Represent.*, 2023.

Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen. DETRs beat YOLOs on real-time object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 16965–16974, 2024. doi: 10.1109/CVPR52733.2024.01598.

Wentao Zhou, Chengtao Cai, Liying Zheng, Chenming Li, and Daohui Zeng. ASSD-YOLO: A small object detection method based on improved yolov7 for airport surface surveillance. *Multimed. Tools Appl.*, 83(18):55527–55548, 2024.