AMOD: A Large-scale Benchmark for RGB-T Multi-view Aerial Military Object Detection

Yechan Kim*
GIST
Republic of Korea
yechankim@gm.gist.ac.kr

Jongmin Joo
GIST
Republic of Korea
luck2u99@gm.gist.ac.kr

JongHyun Park
GIST
Republic of Korea
citizen135@gm.gist.ac.kr

Jongmin Yu
University of Cambridge
United Kingdom
jy522@projectg.ai

Moongu Jeon
GIST
Republic of Korea
mgjeon@gist.ac.kr

Abstract

Existing benchmarks for aerial object detection provide limited support for studying RGB-T perception under controlled conditions. They often lack synchronized multi-view observations of the same scenario, making it hard to analyze viewpoint variation and cross-modal learning in a unified setting, especially under large viewpoint changes and modality discrepancies. In this paper, we introduce AMOD, a new benchmark dataset for RGB-T aerial military object detection, constructed using a game Arma3. AMOD provides paired visible (RGB) and thermal (T) images with aligned annotations, along with multi-view observations of the same area of interest, enabling consistent analysis across viewpoints and modalities. The dataset comprises 73,920 images and 383,212 instances spanning 12 military categories, generated across various background maps with controlled viewpoint configurations. We further observe that, although AMOD is constructed with military object categories, models pretrained on AMOD also transfer effectively to real-world aerial benchmarks containing civilian objects, suggesting its utility beyond military-target detection. The dataset and details are available at https://unique-chan.github.io/AMOD-Project.

CCS Concepts

• **Computing methodologies** → **Object detection**.

Keywords

Synthetic Data, Benchmark, Military, Detection, Visible, Thermal, Multi-angular, Aerial Imagery, Remote Sensing, Earth Vision, Arma3

ACM Reference Format:

Yechan Kim, Jongmin Joo, JongHyun Park, Jongmin Yu, and Moongu Jeon. 2026. AMOD: A Large-scale Benchmark for RGB-T Multi-view Aerial Military Object Detection. In *Proceedings of xxth ACM International Conference on xxxxxxxxxx, xxxx 00–xxxx 00, 2026, xxxxxx, xxxxxxx (xx '26).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

xx '26, xxx xxxx xxxx, xxxxxxx

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

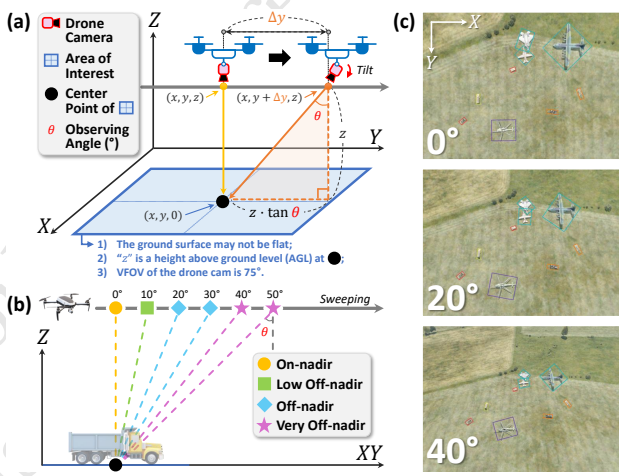


Figure 1: Illustration of how to construct our AMOD benchmark. (a) Geometric definition of the observing angle θ, where the camera shifts laterally by Δy in Y-axis, producing an off-nadir view. (b) Drone camera sweeping setup simultaneously capturing the same area from nadir to very off-nadir viewpoints. (c) Examples from the RGB version of AMOD under different observation angles, showing how target appearance and perspective distortion vary with θ.

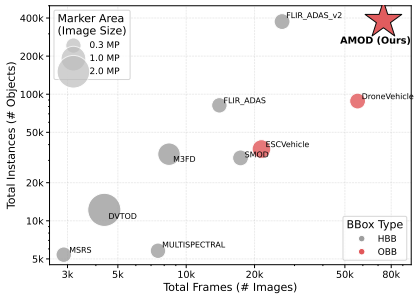
1 Introduction

Aerial object detection is essential for various applications, including visual surveillance, disaster monitoring, and critical security operations. In complex environments, multi-modal perception using visible (RGB) and thermal (T) imagery provides robust recognition [33]. However, advancing this cross-modal learning is severely hindered by a scarcity of paired training data [14, 41, 44]. This data shortage is particularly extreme for security and military targets, where acquiring real-world aerial imagery is strictly limited by operational constraints.

Beyond mere data scarcity, the unconstrained nature of aerial environments presents another critical challenge. Viewpoint variation significantly impacts object appearance due to changes in scale and orientation. However, because existing real-world datasets are collected under unconstrained conditions [43, 54], they lack the controlled variations needed to systematically evaluate both cross-view robustness and cross-modal learning. These datasets

Table 1: Comparison of the AMOD dataset with existing RGB-T object detection benchmarks containing aerial scenes.

Dataset	Scenario	Image Size	Total Frames	Bounding Box Type	Total Instances	Total Categories	Location	ID	Simultaneous Viewpoints
MULTISPECTRAL (MM-17) [35]	Driving (DV)	640x480	7,512	HBB	5.8k	5	Asia	N	Single
FLIR_ADAS (FLIR-18) [2]	DV	480x640	14,000	HBB	81.7k	4	North America	N	Single
DroneVehicle (TCSVT-22) [33]	Drone (DR)	640x512	56,878	OBB	88.3k	5	Asia	N	Single
FLIR_ADAS_v2 (FLIR-22) [2]	DV	640x512	26,442	HBB	375k	15	North America	N	Single
MSRS (IF-22) [36]	DV	480x640	2,888	HBB	5.4k	8	Asia	N	Single
M3FD (CVPR-22) [25]	DR + DV	1024x768	8,400	HBB	33.6k	6	Asia	N	Single
DVTOD (TIV-24) [32]	DR	1920x1080	4,358	HBB	12.2k	3	Asia	N	Single
SMOD (TMM-25) [9]	DV	640x512	17,352	HBB	31.4k	4	Asia	N	Single
ESCVehicle (TGRS-26) [31]	DR	704x704	21,454	OBB	36.9k	7	Asia	N	Single
AMOD (Ours)	**DR**	**1920x1440**	**73,920**	**OBB**	**383.2k**	**12**	**Multiple**	**Y**	**Multiple**



typically suffer from uncontrollable environmental variables, sensor misalignment, and heavily imbalanced viewpoint distributions. Furthermore, simply collecting more real imagery to address these gaps remains a costly and strictly constrained process [23, 29].

To overcome the prohibitive costs and limited controllability of real-world data collection, synthetic data generation offers a scalable alternative. More importantly, its controllability allows us to systematically incorporate two properties that remain insufficiently explored in prior aerial detection benchmarks: ① synchronized multi-view observations and ② strict cross-modal alignment. In this paper, we introduce ***AMOD***, a new benchmark dataset for RGB-T aerial object detection constructed using the Arma3 [1] simulation environment[1] AMOD provides strictly paired visible (RGB) and thermal (T) images with aligned annotations, together with synchronized multi-view observations of the exact same area of interest; see Fig. 1. In contrast to conventional datasets relying on manual labeling or post-hoc processing, annotations in AMOD are obtained directly from the game engine, ensuring geometric consistency and enabling stable large-scale data generation. The dataset comprises 73,920 images and 383,212 instances spanning 12 military categories, generated across diverse backgrounds and six viewing angles. By enabling controlled experiments on viewpoint variation and cross-modal perception, AMOD provides a specialized testbed for military-target detection while also showing utility beyond this domain for broader aerial object detection.

Using AMOD, we conduct a systematic empirical study and summarize our key contributions as follows:

(1) Incorporating multi-angular diversity during training substantially improves model generalization across viewpoints, highlighting the importance of geometric variability for robust aerial object detection.
(2) Models pretrained on AMOD achieve improved performance compared to those pretrained on existing RGB, T, RGB-T datasets.
(3) Despite being constructed with military object categories, AMOD transfers effectively to real-world aerial datasets containing civilian objects, demonstrating its utility beyond the original semantic domain and its applicability to general aerial object detection.

[1]The dataset is generated using our open source software called ***G-MAD*** (https://github.com/unique-chan/G-MAD), which enables structured scenario specification, synchronized multi-view RGB-T capture, and automatic annotation by directly querying engine-level precise geometric information.

2 Related work

2.1 Existing benchmark datasets for RGB-T aerial object detection

Existing RGB-T object detection benchmarks have substantially advanced multi-modal perception as summarized in Tab. 1[2]. However, only a limited subset is directly relevant to aerial scenes. M3FD [25] offers some aerial context, but it mixes ground and drone views. Representative detection benchmarks dedicated to aerial RGB-T imagery include DroneVehicle [33], DVTOD [32], and ESCVehicle [31]. Other datasets such as MULTISPECTRAL [35], FLIR_ADAS [2], MSRS [36], and SMOD [9] are primarily designed for ground-driving scenarios. Even among aerially relevant datasets, most focus on single-viewpoint observations. For example, DroneVehicle [33] has played an important role in cross-modality vehicle detection. Yet, it is still built around single-view drone flights. This is a critical limitation for aerial object detection. Object appearance varies significantly with viewing geometry due to scale changes, foreshortening effects, and self-occlusion. As models trained on single-view datasets struggle to generalize to unseen viewing angles, viewpoint variations are closely entangled with scene composition and environmental conditions. Furthermore, real-world RGB-T aerial datasets inherently suffer from imperfect cross-modal alignment. In the wild, visible and thermal images are rarely registered perfectly at the pixel level due to differences in sensor characteristics and capture timing [18, 48]. DVTOD [32] explicitly highlights this visible-thermal misalignment in drone scenes. Similar spatial mismatches degrade annotation consistency in DroneVehicle [33] and M3FD [25]. To overcome these limitations, AMOD utilizes a simulation game named Arma3 to generate a controlled benchmark with strictly paired RGB-T images and jointly aligned annotations across multiple viewpoints.

2.2 Multi-view aerial image understanding

Several lines of research in Earth vision leverage multi-view signals. The MVOI benchmark [40] reveals how off-nadir imagery degrades building extraction performance. It provides multi-angle RGB satellite imagery captured nearly simultaneously. However, manual labels are only provided for nadir views due to high annotation costs. These annotations are then reused for off-nadir views,

[2]In our comparison, we focus on RGB-T object detection benchmarks for aerial scenes that are not specifically designed for tracking or for predominantly tiny-object detection. Accordingly, tracking-oriented datasets such as VT-UAV [49], as well as tiny-object-focused benchmarks such as Drone-based RGBT Tiny Person Detection [51] and RGT-Tiny [48], are beyond the scope of this work.

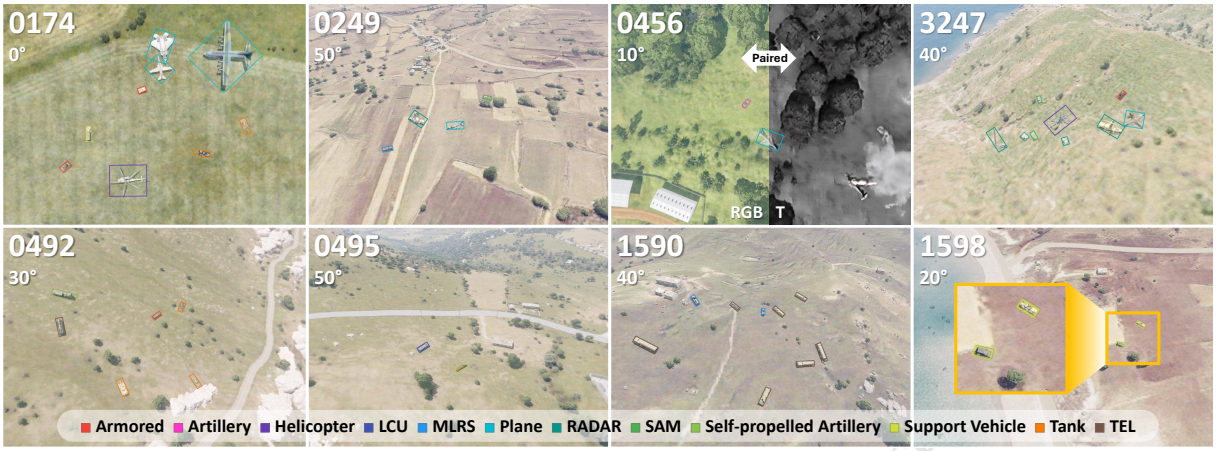


Figure 2: Annotated examples from the AMOD dataset featuring oriented bounding boxes (OBB). The dataset includes paired RGB-T imagery (see region '0456/10°' for instance). For multi-angular variants of region '0174', refer to Fig. 1(c).

limiting the precision of view-specific supervision. Another important direction concerns cross-view geo-localization and matching [17]. Game4Loc [16] utilizes the GTA V [3] environment to facilitate spatial correspondence research. These studies demonstrate the potential of simultaneous multi-view data for geometric reasoning. Still, their primary focus lies in geo-localization rather than category-level detection. AMOD bridges these research gaps by exploring how viewpoint diversity influences category-level detection in both RGB and thermal modalities. By leveraging the Arma3 game environment, we extract precise and simultaneous bounding box labels for all viewing angles, offering a comprehensive perspective on multi-view aerial understanding.

3 AMOD: a novel synthetic benchmark for RGB-T aerial military object detection

We introduce ***AMOD***[3], a new benchmark to provide synchronized multi-view aerial images using a game Arma3, enabling systematic analysis of view variation in aerial detection. We adopt the Arma3 [1] rather than general-purpose game engines such as Unity [4] and Unreal [5]. While those engines offer high-fidelity rendering, they often require custom asset modeling or significant licensing costs of domain-specific 3D models of equipment and terrain [7, 8, 11, 47], for large-scale data construction. In contrast, Arma3 already provides rich 3D object models with various terrain assets. This eliminates the need for heavy 3D modeling.

3.1 General setup

Category design. As Arma3 is a military FPS game, most of its available 3D assets are related to military equipment. We adopt a total of 513 models in Arma3, 258 from the Eastern and 255 from the Western faction. Each model is then categorized into one of 12 classes: **Armored**, **Artillery**, **Helicopter**, **Landing Craft Utility (LCU)**, **Multiple Launch Rocket System (MLRS)**, **Plane**, **RADAR**, **Surface-to-air Missile (SAM)**, **Self-propelled Artillery**, **Support Vehicle**, **Tank**, and **Transporter Erector Launcher (TEL)**; the distribution of Arma3 assets assigned to each category is illustrated in Fig. 3(a).

Objects placement (spawn). Let $\mathcal{A}$ denote an area of interest (or target imaging area) for which simultaneous multi-view captures are provided as in Fig. 1. For each $\mathcal{A}$, our data generator[4] randomly selects and positions items as follows:

- **Step 1.** Sample k classes among the above 12 categories, and store them into a set S. Here, k is chosen between 1 and 8.
- **Step 2.** Randomly determine the number of items e to be placed in the $\mathcal{A}$ (e is between 8 and 14). The e items are selected from the 513 assets, restricted to the classes in S.

To further enhance data diversity, each object is assigned a unique rotation angle. Besides, we enforce that land assets must be located only on land and naval assets for sea, and that no objects overlap with each other.

Recording setup. As shown in Fig. 1, we control the Arma3 to capture six simultaneous multi-view, RGB-T paired images for each $\mathcal{A}$, corresponding to observation angles of 0°, 10°, 20°, 30°, 40°, and 50°. Along with the images, the associated metadata is also extracted to generate bounding-box annotations for each view. Each image has a size of 1920×1440 and the cameras are configured such that the central pixel of the 0°-observing image is a ground sampling distance (GSD) of 0.1 m/pixel. For this nominal GSD, we set the camera altitude (z) to 120 m with Vertical FOV[5] 75° in Arma3. When z exceeds 120 m, shadows largely disappear, resulting in a notable loss of image detail in Arma3. Since the presence of shadows in aerial imagery is known to be of significant research importance [26], we do not consider z >120 m. Besides, we set the time range from 9 to 18 for capturing different lighting conditions[6]. Meanwhile, to enlarge the geographical coverage, we leverage several official Arma3 maps as follows[7]:

- For train/validation splits: Altis, Malden, Stratis, Tanoa, Weferlingen;
- For test split: Malden and Livonia.

[3]It stands for "Aerial Military Object Detection in multi-view RGB-T scenarios."

[4]The values of k and e are empirically chosen and can be changed.

[5]VFOV 75° is a broad viewing angle similar to that of a standard wide-angle camera. As our primary focus is on unprocessed UAV imagery, neither orthorectification nor lens distortion correction is applied. Nevertheless, pretraining with AMOD improves performance over the baseline when evaluated on real-world datasets like DIOR-R, composed of satellite images from Google Earth (Standard Orthophoto).

[6]In Arma3, visible images captured at night are usually rendered entirely black, making them unusable. Therefore, we do not collect RGB-T pairs under nighttime conditions.

[7]Note that all maps are utilized at a nearly equal frequency.

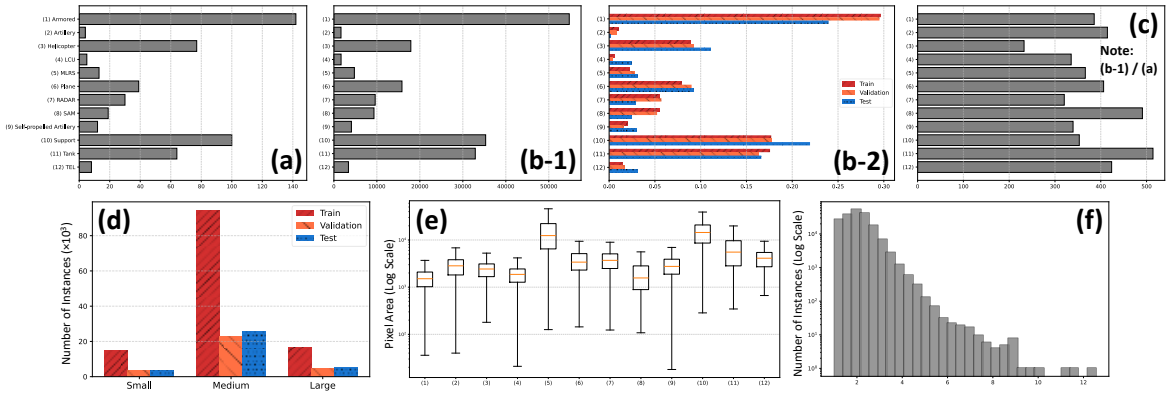


Figure 3: Statistics of the AMOD dataset. (a) Object models per category used in Arma3. (b-1) Instance count per category. (b-2) Instance distribution per category in train/dev/test splits. (c) Normalized instance count by Arma3 models across categories. (d) Size distribution (Small/Medium/Large) in train/dev/test splits. (e) Instance size of OBBs per category. (f) Aspect ratio distribution of OBBs per category. Note that RGB and thermal modalities are perfectly aligned, with identical image counts, instance counts, and bounding box annotations in our benchmark.

Figure 4: Examples of (a) low inter-class and (b) high intra-class variances in Arma3 models we used for our dataset.

Annotation. While capturing each scene, our generator obtains the 3D bounding box using 'boundingBoxReal' function of Arma3; see Supp. Material online for details. It then projects its corner points onto the image plane, and derive both horizontal and oriented bounding box (HBB/OBB) annotations. The HBB is computed by taking the minimum and maximum projected coordinates, while the OBB is obtained by applying a convex hull [13] followed by rotating calipers [37] to find the minimum-area enclosing rectangle. However, OBBs obtained via the rotating calipers algorithm can be suboptimal (i.e., relatively loose) at certain instances. Since tighter bounding boxes are generally more desirable for detector training and also beneficial for fair benchmarking [29, 30, 46], we further refine the initial OBBs. Specifically, we employ the Segment Anything Model (SAM) [21] to extract object masks from the initial OBBs, and subsequently reconstruct tighter OBBs based on these masks. Detailed refinement procedure and its relevant source code are provided in Supp. Material and our online website. Finally, as summarized in Tab. 1, in our benchmark, to facilitate future multi-view studies, the same object observed across different views is assigned a consistent object ID for each region of interest ($\mathcal{A}$).

Data balancing. Class imbalance is undesirable, as it may bias model training toward overrepresented categories and degrade generalization performance [20]. However, achieving perfectly uniform class distributions is infeasible in our data because the number of available 3D assets varies across categories as seen in Fig. 3(a). To address this, we post-process our initially generated data so that the ratio between the number of instances per class and the number of Arma3 items is made as uniform as possible; see Fig. 3(c).

3.2 Statistics, challenges, and applications

Train, validation, and test splits. The dataset is partitioned into three subsets: a training set comprising 47,808 images (64.68%), a validation set with 11,988 images (16.21%), and a test set with 14,124 images (19.11%). We maintain similar class distributions across all splits as shown in Fig. 3(b-2).

Object size and aspect ratio distributions. Following [12], we categorize object sizes into three groups: small ($\leq$ 32×32 pixels), medium (between 32×32 and 96×96), and large ($\geq$ 96×96). The object size distribution of our dataset is illustrated in Fig. 3(d). Meanwhile, Fig. 3(e) and (f) illustrate the average size and aspect ratio of OBBs for each category, respectively.

Distinct characteristics of AMOD. Our dataset is designed with three notable properties as follows.

(1) It provides *simultaneous multi-view observations* of each region of interest ($\mathcal{A}$), enabling research on cross-view consistency in aerial image understanding. Owing to the complex terrain in Arma3 environments, such as cliffs and mountains, certain objects are occluded in some viewpoints[8].
(2) It exhibits both *low inter-class and high intra-class variations*. For instance, visually similar vehicle types such as tanks and armored vehicles challenge fine-grained recognition, while a single class like RADAR includes highly diverse structures, as illustrated in Fig. 4.
(3) It ensures *diverse geographical backgrounds* by leveraging diverse official Arma3 maps, thereby enhancing scene variability and model's generalization ability.

Potential applications of AMOD. Ultimately, AMOD establishes a useful benchmark for aerial object detection across diverse viewing conditions. The provision of paired RGB-T data further makes it a practical testbed for developing and evaluating multi-modal fusion methods. Crucially, AMOD also serves as an effective pre-training source for transfer to real-world aerial domains; as shown in Sec. 4.4, this benefit extends beyond military targets and also improves the detection of civilian objects.

[8]We automatically filter objects as fully occluded if their direct line-of-sight intersects with other scene elements.

Table 2: Class-wise Test AP on AMOD.

	Armored	Artillery	Helicopter	LCU	MLRS	Plane	RADAR	SAM	Self-propelled Artillery	Support	Tank	TEL	Mean
AP_{75}	40.30	74.20	50.40	78.30	65.20	79.30	25.70	30.60	37.40	40.60	47.40	53.70	51.93
$AP_{50:95}$	47.19	64.35	50.47	58.27	56.08	65.20	31.65	38.92	43.53	46.46	48.91	54.76	50.48

Table 3: Cross-angular evaluation results (AP_{75}) on AMOD.

	Test						
Train	0°	10°	20°	30°	40°	50°	Mean
0°	53.71	49.61	46.23	36.81	27.00	18.27	38.61
10°	52.44	48.81	46.88	38.53	32.73	17.50	39.48
20°	51.87	50.50	47.81	44.41	33.97	21.96	41.75
30°	49.39	48.73	46.74	42.49	37.13	23.60	41.35
40°	45.29	45.95	45.71	41.15	34.68	28.20	40.16
50°	32.88	31.43	33.84	33.50	29.56	28.53	31.62
All	**60.89** (+7.18)	**59.62** (+9.12)	**57.35** (+9.54)	**52.78** (+8.37)	**45.24** (+8.11)	**37.83** (+9.30)	**52.29** (+10.54)
All (Reduced to $\frac{1}{6}$)	**56.92** (+3.21)	**52.51** (+2.01)	**53.22** (+5.41)	**50.84** (+6.43)	**39.74** (+2.61)	**32.44** (+3.91)	**47.61** (+5.86)

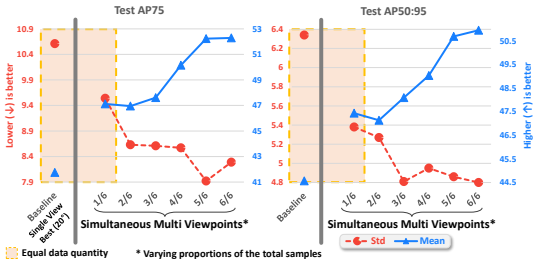


Figure 5: Effect of angular diversity versus data quantity on AMOD (AP_{75}, $AP_{50:95}$). The blue and red dots denote the mean and standard deviation of AP values across all observation angles, respectively.

Table 4: Comparison between AMOD and ImageNet pretraining on two real-world benchmarks, DIOR-R (visible) and HIT-UAV (thermal), evaluated using AP_{50}.

Pretrained from		Finetuned for DIOR-R					Finetuned for HIT-UAV			
Source Data	View Diversity	Airplane	Ship	Vehicle	Windmill	Mean	Person	Car	Others	Mean
ImageNet	–	71.88	81.02	42.59	57.07	63.14	85.30	81.29	57.12	74.57
AMOD° (■)	✗ (0°)	79.41	81.60	43.82	56.55	65.35 (+2.21)	–	–	–	–
	✓ (0°–20°)	78.92	81.83	44.47	56.55	65.44 (+2.30)	–	–	–	–
	✓ (0°–50°)	**79.96**	82.05	44.54	56.79	65.84 (+2.70)	86.59	82.00	58.51	75.70 (+1.13)
AMOD* (■)	✗ (0°)	77.65	86.00	52.48	56.19	68.08 (+4.94)	–	–	–	–
	✓ (0°–20°)	78.22	89.03	**53.27**	57.24	69.44 (+6.30)	–	–	–	–
	✓ (0°–50°)	79.71	**89.04**	53.21	**57.36**	**69.83** (+6.69)	**87.68**	**82.15**	**61.40**	**77.08** (+2.51)

4 Experiments

Following prior work on aerial object detection [19, 22, 39, 52], we adopt Oriented R-CNN [45] with a Swin-S [28] backbone and FPN-1x [24] as our baseline model. In addition, for RGB-T multimodal training, motivated by [50], we adopt channel-wise stacking (early-fusion)[9] of RGB and thermal (T) images at the input level as our baseline[10]. The resulting four-channel input is fed into a Swin-S backbone, and the extracted features are subsequently passed to an Oriented R-CNN detector. This design is efficient in terms of both parameter count and computational cost. All experiments are implemented using the MMRotate framework [53].

[9]To simplify our prototype experiments, we rely only on the thermal split annotations from DroneVehicle [33] during both training and testing.

[10]Since the RGB-T image pairs in our dataset are perfectly aligned at the pixel level, no additional registration or feature-level alignment is required [15]. This makes such a simple early-fusion strategy both feasible and effective.

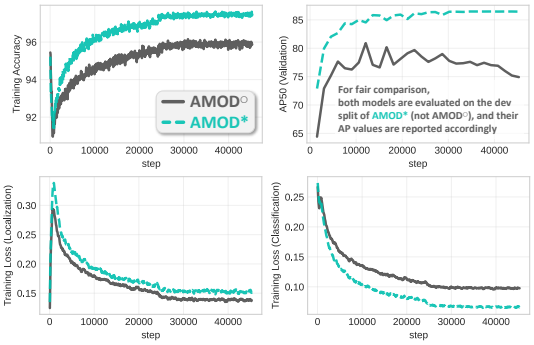


Figure 6: Learning curves for two AMOD variants: AMOD* (■) denotes the corrected version of the initial AMOD labels (■), resulting in tighter bounding box annotations. Unless otherwise specified, AMOD refers to AMOD* (■), a corrected version with refined bounding-box annotations, and all experimental results are reported on this, in our paper.

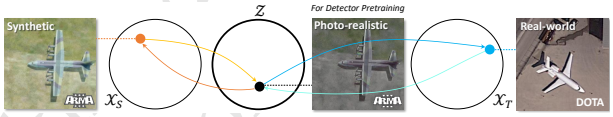


Figure 7: Illustration of the shared latent space ($\mathcal{Z}$) for domain adaptation between our synthetic ($\mathcal{X}_S$) and target real-world ($\mathcal{X}_T$) domains using UNIT [27]. Both domains are mapped to $\mathcal{Z}$. Arma3-rendered photo-realistic images are generated as intermediate samples to bridge $\mathcal{X}_S$ and $\mathcal{X}_T$.

Table 5: Investigating the effectiveness of AMOD-pretraining on a real-world RGB-T benchmark, DroneVehicle (AP_{50}).

Pretrained from	Finetuned for DroneVehicle				
	Car	Van	Truck	Bus	Mean
DroneVehicle (Baseline)	87.81	54.73	64.65	87.39	73.65
AMOD	**89.02**	**55.69**	**69.25**	**88.54**	**75.62** (+1.97)

Figure 8: Qualitative analysis of an AMOD-trained detector on unseen real-world RGB imagery containing military equipment on real-world RGB satellite photographs released through public news coverage of the Russia–Ukraine war: (a) and (b): Successful detections of helicopters and airplanes. (c): A failure case of detecting tanks.

4.1 Class-wise performance on AMOD

We first report class-wise detection performance on the RGB split of AMOD in Tab. 2, which serves as a reference breakdown of category-level performance under the default evaluation setting. Categories such as LCU (78.30 AP_{75}) and Plane (79.30 AP_{75}) achieve relatively high performance, likely because they exhibit distinctive global shapes with limited visual ambiguity. In contrast, categories such as RADAR (25.70 AP_{75}) and SAM (30.60 AP_{75}) show substantially

lower performance. We hypothesize that this is partly attributable to high intra-class variation; as shown in Fig. 4(b).

4.2 Multi-view learning on AMOD

AMOD's synchronized multi-view observations faciliate a controlled analysis of viewpoint generalization in aerial object detection. Using the RGB subset, we investigate how angular diversity during training affects robustness to unseen viewing angles. Detectors trained on a single observation angle show a clear angle-dependent bias; see Tab. 3. Performance drops significantly as the test viewpoint diverges from the training angle. For instance, a model trained at 0° achieves 53.71 AP_{75} at 0° but drops to 18.27 at 50°. The best single-angle model yields a mean AP_{75} of only 41.75. Single-view training clearly lacks the geometric diversity needed for robust cross-angle generalization. In contrast, a multi-view model trained jointly on all six angles outperforms the single-angle baselines across every test angle. It achieves a 52.29 mean AP_{75}, which is 10.54 points higher than the best single-angle model. This gain spans the full angular range from 0° to 50°. To separate the impact of angular diversity from dataset size, we also evaluate a reduced multi-view model trained on one-sixth of the total data. This model still achieves 47.61 mean AP_{75}, outperforming the best single-angle baseline by 5.86 points. Additionally, Fig. 5 shows that increasing multi-view diversity improves the mean test AP and reduces the standard deviation across observation angles. Overall, these results confirm that viewpoint diversity plays a central role in generalization.

Importantly, a similar trend is observed in real-world transfer. As shown in Tab. 4, increasing view diversity during AMOD pretraining consistently improves downstream performance on DIOR-R. For example, the mean AP_{50} increases from 65.35 (0° only) to 65.44 (0°–20°) and further to 65.84 (0°–50°) in the AMOD° setting, and from 68.08 to 69.44 and 69.83 in the AMOD* setting; note that details about AMOD° and AMOD* are described in Sec. 4.3. This indicates that angular diversity learned from synthetic aerial data transfers to real-world detection scenarios. Overall, these results confirm that viewpoint diversity plays a central role in generalization.

4.3 Annotation refinement of AMOD

We analyze the impact of annotation quality by comparing two AMOD variants: the initial version (denoted as AMOD°) and a refined version with tighter bounding box annotations (denoted as AMOD*). Fig. 6 shows that annotation refinement leads to more stable and efficient optimization. The model trained on AMOD* exhibits faster convergence and achieves a higher final validation AP. However, AMOD° achieves a lower final loss of localization than AMOD*, as the latter's ground truth targets are more precise, thus harder to learn. A tighter bounding box (BBox) is more difficult to hit perfectly, leading to a higher, but more meaningful, loss value[11]. Meanwhile, AMOD* achieves a significantly lower classification loss, implying that the noisy BBox labels in the AMOD° hinder the model to learn discriminative features for classification.

This improvement also translates to downstream performance. As shown in Tab. 4, models pretrained on AMOD* outperform those pretrained on the initial AMOD across both DIOR-R and HIT-UAV. For example, on DIOR-R, the mean AP_{50} improves from 65.84 to 69.83 under full view diversity, and on HIT-UAV, it increases from 75.70 to 77.08. These results indicate that annotation quality plays a critical role not only in optimization behavior but also in cross-dataset generalization.

4.4 Real-world domain adaptation of AMOD

Though our synthetic data offers scalable and precisely annotated samples, the visual characteristics of synthetic images significantly differ from real-world data in aspects such as color tone, texture, and background complexity [38]. This discrepancy is generally referred to as *domain gap*. Even in real-world scenarios, domain gaps can still naturally arise due to varying environmental factors, sensors, and imaging modalities [6, 42]. To bridge this gap, we adopt a pipeline consisting of domain transfer (see Fig. 7), pretraining, and finetuning detectors; see Supp. Material online for details.

In this section, we intentionally assume that labeled target-domain data are unavailable during pretraining. Therefore, for UNIT [27]-based domain transfer, we use similar proxy datasets instead of the actual target data: ① DOTA [43] for DIOR-R [10] and DroneVehicle (RGB) [33]; ② DroneVehicle (T) [33] for HIT-UAV [34]; and ③ HIT-UAV for DroneVehicle (T). This setup reflects realistic deployment scenarios and allows us to fairly evaluate the transferability and generalization capacity. The experimental results in Tabs. 4 and 5 indicate that our AMOD serves as an effective pretraining source for real-world aerial detection tasks. As shown in these tables, models pretrained on AMOD consistently outperform ImageNet-pretrained counterparts across multiple downstream benchmarks, including DIOR-R, HIT-UAV, and DroneVehicle.

Furthermore, the qualitative examples in Fig. 8 indicate that AMOD-pretrained detectors with ①-style transfer retain the ability to localize and classify military objects in previously unseen real-world RGB imagery, without finetuning[12]. The successful detections of helicopters and airplanes suggest that AMOD improves not only quantitative performance but also cross-domain generalization to some extent. However, missed detection of tanks highlights a remaining limitation. This issue is expected to be mitigated through fine-tuning with additional real-world military data in future work.

5 Conclusions

In this paper, we introduced AMOD, a large-scale synthetic benchmark for RGB-T multi-view military object detection in aerial scenes. AMOD provides strictly paired RGB-T imagery with consistent annotations across multiple viewpoints, enabling controlled study of viewpoint variation and cross-modal learning. Experiments show that multi-view diversity improves viewpoint generalization and that AMOD pretraining consistently outperforms ImageNet initialization on multiple real-world benchmarks. Despite being built from synthetic military scenarios, AMOD transfers effectively to real-world aerial imagery, including unseen environments and civilian object categories. As future work, we plan to extend AMOD toward more complex settings, including UAV–UGV scenarios and fine-grained classification.

[11] In other words, higher localization loss on AMOD* should not be interpreted as inferior localization learning; rather, it reflects stricter and tighter supervision targets.

[12] Although AMOD focuses on military objects, real-world military validation in this paper is restricted to a small set of unseen qualitative examples because large-scale public aerial datasets with military targets are difficult to obtain and curate.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) Grant by the Korean Government through Ministry of Science and ICT (MSIT) under Grant No. RS-2026-25475324. This work also benefited from high-performance GPU resources provided by HPC-AI Open Infrastructure via GIST SCENT (A100). The authors would like to thank SooYeon Kim (KRIT) for her assistance while working at GIST.

- *Disclaimer: Our AMOD dataset is intended strictly for research purposes, and we will continuously update its license terms and provide additional training materials through our website (https://unique-chan.github.io/AMOD-Project).*

References

[1] [Website]. Arma3. https://arma3.com.
[2] [Website]. Free FLIR Thermal Dataset for Algorithm Training. https://oem.flir.com/solutions/automotive/adas-dataset-form/.
[3] [Website]. Grand Theft Auto V. https://www.rockstargames.com/gta-v.
[4] [Website]. Unity. https://unity.com.
[5] [Website]. Unreal. https://www.unrealengine.com.
[6] Eunsu Baek, Keondo Park, Jiyoon Kim, and Hyung-Sin Kim. 2024. Unexplored faces of robustness and out-of-distribution: Covariate shifts in environment and sensor domains. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 22294–22303.
[7] Maciej Bala, Yin Cui, Yifan Ding, Yunhao Ge, Zekun Hao, Jon Hasselgren, Jacob Huffman, Jingyi Jin, JP Lewis, Zhaoshuo Li, et al. 2024. Edify 3d: Scalable high-quality 3d asset generation. *arXiv preprint arXiv:2411.07135* (2024).
[8] Octavian Blaga and David Scott. 2025. Breaking the Bottleneck: Synthetic Data as the New Foundation for Vision AI. *Synetic AI - Report* (2025), 1–14.
[9] Zizhao Chen, Yeqiang Qian, Xiaoxiao Yang, Chunxiang Wang, and Ming Yang. 2025. AMFD: Distillation via adaptive multimodal fusion for multispectral pedestrian detection. *IEEE Trans. Multimedia* (2025).
[10] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. 2022. Anchor-free oriented proposal generator for object detection. *IEEE Trans. Geosci. Remote Sens.* 60 (2022), 1–11.
[11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Conf. Robot Learn. (CoRL)*. PMLR, 1–16.
[12] Aritra Dutta, Srijan Das, Jacob Nielsen, Rajatsubhra Chakraborty, and Mubarak Shah. 2024. Multiview aerial visual recognition (mavrec): Can multi-view improve aerial visual perception?. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 22678–22690.
[13] Ronald L. Graham. 1972. An efficient algorithm for determining the convex hull of a finite planar set. *Info. Proc. Lett.* 1 (1972).
[14] Shuaihao Han, Tingfa Xu, Peifu Liu, and Jianan Li. 2026. MODA: The First Challenging Benchmark for Multispectral Object Detection in Aerial Images. In *AAAI Conf. Artif. Intell. (AAAI)*, Vol. 40. 4574–4582.
[15] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. 2015. Multispectral pedestrian detection: Benchmark dataset and baseline. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 1037–1045.
[16] Yuxiang Ji, Boyong He, Zhuoyue Tan, and Liaoni Wu. 2025. Game4loc: A uav geo-localization benchmark from game data. In *AAAI Conf. Artif. Intell. (AAAI)*, Vol. 39. 3913–3921.
[17] Yuxiang Ji, Boyong He, Zhuoyue Tan, and Liaoni Wu. 2025. MMGeo: Multimodal Compositional Geo-Localization for UAVs. In *IEEE Int. Conf. Comput. Vis. (ICCV)*. 25165–25175.
[18] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. 2021. LLVIP: A visible-infrared paired dataset for low-light vision. In *IEEE Int. Conf. Comput. Vis. (ICCV) Worksh.* 3496–3504.
[19] Yechan Kim, SooYeon Kim, and Moongu Jeon. 2025. Nbbox: Noisy bounding box improves remote sensing object detection. *IEEE Geosci. Remote Sens. Lett.* 22 (2025), 1–5.
[20] Yechan Kim, Younkwan Lee, and Moongu Jeon. 2021. Imbalanced image classification with complement cross entropy. *Pattern Recognit. Lett.* 151 (2021), 33–40.
[21] Alexander Kirillov et al. 2023. Segment anything. In *IEEE Int. Conf. Comput. Vis. (ICCV)*.
[22] Yujie Lei, Jie Zhang, Wenjie Sun, Wei He, Jiasong Zhu, and Qingquan Li. 2025. MGNet: A Remote Sensing Oriented Object Detector Based on Multi-Cascaded Feature Selection and Geometric Constraints. *IEEE Trans. Geosci. Remote Sens.* (2025).
[23] Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. 2023. Omnicity: Omnipotent city understanding with multi-level and multi-view images. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 17397–17407.
[24] Tsung-Yi Lin et al. 2017. Feature pyramid networks for object detection. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 2117–2125.
[25] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 5802–5811.
[26] Jiahang Liu, Xiaozhen Wang, Mao Guo, Ruilei Feng, and Yue Wang. 2023. Shadow detection in remote sensing images based on spectral radiance separability enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 5 (2023), 3438–3449.
[27] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*. 1–9.
[28] Ze Liu et al. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE Int. Conf. Comput. Vis. (ICCV)*.
[29] Jeffri Murrugarra-Llerena et al. 2022. Can we trust bounding box annotations for object detection?. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Worksh.*
[30] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 658–666.
[31] Jiamin Song, Nan Zhang, Zhenhao Wang, and Tian Tian. 2026. ESCVehicle: A Drone-based Visible-Infrared Vehicle Benchmark with Extensive Scene Coverage. *IEEE Trans. Geosci. Remote Sens.* (2026).
[32] Kechen Song, Xiaotong Xue, Hongwei Wen, Yingying Ji, Yunhui Yan, and Qinggang Meng. 2024. Misaligned visible-thermal object detection: A drone-based benchmark and baseline. *IEEE Trans. Intell. Veh.* (2024).
[33] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. 2022. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans. Circuits Syst. Video Technol.* 32, 10 (2022), 6700–6713.
[34] Jiashun Suo, Tianyi Wang, Xingzhou Zhang, Haiyang Chen, Wei Zhou, and Weisong Shi. 2023. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection. *Sci. Data.* 10, 1 (2023), 227.
[35] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Multispectral object detection for autonomous vehicles. In *ACM Int. Conf. Multimedia (MM) Worksh.* 35–43.
[36] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. 2022. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion* 83 (2022), 79–92.
[37] Godfried T Toussaint. 1983. Solving geometric problems with the rotating calipers. In *IEEE Melecon*, Vol. 83.
[38] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. 2021. Pixel-wise crowd understanding via synthetic data. *Int. J. Comput. Vis.* 129, 1 (2021), 225–245.
[39] Tao Wang, Chenyu Lin, Chenwei Tang, Jizhe Zhou, Deng Xiong, Jianan Li, Jian Zhao, and Jiancheng Lv. 2026. Adaptive image zoom-in with bounding box transformation for UAV object detection. *ISPRS J. Photogramm. Remote Sens.* 233 (2026), 452–466.
[40] Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Sean McPherson, Jacob Shermeyer, Varun Kumar, and Hanlin Tang. 2019. Spacenet mvoi: A multi-view overhead imagery dataset. In *IEEE Int. Conf. Comput. Vis. (ICCV)*. 992–1001.
[41] Xin Wen, Haixu Yin, Kai Li, Wanying Nie, Jianxun Zhao, and Kechen Song. 2025. CMAI-Det: Cross-Modal Alignment and Interaction for RGB-T Object Detection in Drone Scenes. *IEEE Trans. Geosci. Remote Sens.* 63 (2025), 1–11.
[42] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-infrared cross-modality person re-identification. In *IEEE Int. Conf. Comput. Vis. (ICCV)*. 5380–5389.
[43] Gui-Song Xia et al. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*.
[44] Jiuhong Xiao, Roshan Nayak, Ning Zhang, Daniel Tortei, and Giuseppe Loianno. 2025. ThermalGen: Style-Disentangled Flow-Based Generative Models for RGB-to-Thermal Image Translation. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*.
[45] Xingxing Xie, Gong Cheng, Jiabao Wang, Ke Li, Xiwen Yao, and Junwei Han. 2024. Oriented R-CNN and beyond. *Int. J. Comput. Vis.* 132, 7 (2024), 2420–2442.
[46] Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. 2022. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* 190 (2022), 79–93.
[47] Ze Yang, Jingkang Wang, Haowei Zhang, Sivabalan Manivasagam, Yun Chen, and Raquel Urtasun. 2025. GenAssets: Generating in-the-wild 3D Assets in Latent Space. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 22392–22403.
[48] Xinyi Ying, Chao Xiao, Wei An, Ruojing Li, Xu He, Boyang Li, Xu Cao, Zhaoxu Li, Yingqian Wang, Mingyuan Hu, et al. 2025. Visible-thermal tiny object detection: A benchmark dataset and baselines. *IEEE Trans. Pattern Anal. Mach. Intell.* 47, 7 (2025), 6088–6096.

[49] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. 2022. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 8886–8895.

[50] Xue Zhang, Si-Yuan Cao, Fang Wang, Runmin Zhang, Zhe Wu, Xiaohan Zhang, Xiaokai Bai, and Hui-Liang Shen. 2024. Rethinking early-fusion strategies for improved multispectral object detection. *IEEE Trans. Intell. Veh.* (2024).

[51] Yan Zhang, Chang Xu, Wen Yang, Guangjun He, Huai Yu, Lei Yu, and Gui-Song Xia. 2023. Drone-based RGBT tiny person detection. *ISPRS J. Photogramm. Remote Sens.* 204 (2023), 61–76.

[52] Shangdong Zheng, Yang Xu, Peng Zheng, Zhihui Wei, and Zebin Wu. 2026. See Hidden Insight From Transposition: Multi-Axis Feature Aggregation for Aerial Object Detection. *IEEE Trans. Geosci. Remote Sens.* (2026).

[53] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, et al. 2022. Mmrotate: A rotated object detection benchmark using pytorch. In *ACM Int. Conf. Multimedia (MM)*. 7331–7334.

[54] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. 2021. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 11 (2021), 7380–7399.