Multiview Aerial Visual Recognition (MAVREC): Can Multi-view Improve Aerial Visual Perception?

Aritra Dutta[1], Srijan Das[2], Jacob Nielsen[3], Rajatsubhra Chakraborty[2], and Mubarak Shah[4]
[1] AI Initiative, UCF, [2] UNC Charlotte, [3] IMADA, SDU, [4] CRCV, UCF
https://mavrec.github.io

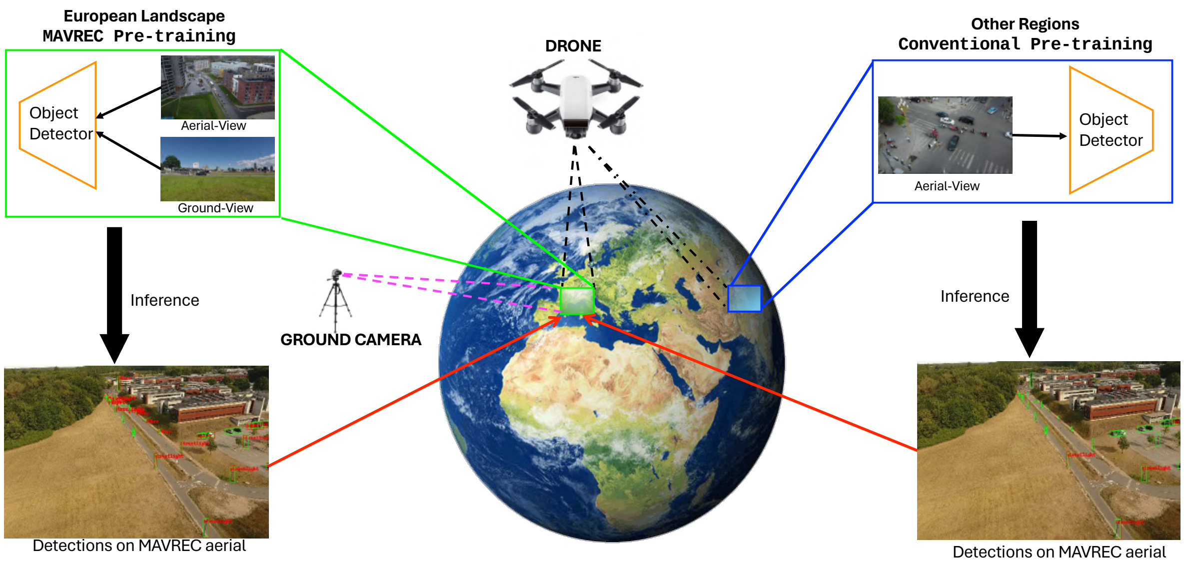

Figure 1. Illustration of the geography-aware model using our proposed MAVREC dataset (green box) collected in the rural and urban European landscape vs. the conventional aerial object detector (blue box) pretrained only on aerial images from VisDrone [91] captured in Asia. The conventional approach fails to detect aerial objects from the MAVREC dataset precisely. In contrast, our object detector pretrained on the ground and aerial images from the MAVREC dataset contextualizes the object proposals of that specific geography and enhances the aerial visual perception, thus outperforming other object detectors pre-trained on popular ground-view dataset (MS-COCO [44]) or other aerial datasets collected from different geographies; also, see Figure 5.

Abstract

Despite the commercial abundance of UAVs, aerial data acquisition remains challenging, and the existing Asia and North America-centric open-source UAV datasets are small-scale or low-resolution and lack diversity in scene contextuality. Additionally, the color content of the scenes, solar zenith angle, and population density of different geographies influence the data diversity. These factors conjointly render suboptimal aerial-visual perception of the deep neural network (DNN) models trained primarily on the ground view data, including the open-world foundational models.

*To pave the way for a transformative era of aerial detection, we present **M**ultiview **A**erial **V**isual **REC**ognition or **MAVREC**, a video dataset where we record synchronized scenes from different perspectives — ground camera and drone-mounted camera. MAVREC consists of around 2.5 hours of industry-standard 2.7K resolution video sequences, more than 0.5 million frames, and 1.1 million annotated bounding boxes. This makes MAVREC the largest ground and aerial view dataset, and the fourth largest among all drone-based datasets across all modalities and tasks. Through our extensive benchmarking on MAVREC, we recognize that augmenting object detectors with ground view images from the corresponding geographical location is a superior pre-training strategy for aerial detection. Building on this strategy, we benchmark MAVREC with a curriculum-based semi-supervised object detection approach that leverages labeled (ground and aerial) and unlabeled (only aerial) images to enhance aerial detection.*

1. Introduction

Object detection and tracking employing UAV (or drone)-based aerial videos are essential in many downstream applications, such as autonomous driving [15, 53], robotics [81], environmental monitoring [60], infrastructure inspection [13], developing livable and safe communities [6, 31, 89], a few to name. Despite having many crucial applications, most visual perception models focus on ground view images. This bias results in suboptimal performance when these models are applied to an aerial perspective — a discrepancy due to a domain shift precipitated by the viewpoint transfer. We hypothesize that the basis of this disparity is primarily twofold.

First, the lack of diversity in the current aerial datasets. Modern DNN-based visual models are data-hungry. However, aerial data collection is intricate due to UAV flight regulations and safety protocol, atmospheric turbulence, and many more [32]. The existing open-source UAV datasets [16, 17, 21, 50, 52, 55, 69, 91] are either small-scale, or low-resolution, and collected primarily in the urban pasture across Asian and North American geographies. These factors contribute to inadequacy in diverse dataset properties and hinder training large DNNs for aerial visual perception.

Second, substandard generalizability of the existing aerial visual models across different geographic locations. An object detection model trained on datasets captured from South Asia underperforms when deployed on videos captured with high-latitude northern European landscapes, characterized by semi-rural pastures and lots of greenery; see Figure 1. Research shows that different geographic factors and climate conditions, including latitude influence the population density [8, 37, 38], hence, the *color-content of the scenes*[1], their complexities, and density and interactions of the foreground objects. These seemingly low-key factors directly affect the data captured from a drone-mounted camera, and the previous studies never considered the inter-domain inference quality of the DNN models trained on these data.

To pave the way for a transformative era of aerial visual perception models, we introduce **M**ultiview **A**erial **V**isual **REC**ognition dataset, **MAVREC**, which uniquely captures time-synchronized aerial and ground view data. MAVREC is collected with consumer-grade handheld cameras (smartphones and GoPro) and drone-mounted cameras, consists of around 2.5 hours of industry-standard 2.7K resolution video sequences, more than 0.5 million frames, covering rural and urban pastures during spring and summer in high-latitude northern European geographies. It makes *MAVREC the largest ground and aerial view dataset*, and *the fourth largest among all drone-based datasets (edited and unedited) across all modalities and tasks that ever existed*; see Table 1.

In this paper, we rigorously assess our hypothesis and explore interesting properties of object detection in aerial images while evaluating MAVREC in a supervised setting. We find that contextual information of the landscape vastly influences aerial object detection, which is not the case for general object detection in ground view; also see [49]. Therefore, when an aerial object detection model is trained exclusively within a specific geographical context, it often exhibits limited generalizability across diverse geographical locations. This limitation necessitates the development of aerial detection models capable of understanding and integrating geography-specific features. Our experimental analysis with MAVREC shows that transfer learning from ground to aerial view induces geography-aware representations in aerial object detection models. This curriculum-based training approach notably surpasses the performance of object detectors pre-trained on alternative aerial or ground datasets, including advanced foundational models such as Grounding DINO [46], trained on extensive data corpora; see details in §4. Furthermore, we reckon significant resource investment in annotating large-scale aerial object detection datasets. Annotations are cost-intensive and demand substantial human intervention and time, rendering them impractical for numerous real-world applications. To this end, we benchmark MAVREC with a curriculum-based semi-supervised object detection approach that leverages labeled and unlabeled images to enhance the detection from an aerial perspective.

We summarize our key technical contributions as follows:

- We introduce MAVREC, which to date represents the most extensive dataset integrating time-synchronized ground and aerial images captured in the European landscape; §3.
- Through benchmarking MAVREC in supervised and semi-supervised settings, we expose the proclivity of existing pre-trained object detectors to exhibit bias toward data captured from ground perspectives; §4.
- We propose a curriculum-based semi-supervised object detection method. Its superior performance shows the importance of training these types of models with ground view images to learn geography-aware representation; §4.2.

2. Related work

Aerial datasets. The last decade witnessed a surge in UAV-based and satellite-based video and image datasets. We list some open-source datasets, curated since 2016, and group their key features according to their downstream tasks.

VisDrone [91] is the most widely used drone dataset for aerial image object detection. It is recorded from 14 cities in China with various drone-mounted cameras, consists of 10 object categories, and is segregated into four task-specific sub-datasets: (a) Image Object Detection (10,209 images), (b) Video Object Detection (96 videos, 40,001 images), (c) Single-Object Tracking (139,276 images), and (d) Multi-Object Tracking (40,000 images). *Campus* [64], is the largest aerial dataset for multi-target tracking, activity comprehension, and trajectory prediction, focuses solely on the university campus, in contrast to our MAVREC. *UAVDT* [21] dataset consists of 80,000 frames and 3 subsets, focusing on single and multi-object detection and tracking, under different weather conditions, lighting, and altitude of the drone. *MOR-UAV* [52] is an aerial dataset consisting of 10,948 images, all annotated, designed for moving object detection under various challenges, such as illumination, camera movement, etc. *UAV123* [55] is a low-altitude aerial dataset consisting of 112,578 fully-annotated images across

[1]European vehicles are comprising of mainly three colors [5]; also, see B.2 for an analysis.

Table 1. State-of-the-art UAV-based datasets since 2016 in chronological order. For viewpoints, G denotes *ground-view*, A denotes *aerial-view*, and AG denotes both. Thermal IR datasets are not included.

Dataset	Total Frames	Resolution	Total Annotations	Instances per Annotated Frame	Categories	Viewpoints	Region	Year
Campus [64]	929,499	1400×2019	19,564	0.02	6	Single (A)	North America	2016
UAV123 [55]	110,000	720×720	110,000	1.0	6	Multi (A)	Middle East	2016
CarFusion[62]	53,000	$1,280 \times 720$	–	–	4	Multi	North America	2018
DAC-SDC [85]	150,000	640×360	NA	NA	12	Single	Asia	2018
UAVDT [21]	80,000	1080×540	841,500	10.52	3	Single	Asia	2018
MDOT [90]	259,793	–	–	–	9	Multi (A)	Asia	2019
VisDrone DET [91]	10,209	3840×2160	471,266	53.09	10	Single (A)	Asia	2019
VisDrone MOT [91]	40,000	3840×2160	1,527,557	45.83	10	Single (A)	Asia	2019
DOTA[82]	2806	4000×4000	188,282	67.09	15	Single (A)	Multiple	2019
DOTA V2.0 [20]	11,268	4000×4000	1,793,658	159.18	18	Single (A)	Multiple	2021
MOR-UAV [52]	10,948	1280×720, 1920×1080	89,783	8.20	2	Single	Asia	2020
AU-AIR [17]	32,823	1920×1080	132,034	4.02	8	Multi	Europe	2020
UAVid [50]	300	3840×2160	–	–	8	Single	Europe	2020
MOHR [87]	10,631	5472×3078, 7360×4192, 8688×5792	90,014	8.47	5	Multi (A)	Asia	2021
MAVREC (This paper)	537,030	2700×1520	1,102,604	50.01	10	**Multi (AG)**	Europe	2023

123 video sequences (simulated and recorded), designed for object tracking, with a subset intended for long-term aerial tracking. *MDOT* [90] is a *multi-drone based single object tracking dataset* with 259,793 frames across 155 groups of video clips, and 10 different annotated attributes. *Au-Air* [17] is a medium-scale, multi-sensor, aerial data designed for real-time object detection, with the aim of bridging the gap between computer vision and robotics. *DAC-SDC* [85] is a single-object detection dataset with 150,000 images collected from *DJI* [7] with 12 categories. *DOTA* [82] is an aerial dataset (2,806 images, 15 categories) for object detection in earth vision. *DOTA V2.0* [20], an upgraded version of DOTA, is a single view dataset collected from Google Earth, GF-2 satellite and aerial imagery (11,268 images, 18 categories) for object detection. The UG^2+ Challenges provide *A2I2-Haze* [57], the first real haze dataset, consisting of 229 pairs of hazy and clean images (197 training pairs, 32 testing pairs) from 12 videos, focusing on detection in visually degraded environments with smoke and haze with mutually exclusive aerial and ground images.

In an orthogonal line of work, *GRACO* [94] is a multimodal dataset for synchronized ground and aerial collaborative simultaneous localization and mapping (SLAM) algorithms (6 ground and 8 aerial sequences collected in China within a university campus) by a group of ground and aerial robots equipped with light detection and ranging (LiDAR), cameras, and global navigation satellite/inertial navigation systems (GNSS/INS) that capture images at 20Hz with 2182.52 and 2675.54 seconds duration in the ground and aerial, respectively. *S3E* [23] is a multimodal dataset for collaborative SLAM, consists of 7 outdoor and 5 indoor synchronized ground sequences, each longer than 200 seconds, and collected in 5 locations within a university campus in China. *DVCD18K* [12] is a cinematographic dataset (with corresponding camera paths) consisting of 18,551 edited drone clips spanning 44.3 hours.

Our proposed *MAVREC* is inherently different from the above datasets, because: (*a*) compared to other small-scale (e.g., UAVid, DOTA, MOR-UAV, AU-AIR), and low-resolution datasets (e.g., UAV123, UAVDT, DroneVehicle, BIRDSAI), MAVREC is the first-ever *large-scale*, unscripted, multiview video dataset (fourth largest among all UAV-based datasets ever after DVCD18K, Highway-drone, and Campus; Campus lacks object boundaries) recorded in *industry-standard* 2.7K resolution; (*b*) its multiview presents the same scenes through the lens of one or more ground cameras, and a medium altitude (flight height 25–45 meters, compared to low or high-altitude datasets, e.g., UAV123 with flight height 5–25 meters, MOHR with flight height 200 meters or above) drone-mounted camera (this perspective is unique compared to the existing multiview drone datasets, e.g., MDOT, UAV123, MVDTD, MCL) to have balanced variations of small and medium objects from both perspectives; (*c*) its high variance in object distribution across different scenes is complementary to datasets like VisDrone where object detection is relatively straightforward due to their biased object distribution (dense), reflecting its demographic characteristics; (*d*) *MAVREC is the first multiview, drone-based dataset curated in the wild, with a central focus on object detection*. GRACO and S3E are multiview but confined to university campuses and used for SLAM algorithms; S3E does not incorporate drone-based data acquisition. Alongside, A2I2-Haze dataset a tiny subgroup of [57] challenges, consists of mutually exclusive, non-synchronized aerial and ground images, while DVCD18K is an *human-edited cinematographic dataset with drone camera paths*, and vastly different from MAVREC in many aspects. §3.2 explains more unique challenges of MAVREC. There are UAV-

Figure 2. Different sample scenes (with annotation) from our dataset; the first row is the aerial view, and the second row presents the same scenes from a ground camera. See more sample frames in §B, Figure 9. For cropped-out, zoomed-in versions, see §B, Figure 10.

based datasets with downstream tasks primarily orthogonal to MAVREC. For completeness, we list some UAV-based datasets for action detection, counting, geo-localization, 3D reconstruction, and benchmarking in §A; also, see [81].

Object detection. CNN-based object detectors are divided into two categories: two-stage and one-stage. Two-stage detectors such as RCNN [27], Fast RCNN [26], Faster RCNN [63], employ a class-agnostic region proposal module followed by simultaneously regressing the object boundaries and their classes. In contrast, one-stage detectors like SSD [47], YoloV4 [74], YoloV6 [40], YoloV7 [75], YoloX [25], FCOS [70], directly predicts the image pixels as objects, leading to models that offer fast inference. Recently, by using neural architecture search, Yolo-NAS [11] outperforms previous Yolo models in real-time object detection. With the success of transformers, DETR [18] was the first transformer-based, end-to-end object detector. Following this, Deformable-DETR (D-DETR) [92] introduces a sparse attention module, computationally $6\times$ faster than DETR, and robust in detecting small objects. The majority of the aerial object detectors are based on the foundational principles established by the aforementioned popular object detectors [77, 83, 84]; TPH-YoloV5 [93] combines YoloV5 with a transformer prediction head to solve the varying object scales and motion blur for drone-captured scenarios. Our analysis utilizes the MAVREC dataset to benchmark some of these well-established methods, prioritizing factors such as fast inference, high precision, and the effective detection of small-scale objects.

3. MAVREC dataset

In this section, we start with the data acquisition process; and then explain annotation, statistical attributes, and unique challenges of MAVREC.

3.1. General setup

Recording set-up. We conduct the recording in public spaces in compliance with the European Union's drone safety and Scandinavian video surveillance regulations; see §D for a detailed discussion on reproducibility, licensing, privacy, safety, maintenance plan, and broader impact of the dataset. We record our dual-view aerial-ground dataset with a drone-mounted camera (DJI Phantom 4, DJI mini 2) and a consumer-grade static ground camera (GoPro Hero 4, GoPro Hero 6, iPhone 11 and 13-Pro) placed on a tripod; see details in Table 5. The drone is kept semi-static, hovering approximately 25–45 meters above the ground; see the relative positions and viewing angles of the drone and the ground camera in Figure 3. Based on that, we identify three recording scenarios (P1, P2, and P3). In P3, we better capture the objects as the drone gets a wider viewing angle. However, we keep all views *not to amplify biases* from any particular view. For some recordings in the city center, railroad, or crowded intersections, we could not operate a drone due to the UAV flight regulations; hence, we used a user-grade handheld camera set-up on the balcony of a high-riser to capture aerial views.

Recording locations and scenes. To avoid locational bias, we collected our data in 11 different geographical locations (European outdoors, rural, and urban) with mixed pastures, in spring and summer (with the sun hitting the cameras from different angels), and when there is an encyclopedic spectrum of green and yellow intertwined in the background; see Figures 3 and 2 (also, see B.2 for an analysis). We choose the parking lots, and busy traffic intersections in the city, during the peak traffic hours to create more nuanced and complex interactions, in which multiple foreground objects interact and create enormous visual challenges. Alongside, we choose harbor, single-lane roads in the countryside, asphalt roads, and bicycle lanes, in moderate traffic conditions, to collect simple scenarios that might have sparse to dense foreground objects (see sample frames in Figure 2).

Alignment of dual-views. Human operators simultaneously record the scenes from dual views; although a minute time-lapse is unavoidable. After recording, clips are loaded into the `QuickTime` player, and a human operator manually synchronizes the frames to alleviate the time lapses.

Annotation and categories. MAVREC, as highlighted previously, stands as one of the largest drone datasets, encompassing millions of objects within its distribution. However, annotating each object is a resource-intensive task. Inspired by the recent success of the semi-supervised learning paradigm in the computer vision community [19, 39, 48, 54, 76, 78], we

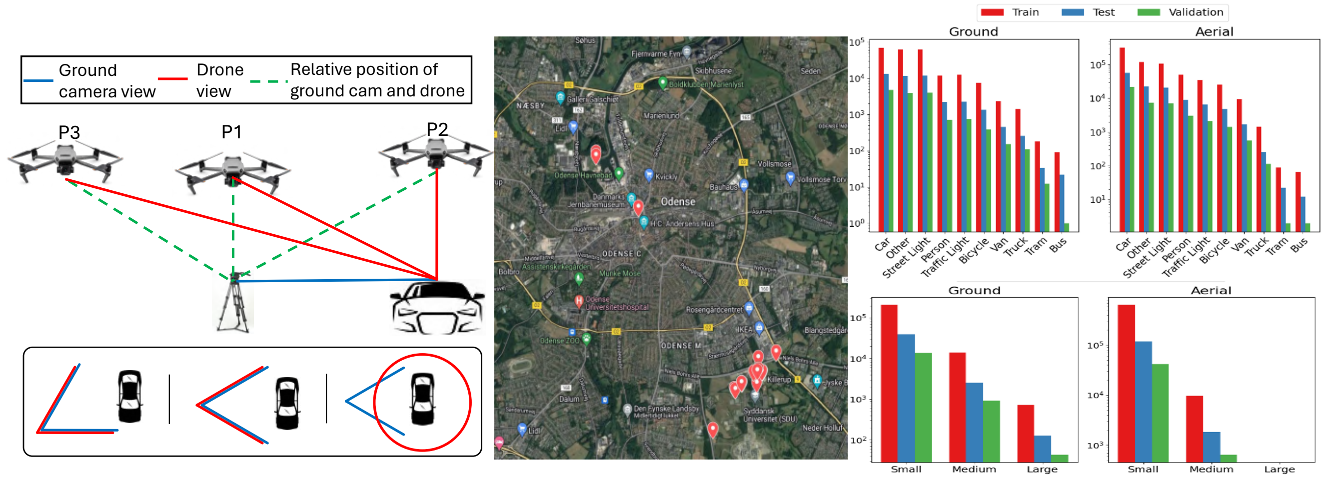

Figure 3. **Left**: Recording instances are divided into three different scenarios (P1, P2, and P3) based on the relative positions and the field-of-view (FOV) of the ground camera and the drone. The drone operates directly on top of the object (P2), and two oblique views—directly on top of the ground camera (P1), and behind the ground camera (P3). **Middle**: Recording locations as red dropped pins on the Google map's satellite view. **Right (top)**: Total numbers of objects in each category in the ground and aerial view. **Right (bottom)**: Number of large, medium, and small objects in the test, train, and validation sets of two views; aerial view has no *large* object annotation.

split the videos from different views into two categories; an annotated set and an unannotated set. After pre-processing, we select the first 30 seconds of the synchronized videos and annotate the frames through a semi-automatic, open-source annotation platform by Intel, called `CVAT` [2], and leave the rest of the video unannotated; see `CVAT` interface in Figure 11. We provide an annotation interface with 10 categories in `CVAT`: `tram`, `truck`, `bus`, `van`, `car`, `bicycle`, `person`, `street light`, `traffic light`, and `other`. In `other` category, we annotated objects that share visual similarities with objects from the remaining categories; e.g., blocks of concrete from aerial view might look like cars, or white divider and marker posts from aerial view might look like a person with a white T-shirt, and so on. We created this category for the models to learn to disambiguate the *look-alike* objects from different categories. The in-built tracker in `CVAT` tracks an object through multiple frames. We annotated by skipping forward 10 frames; thus speeding up the annotation process. Nevertheless, to ensure high annotation quality, a human annotator reviews each frame, and 5 human non-annotators have reviewed the dataset annotations. The annotation reviewers check that the bounding box encapsulates the object or its parts, has minimal overlap with other objects, and that all instances of the class in the frame are labeled. This process is performed on 600 annotated images randomly sampled from the annotated data. During the review process, we find that the error is around 6% which is comparable to the benchmark datasets in this domain. Similar to other benchmarks [44, 91], annotated frames are assembled into `COCO-json` format to give a unique identifier for each object class.

3.2. Structuring, statistics, and challenges

This section discusses the size, statistical properties, and challenges of the training distribution of MAVREC. We focus on two key points: (*i*) distribution of different categories, and (*ii*) distribution of the annotated object size.

Structuring the dataset. We divided the annotated data from both views into three subsets—train, validation, and test sets. Following [17, 55], to ensure the distributions of the different objects are approximately the same throughout these three sets, we split each video sequence into three fragments, and then randomly select samples for each set.

Distribution of different categories. We show the distribution of categories from both views in Figure 3; also, see Figure 12. MAVREC contains over 1.1 million bounding box annotations in both views combined, rendering ~ 50.01 annotations per frame; see Figure 3 and details in Table 6. The distribution is *long-tailed* where cars are more frequent than trams and buses. The slight inconsistency in the object distribution from both views is natural as some recordings were conducted with the P3 setup, and in this setup, the drone has a wider viewing angle than the ground camera.

Object size distribution. To better illustrate the challenges in MAVREC, we divide the object sizes present in the videos into *three* categories: small ($< 32 \times 32$ pixels), medium (lies inclusively between 32×32 and 96×96 pixels), and large ($> 96 \times 96$ pixels). Figure 3 (also, see Figure 13) presents the number of annotated object sizes in both views. Large objects, such as trams, buses, and trucks, are present in fewer frames compared to the other objects. Also, the drone is maneuvered at a higher altitude, and the aerial view has a higher percentage of small objects compared to the ground view, creating a natural bias in object sizes. We also observe that the distribution for the split into the train, validation, and test set has almost an equal distribution of the different object sizes for both views; see Figure 13-(c) for distribution for the object sizes.

Unique properties of MAVREC. MAVREC contains typical outdoor activities characterized by real-world properties like long-tail distribution, objects with similar appearance, viewpoint changes, varying illumination, etc. Additionally, MAVREC exhibits some unique properties, not found in other datasets: (*i*) Ground view contains occluded objects; these objects can be recovered due to the wide aerial FOV. *Dual-view feature of the MAVREC* has the potential to offer a wide range of solutions for scenes with occlusion, which remains a significant challenge in video surveillance. (*ii*) *MAVREC's color distribution* reflects European demographics, which may influence object detection algorithms that incorporate scene-contextual information, particularly those pre-trained on general object detection datasets; see a comparison in Figure 14. (*iii*) Historically, vehicle color distributions vary across Europe, North America, and the Asia-Pacific; see Figure 8. The existing datasets collected in Asia and North America appear to be more colorful. E.g., in 2021, Europe's top car colors were gray (27%), white (23%), and black (22%), contrasting with North America's gray (21%), black (20%), blue (10%), red (10-11%), and silver (10%), and China's predominance of white (50%) and brown (10%) cars [4]. (*iv*) MAVREC was collected at *high latitudes*. The elevation of the sun in these areas (see Figure 7) during the peak traffic times is high, creating a *mirage-like* reflection on one of the sensors in many scenes, thereby, causing significant disparities between the two views. The second column of rows 1 and 2 in Figure 2 shows this effect. (*v*) The aerial perspective inherent in MAVREC leads to *small objects* inclusion; their presence is susceptible to misdetection by detection algorithms. (*vi*) MAVREC is characterized by both *sparse* and *dense* distribution of objects. Our empirical findings suggest that such a large disparity in object distribution presents challenges in training object detectors, compared to scenes exhibiting only uniformly dense annotations.

4. Baselines and evaluation

This section presents the benchmarking results on MAVREC in supervised and semi-supervised settings. We also present our observations concerning the prevailing trends in object detectors employed on aerial images.

Datasets and evaluation metric. For supervised and semi-supervised benchmarking with MAVREC, we use a total of 8,605 labeled frames, and at most 8,605 unlabeled frames from each view at training. The validation and test set (extracted from disjoint video sequences) for each view contains 805 and 1,614 annotated images, respectively. We evaluate the models with the widely used metric for object detection, mean average precision (mAP) [44]; see a detailed discussion in §C.2.

Object detector baselines. For *supervised benchmarking*, we use CNN-based YoloV7 [75], and transformer-based DETR [18] and D-DETR [92]. Additionally, we use Yolo-NAS [11]. For *semi-supervised benchmarking*, we propose a curriculum based semi-supervised baseline using D-DETR. We provide the implementation details and computing environment in the §C.1; we refer to Tables 7 and 8 for other model-specific implementation details.

4.1. Supervised benchmarking

Table 2 presents the supervised baseline results on the MAVREC dataset for both ground and aerial perspectives. Despite an equal number of training samples from different views, we observe that all the baselines exhibit superior performance on the ground perspective compared to the aerial perspective. This discrepancy highlights the challenge associated with object detection in aerial views due to their smaller sizes, as indicated by the $\mathtt{AP_S}$ metric. Notably, YoloV7 demonstrates the best performance on aerial images, while D-DETR pre-trained on MSCOCO surpasses other models on the ground view. Interestingly, Yolo-NAS, which surpasses other Yolo-based detectors on ground images according to [11], exhibits lower performance than YoloV7 on aerial images indicating that the learned Yolo-NAS architecture is suboptimal for aerial images. We show some qualitative results in Figure 17.

4.2. Curriculum learning based semi-supervised object detection

Can ground view images improve object detection in aerial perspective? To answer this, we trained D-DETR and YoloV7 by augmenting the existing aerial view sample set with ground view samples. We achieve this by simply adding the two sets of aerial- and ground view samples along with their corresponding annotations. Our findings demonstrate that including ground view samples substantially improves object detection performance.

Graphs presented in Figure 4 illustrate that D-DETR outperforms the CNN-based YoloV7 when the extra ground view samples enrich the training distribution. While YoloV7 requires an equal number of ground view samples as aerial samples to achieve its peak performance, D-DETR achieves a relative improvement up to 270% with a subset of ground view samples ($\sim$ 2K ground view images). Interestingly, further augmentation of ground view images during D-DETR training does not enhance its performance, indicating the sensitivity of D-DETR's training process to ground view image sampling.

Thus, experimental analysis of MAVREC within a supervised framework suggests that (*i*) CNN-based models, such as Yolo, demonstrated superior performance over transformer-based models, like D-DETR, when trained from scratch. However, the transformer-based model outperforms when pretrained on large-scale ground images in MSCOCO. This shows the superiority of these architectures when large-scale data is available. (*ii*) These transformer-based object detectors augmented with even 25% of MAVREC's ground view images, surpass the model pretrained on MSCOCO. This shows the importance of learning geography-aware representation for aerial visual perception, suggesting a new direction for enhancing object detection in aerial perspective.

Model generalization. In Table 3, we provide the object

Table 2. Supervised benchmark of MAVREC. D-DETR* denotes a MSCOCO pre-trained D-DETR.

Trained DNN Models	**Validation Set**								**Test Set**							
	Ground				**Aerial**				**Ground**				**Aerial**			
	AP	AP_{50}	AP_S	AP_M	AP	AP_{50}	AP_S	AP_M	AP	AP_{50}	AP_S	AP_M	AP	AP_{50}	AP_S	AP_M
DETR	21.8	36.9	21.9	23.9	24.9	39.7	27.6	45.3	20.8	35.4	21.3	24.0	23.6	40.1	23.4	44.9
D-DETR	27.5	51.4	28.1	43.7	13.1	28.3	14.2	38.1	18.2	46.8	17.9	36.0	10.3	25.0	10.1	29.4
D-DETR*	**59.6**	**82.7**	**59.7**	**79.6**	31.0	**61.7**	31.7	55.1	**58.6**	**81.4**	**59.0**	**80.2**	**33.2**	**61.9**	**31.5**	51.0
Yolo-NAS (L)	41.4	61.7	36.8	72.9	30.3	49.8	29.2	**61.5**	41.2	63.4	37.8	74.3	27.0	43.3	25.9	58.0
YoloV7	45.6	72.1	40.6	74.9	**31.3**	57.7	**34.2**	61.2	45.0	72.5	42.4	74.4	31.9	58.8	31.4	**63.1**

Training Protocol	**Pre-training**	**Test Set**			
		AP	AP_{50}	AP_S	AP_M
Trained from scratch	×	10.3	25.0	10.1	29.4
Grounding-DINO	O365, GoldG, Cap4M	20.4	40.9	18.6	32.5
FT on MAVREC Aerial view	Visdrone [91]	20.9	41.9	20.6	43.8
FT on MAVREC Aerial view	MS-COCO [44]	33.2	61.9	31.5	51.0
FT on MAVREC Aerial view	MAVREC Ground view	**44.8**	**71.5**	**42.9**	**72.4**

Table 3. Object detection using D-DETR [92] on the aerial view images of the MAVREC dataset; FT indicates finetuning.

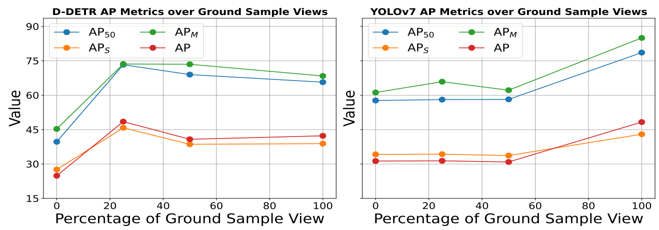

Figure 4. Supervised benchmark on the aerial view of MAVREC (Validation Set).

detection performance of MAVREC with models pretrained with different strategies. Our empirical evaluations indicate that state-of-the-art object detection models, including the *open-world foundational models* like Grounding-Dino [46], which fail to achieve the expected performance level on MAVREC; see Figure 5. This observation validates an inherent bias of these models towards ground view data. Moreover, Table 3, combined with the visual evidence in Figure 5, shows an object detector pre-trained on popular ground view dataset (MS-COCO [44]) or other aerial datasets collected from different geographies (e.g., VisDrone [91] from China) has diminished efficacy on aerial images obtained from disparate geographical regions (for our case, Europe). Therefore, unlike classical object detection, training a sophisticated DNN model on a large dataset (e.g., ImageNet [65] or MS-COCO [44]) does not offer the best overall solution. We find augmenting object detectors with ground view images from the corresponding geographical context is a superior strategy that boosts detection performance.

In this section, we introduce a curriculum-based learning strategy for semi-supervised object detection. Curriculum learning [67] provides a systematic strategy to enhance model performance by incrementally introducing complexity into the training regime. We adopt curriculum learning in semi-supervised object detection using a D-DETR. In our semi-supervised baseline, we train the object detector using both labeled and unlabeled image sets. The foundation of our semi-supervised baseline is a consistency regularization framework based on a `teacher-student` model [76]. This framework encompasses two distinct training phases: (*i*) the 'burn-in' stage, where the `teacher` model is trained exclusively with labeled images, and (*ii*) the semi-supervised training stage, where the `teacher-student` model engages with unlabeled images. Given a trained `teacher` network from the burn-in stage, this framework leverages weak-to-strong consistency regularization [22] to leverage unlabeled aerial images as shown in Figure 6. In this stage, the `teacher` network processes the weakly augmented unlabeled aerial view images to generate bounding boxes. These bounding boxes serve as pseudo-labels for the strongly augmented counterpart of the image, which is the input of the `student`. The `teacher` network is updated through an exponential moving average (EMA) of the `student`'s updates, and the `student` network only gets back-propagated.

However, the performance of this semi-supervised baseline partially relies on the effectiveness of the `teacher` network to generate pseudo-labels. Inspired by our insights in transferring geography-aware knowledge from the ground view to the aerial view, we employ a curriculum learning strategy in the burn-in stage of the semi-supervised framework. Thus, the `teacher` network is initially trained on simpler examples, followed by complex examples. The pseudo-labels generated by the `teacher` on the more challenging aerial samples are deemed more reliable than those from a baseline model trained randomly on aerial and ground samples. The ordering criteria are determined by a difficulty scoring function, which is the average confidence level of all generated proposals for a given input image. Subsequently, our findings reveal that the effective approach involves training the `teacher` network initially with m labeled ground view images, followed by n labeled aerial images. The outcome is a trained geography-aware `teacher` network that facilitates the second phase of training the semi-supervised framework by generating precise object proposals. We observe further improvements when a similar sequential training strategy, first with unlabeled ground images and then with unlabeled aerial images, is applied to the `student` network.

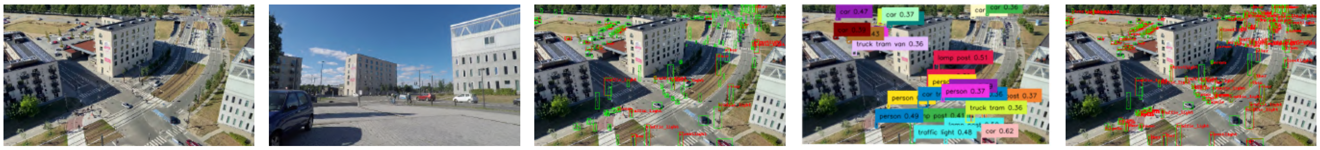

Figure 5. **Left to right**: Sample frames from time-synchronized MAVREC showing aerial and ground views; D-DETR [92] trained on aerial VisDrone DET [91] inference results on MAVREC (GT bounding boxes are green, and detection are in red); Grouding-Dino [46] inference result on MAVERC; inference results of D-DETR trained on aerial MAVREC has fewer missed detections (Zoom in for a better view).

Table 4. Semi-supervised Omni-DETR [76] benchmark on MAVREC. In the table, G and A denote the number of ground and aerial view images, respectively. During the burn-in, we only use the labeled subset.

Training Technique	**Labelled** $G(m)$	**Labelled** $A(n)$	**Unlabelled** G	**Unlabelled** A	**Test perspective**	**Validation Set** AP	**Validation Set** AP_{50}	**Validation Set** AP_{S}	**Validation Set** AP_{M}	**Test Set** AP	**Test Set** AP_{50}	**Test Set** AP_{S}	**Test Set** AP_{M}
Semi-Supervised	8605	0	8605	0	G	56.9	83.3	54.8	**74.9**	45.8	75.5	45.4	58.7
Semi-Supervised	0	8605	0	8605	A	29.3	49.3	24.8	60.6	19.8	38.4	19.5	35.0
Curriculum Learning (Ours)	2151	8605	0	8605	A	37.8	**64.9**	35.4	68.3	23.2	45.4	21.8	43.6
Curriculum Learning (Ours)	2151	8605	2151	8605	A	**38.0**	64.8	**35.7**	67.6	**26.7**	**54.1**	**24.5**	42.4

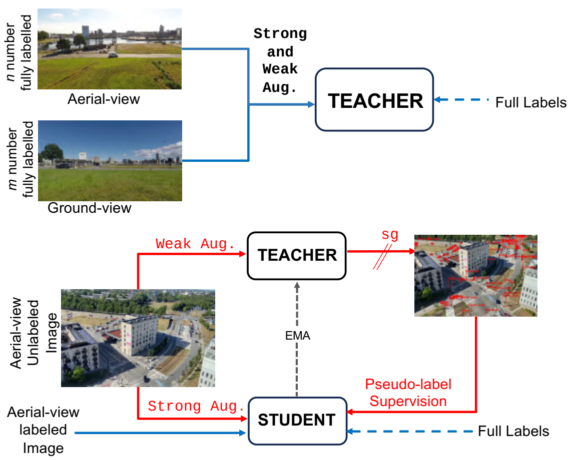

Figure 6. Semi-Supervised object detection framework based on curriculum learning approach. Here blue represents the initial supervised stage, red represents the later unsupervised stage, SG represents stop gradient.

In Table 4, we showcase the results of semi-supervised benchmarking on the MAVREC dataset, utilizing D-DETR as the backbone object detection model within the teacher-student framework. Note that all the models are trained on 39 epochs with 20 epochs for the burn-in stage and 19 epochs for the semi-supervised training stage. Our semi-supervised baseline results demonstrate that by utilizing the same number of unlabeled aerial images as labeled images, we achieve a substantial boost in object detection performance—from 13.1% to 29.3% and 10.3% to 19.8% on validation and test set, respectively. We observe a consistent improvement in the ground view. This warrants the importance of using semi-supervised approaches for vision tasks, particularly where the annotation process is labor-intensive. Subsequently, we demonstrate the performance of our curriculum-based semi-supervised object detector on the MAVREC dataset. This approach outperforms the baseline model by 8.5% and 7% on validation and test sets, respectively. Moreover, we observed additional improvements in object detection accuracy when the second training phase was augmented with unlabelled ground view images. This shows the importance of learning geography-aware representation from the ground view for enhancing aerial visual perception.

5. Conclusion

In this paper, we introduce a large-scale, high-definition ground and aerial-view video dataset, MAVREC. To the best of our knowledge, MAVREC is the first drone-based aerial object detection dataset that exploits the multi-view of the data coming from orthogonal views, aerial and ground to offer enhanced detection capacity for an aerial view. In our extensive benchmarking on MAVREC, we employed both supervised and semi-supervised learning methods, along with our proposed curriculum-based ground-view pre-training strategy. Our findings highlight several key insights, particularly the importance of geographic awareness in aerial visual models. We discovered that models trained on aerial images from one geographic location often struggle to generalize to different regions. However, integrating ground-view data from the same geographic area significantly enhances the model's ability to learn more distinctive visual representations. We envision that this dataset and benchmarking will benefit: (*i*) researchers, who will use it as the basis for consistent implementation and evaluation; and (*ii*) practitioners, who need an appropriate, large-scale, industry-standard dataset for training DNN models for aerial images. We also recognize that greater data diversity is essential and in the future, we plan to diversify MAVREC by adding more domains and geographies.

Acknowledgements. We thank Jonathan Lorray, Thomas Lindal Winther, and Peter Schneider-Kamp for their support during different stages of this project. This work is supported in part by the National Science Foundation (IIS-2245652).

References

[1] Color difference. https://en.wikipedia.org/wiki/Color_difference. 15

[2] CVAT annotation tool. https://www.cvat.ai. 5, 14, 17

[3] PNNL Parking Lot 1 and 2 and Pizza sequences. https://www.crcv.ucf.edu/data/ParkingLOT/. 14

[4] The Most Popular Car Color: Can You Guess Which One?, . https://www.motorbiscuit.com/most-popular-car-color-guess-color/. 6

[5] Most popular car-colors by country, . https://haynes.com/en-us/tips-tutorials/most-popular-car-colors-country-or-don-t-buy-black-car-india. 2

[6] Innovation built through partnerships to improve life on the streetscape for all. https://cs3-erc.org/. 2, 22

[7] DJI. https://www.dji.com. 3

[8] Mapped: The World's Population Density by Latitude. https://www.visualcapitalist.com/cp/mapped-the-worlds-population-density-by-latitude/. 2, 14

[9] Car colour popularity. https://en.wikipedia.org/wiki/Car_colour_popularity, . 15

[10] Solar zenith-angle, . https://en.wikipedia.org/wiki/Solar_zenith_angle. 15

[11] Yolo-NAS. https://github.com/Deci-AI/super-gradients/blob/master/YOLONAS.md. 4, 6, 20, 21

[12] Amirsaman Ashtari, Raehyuk Jung, Mingxiao Li, and Junyong Noh. A drone video clip dataset and its applications in automated cinematography. In *Computer Graphics Forum*, pages 189–203, 2022. 3

[13] Naeem Ayoub and Peter Schneider-Kamp. Real-time on-board detection of components and faults in an autonomous uav system for power line inspection. In *Proceedings of the International Conference on Deep Learning Theory and Applications*, pages 68–75, 2020. 2, 22

[14] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the Conference on computer vision and pattern recognition workshops*, pages 28–35, 2017. 14

[15] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. 1, 22

[16] Elizabeth Bondi, Raghav Jain, Palash Aggrawal, Saket Anand, Robert Hannaford, Ashish Kapoor, Jim Piavis, Shital Shah, Lucas Joppa, Bistra Dilkina, et al. BIRDSAI: A dataset for detection and tracking in aerial thermal infrared videos. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1747–1756, 2020. 2, 14

[17] Ilker Bozcan and Erdal Kayacan. Au-air: A multimodal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *IEEE International Conference on Robotics and Automation*, pages 8504–8510, 2020. 2, 3, 5

[18] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 213–229, 2020. 4, 6, 17, 20, 21

[19] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 4815–4824, 2022. 4

[20] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7778–7796, 2021. 3

[21] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision*, pages 370–386, 2018. 2, 3, 19

[22] Yue Fan, Anna Kukleva, Dengxin Dai, and Bernt Schiele. Revisiting consistency regularization for semi-supervised learning. *International Journal of Computer Vision*, 131(3):626–643, 2023. 7

[23] Dapeng Feng, Yuhua Qi, Shipeng Zhong, Zhiqiang Chen, Yudu Jiao, Qiming Chen, Tao Jiang, and Hongbo Chen. S3e: A large-scale multimodal dataset for collaborative slam. *arXiv preprint arXiv:2210.13723*, 2022. 3

[24] Jan Gasienica-Jozkowy, Mateusz Knapik, and Bogusław Cyganek. An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance. *Integrated Computer-Aided Engineering*, 28(3):221–235, 2021. 14

[25] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 4

[26] Ross Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision*, pages 1440–1448, 2015. 4

[27] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 4

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 21

[29] Yu Hongyang, Guorong Li, Weigang Zhang, Qingming Huang, Dawei Du, Tian Qi, and Sebe Nicu. The unmanned aerial vehicle benchmark: Object detection, tracking and baseline. *International Journal of Computer Vision*, 128(5):1141–1159, 2020. 14

[30] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the International Conference on Computer Vision*, pages 4145–4153, 2017. 14

[31] Zubayer Islam, Mohamed Abdel-Aty, Amrita Goswamy, Amr Abdelraouf, and Ou Zheng. Effect of signal timing on vehicles' near misses at intersections. *Scientific reports*, 131:9065, 2023. 2, 22

[32] Efstratios Kakaletsis, Charalampos Symeonidis, Maria Tzelepi, Ioannis Mademlis, Anastasios Tefas, Nikos Nikolaidis, and Ioannis Pitas. Computer vision for autonomous UAV flight safety: an overview and a vision-based safe landing pipeline example. *ACM Computing Surveys*, 54(9):1–37, 2021. 2

[33] Isha Kalra, Maneet Singh, Shruti Nagpal, Richa Singh, Mayank Vatsa, and P. B. Sujit. DroneSURF: Benchmark Dataset for Drone-based Face Recognition. In *proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 1–7, 2019. 14

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015. 20

[35] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125, 2018. 14

[36] S.V. Aruna Kumar, Ehsan Yaghoubi, Abhijit Das, B.S. Harish, and Hugo Proença. The P-DESTRE: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security*, 16:1696–1708, 2020. 14

[37] Matti Kummu and Olli Varis. The world by latitudes: A global analysis of human population, development level and environment across the north–south axis over the past half century. *Applied geography*, 31(2):495–507, 2011. 2, 14

[38] Matti Kummu, Hans De Moel, Gianluigi Salvucci, Daniel Viviroli, Philip J Ward, and Olli Varis. Over the hills and further away from coast: global geospatial patterns of human and environment over the 20th–21st centuries. *Environmental Research Letters*, 11(3): 034010, 2016. 2

[39] Aoxue Li, Peng Yuan, and Zhenguo Li. Semi-supervised object detection via multi-instance alignment with global class prototypes. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 9809–9818, 2022. 4

[40] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022. 4

[41] Jing Li, Dong Hye Ye, Timothy Chung, Mathias Kolsch, Juan Wachs, and Charles Bouman. Multi-target detection and tracking from a single camera in unmanned aerial vehicles (UAVs). In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 4992–4997, 2016. 14

[42] Jingtong Li, Jesse Murray, Dorina Ismaili, Konrad Schindler, and Cenek Albl. Reconstruction of 3D flight trajectories from ad-hoc camera networks. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 1621–1628, 2020. 14

[43] Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 14

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014. 1, 5, 6, 7, 20

[45] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 5007–5015, 2015. 14

[46] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 7, 8

[47] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander Berg. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, pages 21–37, 2016. 4

[48] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for

anchor-free and anchor-based detectors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022. 4

[49] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods*, pages 324–331, 2017. 2, 14

[50] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. UAVid: A semantic segmentation dataset for uav imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020. 2, 3

[51] András L Majdik, Damiano Verda, Yves Albers-Schoenberg, and Davide Scaramuzza. Air-ground matching: Appearance-based GPS-denied urban localization of micro aerial vehicles. *Journal of Field Robotics*, 32(7):1015–1039, 2015. 14

[52] Murari Mandal, Lav Kush Kumar, and Santosh Kumar Vipparthi. Mor-UAV: A benchmark dataset and baselines for moving object recognition in UAV videos. In *Proceedings of ACM International Conference on Multimedia*, pages 2626–2635, 2020. 2, 3

[53] Aboli Marathe, Pushkar Jain, Rahee Walambe, and Ketan Kotecha. RestoreX-AI: A contrastive approach towards guiding image restoration via explainable AI systems. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, pages 3030–3039, 2022. 1, 22

[54] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 14482–14491, 2022. 4

[55] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for UAV tracking. In *Proceedings of the European Conference on Computer Vision*, pages 445–461, 2016. 2, 3, 5, 17, 19, 20

[56] Arjun Nagendran, Don Harper, and Mubarak Shah. New system performs persistent wide-area aerial surveillance. *SPIE Newsroom*, 5:20–28, 2010. 14

[57] Priya Narayanan, Xin Hu, Zhenyu Wu, Matthew D. Thielke, John G. Rogers, Andre V Harrison, John A. D'Agostino, James D Brown, Long P. Quang, James R. Uplinger, Heesung Kwon, and Zhangyang Wang. A multi-purpose realistic haze benchmark with quantifiable haze levels and ground truth. *IEEE Transactions on Image Processing*, 32:3481–3492, 2023. 3

[58] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science and Business Media, 2003. 20

[59] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 3153–3160, 2011. 14

[60] Yaoru Pan, Mogens Flindt, Peter Schneider-Kamp, and Marianne Holmer. Beach wrack mapping using unmanned aerial vehicles for coastal environmental management. *Ocean and Coastal Management*, 213, 2021. 2, 22

[61] Anne-Flore Perrin, Vassilios Krassanakis, Lu Zhang, Vincent Ricordel, Matthieu Perreira Da Silva, and Olivier Le Meur. EyetrackUAV2: A large-scale binocular eye-tracking dataset for UAV videos. *Drones*, 4 (1):2, 2020. 14

[62] N Dinesh Reddy, Minh Vo, and Srinivasa G Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3D reconstruction of vehicles. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1906–1915, 2018. 3, 14

[63] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Proceedings of Advances in Neural Information Processing Systems*, 28, 2015. 4

[64] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proceedings of the European Conference on Computer Vision*, pages 549–565, 2016. 2, 3, 17, 19, 20

[65] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 7

[66] Qi Shan, Changchang Wu, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M Seitz. Accurate geo-registration by ground-to-aerial image matching. In *Proceedings of International Conference on 3D Vision*, pages 525–532, 2014. 14

[67] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022. 7

[68] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 14

[69] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. 2, 14

[70] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1922–1933, 2020. 4

[71] Andrea Vallone, Frederik Warburg, Hans Hansen, Søren Hauberg, and Javier Civera. Danish airs and grounds: A dataset for aerial-to-street-level place recognition and localization. *IEEE Robotics and Automation Letters*, 7(4):9207–9214, 2022. 14

[72] Leon Amadeus Varga, Benjamin Kiefer, Martin Messmer, and Andreas Zell. Seadronessee: A maritime benchmark for detecting humans in open water. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 2260–2270, 2022. 14

[73] Chuanyun Wang, Yang Su, Jingjing Wang, Tian Wang, and Qian Gao. UAVSwarm dataset: An unmanned aerial vehicle swarm dataset for multiple object tracking. *Remote Sensing*, 14(11), 2022. 14

[74] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 13029–13038, 2021. 4

[75] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YoloV7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023. 4, 6, 20, 21

[76] Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. Omni-DETR: Omni-supervised object detection with transformers. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 9367–9376, 2022. 4, 7, 8

[77] Xin Wang, Ning He, Chen Hong, Qi Wang, and Ming Chen. Improved YOLOX-X based UAV aerial photography object detection algorithm. *Image and Vision Computing*, 135:104697, 2023. 4

[78] Xinjiang Wang, Xingyi Yang, Shilong Zhang, Yijiang Li, Litong Feng, Shijie Fang, Chengqi Lyu, Kai Chen, and Wayne Zhang. Consistent-Teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 3240–3249, 2023. 4

[79] Jan D Wegner, Steven Branson, David Hall, Konrad Schindler, and Pietro Perona. Cataloging public objects using aerial and street-level images-urban trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6014–6023, 2016. 14

[80] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 7808–7817, 2021. 14

[81] Xin Wu, Wei Li, Danfeng Hong, Ran Tao, and Qian Du. Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 10(1):91–124, 2022. 1, 4, 14, 22

[82] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 3

[83] Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:79–93, 2022. 4

[84] Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. RFLA: Gaussian receptive field based label assignment for tiny object detection. In *Proceedings of the European Conference on Computer Vision*, pages 526–543, 2022. 4

[85] Xiaowei Xu, Xinyi Zhang, Bei Yu, Xiaobo Sharon Hu, Christopher Rowen, Jingtong Hu, and Yiyu Shi. DAC-SDC low power object detection challenge for UAV applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):392–403, 2021. 3

[86] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. In *Proceedings of the International Conference on Learning Representations*, 2018. 20, 21

[87] Haijun Zhang, Mingshan Sun, Qun Li, Linlin Liu, Ming Liu, and Yuzhu Ji. An empirical study of multi-scale object detection in high resolution UAV images. *Neurocomputing*, 421:173–182, 2021. 3

[88] Wang Zhang, Chunsheng Liu, Faliang Chang, and Ye Song. Multi-scale and occlusion aware network for vehicle detection and segmentation on uav aerial images. *Remote Sensing*, 12(11):1760, 2020. 14

[89] Ou Zheng, Mohamed Abdel-Aty, Lishengsa Yue, Amr Abdelraouf, Zijin Wang, and Nada Mahmoud. CitySim: A drone-based vehicle trajectory dataset for safety oriented research and digital twins. *arXiv preprint arXiv:2208.11036*, 2022. 2, 22

[90] Pengfei Zhu, Jiayu Zheng, Dawei Du, Longyin Wen, Yiming Sun, and Qinghua Hu. Multi-drone-based single object tracking with agent sharing network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):4058–4070, 2020. 3

[91] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection

and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7380–7399, 2022. 1, 2, 3, 5, 7, 8, 16, 17, 19, 20

[92] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proceedings of the International Conference on Learning Representations*, 2020. 4, 6, 7, 8, 17, 20, 21

[93] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the International Conference on Computer Vision*, pages 2778–2788, 2021. 4

[94] Yilin Zhu, Yang Kong, Yingrui Jie, Shiyou Xu, and Hui Cheng. Graco: A multimodal dataset for ground and aerial cooperative localization and mapping. *IEEE Robotics and Automation Letters*, 2023. 3

[95] Lojze Žust, Janez Perš, and Matej Kristan. Lars: A diverse panoptic maritime obstacle detection dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20304–20314, 2023. 14