Dual-View Drone Dataset: Can Multi-view Improve Aerial Visual Perception?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Object detection from aerial images are essential in many applications, such as self-driving cars, robotics, surveillance, to mention a few. Despite commercial abundance of unmanned aerial vehicles (UAVs), aerial data acquisition remains challenging due to UAV-flight regulations, atmospheric turbulence, sensor failure, etc. Additionally, the color content of the scenes, solar-zenith angle, and population density of different geographies are latitude-dependent, and influence the diversity of the collected data and generalization capacity of the deep neural network (DNN) models trained on it. The existing drone datasets collected primarily in the urban pasture of Asia and North America, do not exhaustively represent different geographies.

In this work, we present **D**ual **V**iew **D**rone or **DVD**, a video dataset where synchronized scenes are recorded from different perspectives — ground camera and drone-mounted camera. DVD 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, covering rural and urban pastures during spring and summer in high-latitude European geographies. This makes DVD the largest ground and aerial-view dataset, and the second largest among all drone-based datasets across all modalities and tasks. Additionally, we recognize that annotating raw videos is a resource-heavy task. Therefore, while benchmarking, we focus on both supervised and semi-supervised object detection baselines that can distill knowledge from multiple views and learn collaboratively. We release the DVD dataset for research use[1].

1 Introduction

Object detection and tracking employing aerial videos captured from UAVs or drones are essential in many downstream applications, such as autonomous driving [44, 14], robotics [68], environmental monitoring [50], infrastructure inspection [12], developing livable and safe communities [26, 5, 76], a few to name. The presence of small-scale objects in a vast background, occlusions, complex backgrounds, and variations in lighting and shadows present formidable challenges in this task. While computer vision has been rapidly advancing with the advent of modern DNN models that are data-hungry, the same does not hold for data representing the aerial perspective. Aerial data collection is complicated due to UAV-flight regulations and safety protocol, atmospheric turbulence, and many more [27]. Furthermore, if the camera or the sensor in the UAV is erroneous, the data

[1] https://dvd-dataset.github.io/dvd.github.io/

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

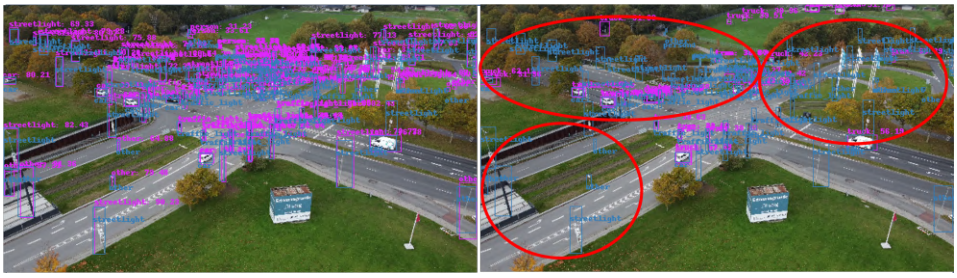


Figure 1: Inference results of D-DETR (900 Queries, 16 attention heads) on DVD; GT bounding boxes are blue, and detection are in purple. D-DETR trained on aerial DVD has much robust detection (left) compared to D-DETR trained on aerial VisDrone DET; it has fewer and failed detection (red circles) on DVD.

collection may suffer. Hence, the availability of high-quality, large-scale, and diverse aerial datasets is limited. The existing open-source UAV datasets [77, 41, 43, 16, 46, 19, 57, 15] are either small-scale, or low-resolution, and collected primarily in the urban pasture across different geographies (primarily in Asia and North America). Many studies reveal latitude influences the population density [31, 6], hence, the *color-content of the scenes*[2], their complexities, and density and interactions of the foreground objects. Moreover, the ambient light level and variation of outdoor illumination is a function of solar elevation and depends on a multitude of factors, such as the presence of clouds and haze, pollution, and atmospheric turbidity [56]; see solar zenith angle in Figure 6. 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. For instance, DNN trained on aerial data from urban South Asian demographics struggles to detect objects accurately in videos captured in high-latitude European demographics, characterized by semi-rural pastures and lots of greenery; see Figure 1.

Therefore, in this paper, we address the following questions:

- Can we learn precise object representation in scenes captured in high-latitude, low-ambient light, sparsely populated European geographies with intertwined rural and urban pasture?
- Can we address the inherent challenges of object detection in aerial views by augmenting them with alternative views that possess enhanced visual perception?
- Can we address the challenge of sensor failure or unresponsive agents during aerial surveillance, in general, by adding one or more sensors on the ground, recording the same scene?

To answer these questions, we introduce **D**ual **V**iew **D**rone dataset, **DVD**, which captures synchronized aerial and ground view data for the first time. DVD is collected with consumer-grade handheld cameras (smartphones and GoPro) and drone-mounted cameras; see Section 3. It consists of around 2.5 hours of industry-standard 2.7K resolution video sequences, more than 0.5 million frames, covering both rural and urban pastures during spring and summer in high-latitude European geographies. This makes *DVD the largest ground and aerial-view dataset*, and *the second largest among all drone-based datasets across all modalities and tasks that ever existed*; see Table 1. The dual-view adds inference validation (if one view sees an object, then the object is present in the scene) in tracking and detection, and we expect DNNs trained on a multi-view dataset such as DVD would allow independent and combined detection.

In this study, we explore many unique properties of object detection in aerial images while evaluating DVD in a supervised setting. Our findings reveal that, compared to densely populated object scenes, detecting objects in scenes with varying object distribution (both sparse and dense) poses a greater challenge. Additionally, we observe that the top-performing object detectors in popular datasets may not yield optimal performance on DVD, emphasizing the significance of contextual information for accurate object detection. Furthermore, we demonstrate that augmenting object detectors with ground-view images could be the most effective strategy for achieving high detection performance compared to other pre-training approaches. Additionally, we benchmark DVD in a semi-supervised setting,

[2]European vehicles are comprising of mainly three colors [7]; 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. Okutama-Action and UCF-ARG are scripted dataset for human action recognition.

Dataset	Total Frames	Resolution	Total Annotations	Instances per Frame	Categories	Viewpoints	Region	Year
Campus [54]	929,499	1400×2019	19,564	0.02	6	Single (A)	North America	2016
UAV123 [46]	110,000	720×720	110,000	1.0	6	Multi (A)	Middle East	2016
Okutama-Action [13]	77,365	3840×2160	422,100	5.45	12	Single (A)	Asia	2017
CARPK[25]	1,500	$1,280 \times 720$	89,777	59.85	1	Single	Asia	2017
CarFusion[52]	53,000	$1,280 \times 720$	–	–	4	Multi	North America	2018
DAC-SDC [72]	150,000	640×360	NA	NA	12	Single	Asia	2018
UAVDT [19]	80,000	1080×540	841,500	10.52	3	Single	Asia	2018
MDOT [78]	259,793	–	–	–	9	Multi (A)	Asia	2019
Visdrone DET [77]	10,209	3840×2160	471,266	53.09	10	Single (A)	Asia	2019
Visdrone MOT [77]	40,000	3840×2160	1,527,557	45.83	10	Single (A)	Asia	2019
DOTA[69]	2806	4000×4000	188,282	67.09	15	Single (A)	Multiple	2019
MOR-UAV [43]	10,948	1280×720, 1920×1080	89,783	8.20	2	Single	Asia	2020
AU-AIR [16]	32,823	1920×1080	132,034	4.02	8	Multi	Europe	2020
UAVid [41]	300	3840×2160	–	–	8	Single	Europe	2020
UCF-ARG [47]	-	1920×1080	–	–	10	Multi (ARG)	North America	2020
MOHR [73]	10,631	5472×3078, 7360×4192, 8688×5792	90,014	8.47	5	Multi (A)	Asia	2021
DVD (This paper)	537,030	2700×1520	1,102,604	50.01	10	Multi (AG)	Europe	2023

utilizing the unlabeled frames to enhance the detection performance. This approach encourages the computer vision community to explore the *label-efficient object detection* methods for aerial images.

In summary, our work sheds light on the challenges associated with object detection in aerial images, introduces the Dual View Drone (DVD) Dataset, and offers valuable insights into enhancing performance by utilizing ground-view images, and semi-supervised learning techniques.

2 Related work

In this section, we briefly review publicly available drone-based datasets and state-of-the-art object detection algorithms, focusing on aerial images.

UAV-based datasets. The last decade witnessed a surge in UAV-based video and image datasets. We list some open-source UAV datasets, curated since 2016, and group their key features according to their downstream tasks; see Table 1 in [46] for a summary of pre-2016 UAV-based datasets.

VisDrone [77] 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 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* [54], is the largest aerial dataset for multi-target tracking, activity comprehension, and trajectory prediction, focuses solely on the university campus, in contrast to our DVD. *UAVDT* [19] dataset consists of 80,000 frames and 3 subsets, focusing on single and multi-object detection and tracking, under different weather condition, lighting, and altitude of the drone. *MOR-UAV* [43] 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* [46] is a low-altitude aerial dataset consisting of 112,578 fully-annotated images across 123 video sequences (simulated and recorded), designed for object tracking, with a subset intended for long-term aerial tracking. *MDOT* [78] 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* [16] 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* [72] is a single-object detection dataset with 150,000 images collected from *DJI* [4] with 12 categories. DOTA [69] is an aerial dataset (2,806 images, 15 categories) for object detection in earth vision.

Among others, *UVSD* [75] is a small-scale (5,874 images), multi-view, aerial dataset for vehicle detection and segmentation. *DroneVehicle* [57] (thermal infra-red+RGB) and *BIRDSAI* [15] (thermal infra-red) are small-scale, low-resolution datasets used for detection, tracking, and counting.

Figure 2: Left: Recording instances are classified 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 on three different locations, 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 sattelite view. Right: A sample recording arrangement for two-views with human operators.

In another line of work, *MVDTD* [35] is a collection of datasets to estimate 3D drone trajectories from multiple unsynchronized cameras. *UAVSwarm* [63] detects and tracks UAVs, [34] proposes drone-to-drone detection and tracking from a single drone-camera. *EyeTrackUAV2* [51] tracks drones from a ground perspective, specifically, from a *binocular* viewpoint.

Our proposed *DVD* 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), DVD is the first-ever *large-scale*, unscripted, multi-viewpoint video dataset (second largest among all UAV-based datasets ever after Campus which lacks object boundaries) recorded in *industry-standard* 2.7K resolution; (*b*) its multi-viewpoint 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); (*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. Section 3.2 explains more unique challenges of DVD.

There are UAV-based datasets with downstream tasks primarily orthogonal to DVD. For completeness, we list some UAV-based datasets for action detection, counting, geo-localization, 3D reconstruction, and benchmarking in Appendix A; also see [68].

(*iii*) **Object detection.** Object detectors based on CNNs are divided into two categories: two-stage and one-stage detectors. Two-stage detectors such as RCNN [22], Fast RCNN [21], Faster RCNN [53], employ a class-agnostic region proposal module followed by simultaneously regressing the object boundaries and their classes. In contrast, one-stage detectors like SSD [39], YoloV4 [61], YoloV6 [33], YoloV7 [62], YoloX [20], FCOS [58], directly predicts the image pixels as objects, leading to models that offer fast inference. Recently, by using neural architecture search, Yolo-NAS [11] claims to outperform previous Yolo models in real-time object detection. However, with the success of transformers, DETR [17] was the first transformer-based, end-to-end object detector. Following this, Deformable-DETR (D-DETR) [80] introduces a sparse attention module, computationally $6\times$ faster than DETR, and robust in detecting small objects. The majority of object detectors designed for aerial imagery draw upon the foundational principles established by these aforementioned popular object detectors [70, 71, 65]. Along this line, TPH-YoloV5 [79] combines YoloV5 with a transformer prediction head to solve the varying object scales and motion blur for drone-captured scenarios. As a result, our analysis utilizes the DVD dataset to benchmark these well-established methods, prioritizing factors such as fast inference, high precision, and the effective detection of small-scale objects. We also benchmark DVD with a semi-supervised object detection framework, Omni-DETR [64], to leverage available unlabelled aerial images to boost the detection performance. Omni-DETR

Figure 3: Different sample scenes (with annotation) from our dataset; the first row is the aerial-view, second row presents the same scenes from a ground camera. Similarly, the third row is the aerial-view, and the fourth row presents the same scenes from a ground camera. See more sample frames in Appendix B, Figure 8.

is a D-DETR-based student/teacher network that supports various forms of weak augmentations to generate pseudo-labels.

3 Dataset

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

3.1 General setup

Recording set-up. 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 6. 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 2. Based on that, we identify three recording scenarios (P1, P2, and P3). Out of them, 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 were unable to operate a drone due to the UAV-flight regulations; hence, we used a user-grade handheld camera set-up in 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 2 and 3 (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 are interacting and creating enormous visual challenges. Alongside, we choose harbor, single-lane roads in the countryside, asphalt roads, and bi-cycle lanes, in moderate traffic conditions, to collect simple scenarios which might have sparse to dense foreground objects (see sample frames in Figure 3).

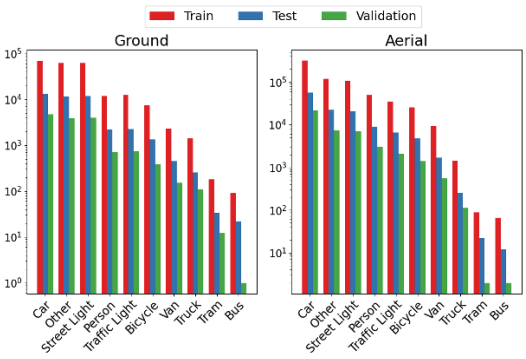


Figure 4: Total numbers of objects in each category in the ground and aerial view.

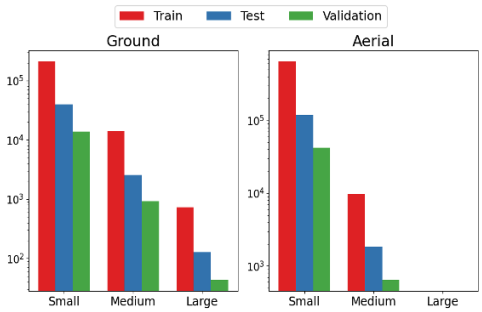


Figure 5: Number of large, medium, and small objects in the test, train and validation sets of two views; aerial view has no *large* object annotation.

Alignment of dual-views. Human operators simultaneously record the scenes from dual views; although a minute time-lapse is unavoidable. Consequently, after recording, clips are loaded into the `QuickTime` player, and a human operator manually synchronizes the frames to alleviate the time lapses. We note that for 12.5% of the clips (22.3% of image frames), we captured an extra ground view. Thus, these video clips offer three perspectives in total—two from ground and one from aerial.

Annotation and categories. DVD, as highlighted previously, stands as one of the largest drone datasets, encompassing millions of objects within its distribution. However, annotating each object within these images is a resource-intensive task that potentially restricts the assimilation of rich data. Inspired by the recent success of the semi-supervised learning paradigm in the computer vision community [64, 18, 40, 66, 45, 32], we opted to 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` [3], and leave the rest of the video unannotated; see `CVAT` interface in Figure 9. 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, we review these annotated videos. Similar to other benchmarks [77, 38], the annotated frames are assembled into `COCO-json` format to give a unique identifier for each object class.

3.2 Structuring, statistics, and challenges of DVD

In this section, we discuss the size, statistical properties, and challenges of the training distribution of DVD. We focus on two key points: (*i*) the distribution of different categories, and (*ii*) the 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. 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 4; also, see Figure 10. DVD contains over 1.1 million bounding box annotations in both views combined, rendering ~ 50.01 annotations per frame; see Figure 11-(a) and details in Table 7. The distribution follows a *long-tail pattern* 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.

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

Trained	Validation Set								Test Set							
DNN	Ground				Aerial				Ground				Aerial			
Models	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**

Object size distribution. To better illustrate the challenges in DVD, 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 5 (also, see Figure 11) 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 11-(c) for distribution for the object sizes.

Unique properties of DVD. DVD contains typical outdoor activities characterized by real-world properties like long-tail distribution, objects with similar appearance, viewpoint changes, varying illumination, etc. Additionally, DVD exhibits some unique properties, not found in other datasets: (*i*) Ground view contains occluded objects. Nevertheless, these objects can be recovered due to the wide field-of-view of the aerial view. This *dual-view feature of the DVD* has the potential to offer a wide range of solutions for scenes with occlusion, which remains a significant challenge in video surveillance. (*ii*) *DVD'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 12. (*iii*) Historically, vehicle color distributions vary across Europe, North America, and the Asia-Pacific; see Figure 7. 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 [10]. (*iv*) DVD was collected at *high latitudes*. The elevation of the sun in these areas (see Figure 6) 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 3 and 4 in Figure 3 shows this effect. (*v*) The aerial perspective inherent in DVD leads to *small objects* inclusion; their presence is susceptible to miss-detection by detection algorithms. (*vi*) DVD is characterized by both *sparse* and *dense* distribution of objects. Our empirical findings suggest that such a large variance in object distribution presents challenges in training object detectors, compared to scenes exhibiting only dense annotations.

4 Baselines and Evaluation

This section presents the benchmarking results on DVD 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 our dataset, **DVD**, 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 for each view contain 805 and 1,614 annotated images, respectively. We evaluate the models with the widely used metric for object detection, mean average precision (mAP) [38].

Object detector baselines. For *supervised benchmarking*, we use CNN-based YoloV7 [62], and transformer-based DETR [17] and D-DETR [80]. Additionally, we use Yolo-NAS [11]. For *semi-*

Table 3: Supervised benchmark on aerial view of DVD. The first column indicates percentage of infused ground-view samples with the aerial-view train set. The last column indicates the relative change in mAP compared to the baseline model that is trained exclusively on aerial-view training set from DVD.

Extra ground view samples	AP	AP_{50}	AP_S	AP_M	Relative(↑↓) change
12.5%	34.4	63.8	31.6	64.3	162.6% ↑
25%	**48.5**	73.3	45.8	73.6	270.2% ↑
37%	44.4	71.0	41.9	71.9	238.9% ↑
50%	40.8	69.0	38.6	73.5	211.5% ↑
75%	44.2	66.6	40.8	79.5	237.4% ↑
100%	42.3	65.7	38.9	68.4	222.9% ↑

(a) D-DETR

Extra ground view samples	AP	AP_{50}	AP_S	AP_M	Relative(↑↓) change
12.5%	30.9	57.7	33.7	59.4	1.3% ↓
25%	31.4	58.1	34.3	65.9	0.3% ↑
37%	35.8	68.4	34.7	66.8	14.4% ↑
50%	30.9	58.2	33.7	62.2	1.3% ↓
75%	45.3	79.1	43.0	79.6	44.9% ↑
100%	**48.3**	78.6	43.0	85.0	54.5% ↑

(b) YoloV7

Table 4: Semi-supervised Omni-DETR [64] benchmark on DVD. In the table, G and A denote number of ground- and aerial-view images, respectively. During the burn-in, we only use the labelled subset.

Labelled		Unlabelled		Test	Validation Set				Test Set			
G	A	G	A	perspective	AP	AP_{50}	AP_S	AP_M	AP	AP_{50}	AP_S	AP_M
0	8605	0	8605	A	29.3	49.3	24.8	60.6	19.8	38.4	19.5	35.0
8605	0	8605	0	G	56.9	83.3	54.8	74.9	45.8	75.5	45.4	58.7
8605	8605	8605	8605	A	34.9	59.2	32.1	**71.0**	24.2	48.7	23.7	**46.0**
2151	8605	0	8605	A	37.8	**64.9**	35.4	68.3	23.2	45.4	21.8	43.6
2151	8605	2151	8605	A	**38.0**	64.8	**35.7**	67.6	**26.7**	**54.1**	**24.5**	42.4

supervised benchmarking, we adapt transformer-based Omni-DETR [64] with D-DETR. We provide the implementation details and computing environment in the Appendix C.1; we refer to Tables 8 and 9 for other model specific implementation details.

4.1 Supervised benchmarking

Table 2 presents the supervised baselines results on DVD 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 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. These findings suggest that (*i*) object detectors pre-trained on the widely used MSCOCO dataset exhibit better generalization on the ground view, whereas training from scratch is more effective for achieving superior performance in the aerial view; and that (*ii*) CNN-based architectures, such as Yolo, outperform transformer-based architectures, indicating that fully convolutional architectures are more adept at handling small-scale objects. 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.

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 concatenating two sets of aerial- and ground-view samples along with their corresponding annotations. Our findings demonstrate that the inclusion of ground-view samples substantially improves the object detection. Table 3 illustrates 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 of 270% even 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. This highlights the need for future research to explore and develop effective sampling strategies for improved performance. Similar observation holds for DVD test set; see Table 10. We show the qualitative results in Figure 15.

Table 5: Domain tests on DVD using D-DETR, evaluated on validation and test set.

Training Protocol	Validation Set				Test Set			
	AP	AP_{50}	AP_S	AP_M	AP	AP_{50}	AP_S	AP_M
Trained from scratch on DVD	13.1	28.3	14.2	38.1	10.3	25.0	10.1	29.4
Pre-trained on VisDrone, fine-tuned on DVD	23.4	45.8	25.6	51.0	20.9	41.9	20.6	43.8
Pre-trained on DVD_{ground}, fine-tuned on DVD_{aerial}	30.0	55.9	26.8	46.6	32.3	59.4	29.0	43.8
Pre-trained on VisDrone, fine-tuned on DVD (8k aerial, 2k ground)	**38.4**	**65.0**	**34.8**	**77.2**	**35.1**	**64.4**	**33.6**	**71.6**

4.2 Semi-supervised benchmarking

To exploit the unlabelled aerial images, we evaluate the DVD dataset using a semi-supervised framework in Table 4. For this framework, we employ Omni-DETR [64] by adapting the object detector to D-DETR. The approach involves a two-stage process: a *Burn-in stage* where we train a D-DETR with available labels and a *consistency learning stage* following [64]. In our experiments, we utilize all labeled images in the burn-in stage and an equal number of unlabeled images in the second stage. Our 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. Furthermore, when employing the Omni-DETR framework with all labeled and unlabelled image frames, we achieve a significant improvement. Building upon the insights gained from the supervised benchmarking, we utilize all labeled aerial images and 25% of labeled ground view images in the burn-in stage. However, while this model performs well on the validation set, it underperforms on the test set. Interestingly, incorporating an additional 25% of unlabelled ground images in the second stage leads to superior performance compared to all other models, including the one utilizing all unlabelled ground images.

4.3 Transfer learning on DVD dataset

Table 5 presents an analysis of various pre-training strategies and knowledge transfer trends on the DVD dataset. We observe that pre-training the model on Visdrone leads to a 78.6% improvement in object detection performance on the DVD dataset. However, pre-training the model on the ground view images of DVD yields an even greater improvement of 129%. Similarly, training a Visdrone pre-trained model on a complete set of aerial images and 25% of ground view images yields an object detection model that surpasses all other representative models in performance.

5 Conclusion and Future Work

In this paper, we introduce a large-scale, high-definition ground and aerial-view video dataset, DVD. To the best of our knowledge, DVD is the first drone-based aerial object detection dataset that exploits the multi-modality of the data coming from orthogonal views, aerial and ground to offer enhanced detection capacity for an aerial view. We used supervised and semi-supervised learning techniques with convolution and transformer-based DNN models to perform an extensive benchmarking on DVD and report many interesting findings attributed to the dual view of the dataset. 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.

During the data acquisition, we observed and analyzed how a spectrum of low-key factors, e.g., ambient light, latitude, altitude, atmospheric turbidity, etc., are related and directly influence the colorfulness of the scenes. These factors, together with the dual view, add several unique challenges to DVD. Nevertheless, providing a solution involving all these factors is not in the scope of this work. Also, annotating a large video dataset is a resource-intensive task. By providing partial annotation of the DVD, and by benchmarking the DVD in semi-supervised setting, we encourage the machine learning community to actively design *label-efficient models* where the multi-view of the data may provide a better solution to diverse video understanding tasks when annotations are scarce.

References

[1] Car colour popularity. https://en.wikipedia.org/wiki/Car_colour_popularity.

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

[3] CVAT annotation tool. https://www.cvat.ai.

[4] DJI. https://www.dji.com.

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

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

[7] 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.

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

[9] Solar zenith-angle. https://en.wikipedia.org/wiki/Solar_zenith_angle.

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

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

[12] 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*, volume 1, pages 68–75, 2020.

[13] 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] 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. 2016.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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*, volume 31, 2017.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

[48] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science and Business Media, 2003.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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*, volume 1, pages 525–532, 2014.

[56] Manuel Spitschan, Geoffrey K. Aguirre, David H. Brainard, and Alison M. Sweeney. Variation of outdoor illumination as a function of solar elevation and light pollution. *Scientific reports*, 6(1):1–14, 2016.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

Contents

A Related Work—Continued

This section extends the discussion in Section 2 of the main paper by including additional UAV-based datasets that focus on different downstream tasks such as action detection, counting, geo-localization, 3D reconstruction, and benchmarking; also, see [68].

(*i*) **Human tracking.** PNNL 1 and 2 [8] are unannotated datasets consisting of 1,000 and 1,500 frames, respectively, designed for human tracking from a fixed perspective with long-term inter-object occlusion.

(*ii*) **Action detection from aerial viewpoints.** UCF-ARG [47] is a multi-view, scripted dataset, designed for 10 different human action detection, where the scenes are recorded from 3 different views—a rooftop camera, a ground camera, and an aerial camera. Okutama-Action [13] is an aerial dataset consisting of 77,365 annotated frames, designed for 12 concurrent human action detection.

(*iii*) **Counting and 3D reconstruction.** CARPK [25] is a single-view video dataset, captured from a moving drone, contains nearly 90,000 cars from 4 different parking lots, and is used for predicting the car-counts in a scene. CarFusion [52] is a multi-view dataset consisting of 53,000 fully-annotated

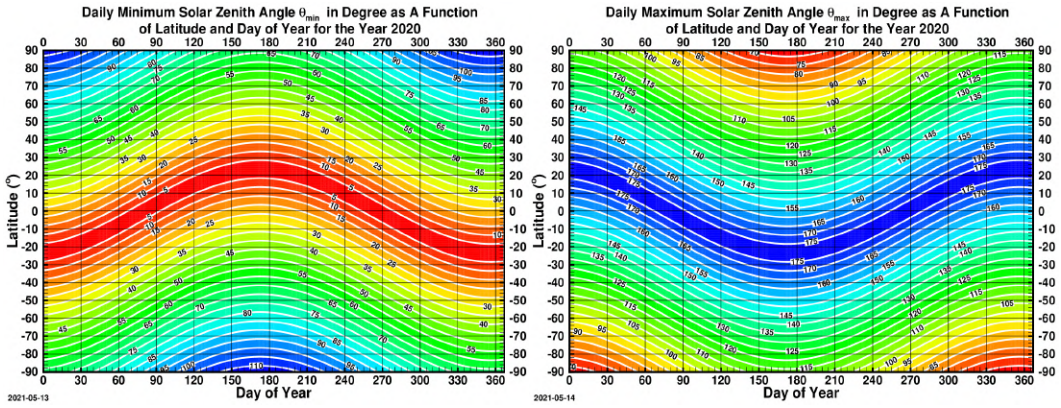


Figure 6: The daily minimum and maximum of the solar zenith angle as a function of latitude and day of year for the year 2020. In the Earth-Centered Earth-Fixed (ECEF) geocentric Cartesian coordinate system, let (ϕ_s, λ_s) and (ϕ_o, λ_o) be the latitudes and longitudes of the subsolar point and the observer's point, then the upward-pointing unit vectors at the two points, $\mathbf{S}$ and $\mathbf{V}_{oz}$, are $\mathbf{S} = \cos\phi_s \cos\lambda_s \mathbf{i} + \cos\phi_s \sin\lambda_s \mathbf{j} + \sin\phi_s \mathbf{k}$, and $\mathbf{V}_{oz} = \cos\phi_o \cos\lambda_o \mathbf{i} + \cos\phi_o \sin\lambda_o \mathbf{j} + \sin\phi_o \mathbf{k}$, where $\mathbf{i}, \mathbf{j}$ and $\mathbf{k}$ are the basis vectors in the ECEF coordinate system. Consequently, cosine of the solar zenith angle, θ_s, is the inner product between $\mathbf{S}$ and $\mathbf{V}_{oz}$. Source: [9].

frames, 100,000 car instances with 14 semantic key points, captured from 18 moving cameras at multiple locations, designed for 3D reconstruction of cars.

(*iv*) **Geo-localization** is a challenging problem, and over the past years, some dedicated datasets were proposed to devise efficient solutions to this problem. Danish airs and grounds (DAG) dataset [59] is a large collection of ground-level and aerial images covering about 50 kilometers in urban and rural environments with the extreme viewing-angle difference between query and reference images is a dataset for place recognition and visual localization. Similar to DAG, [42] assembled a much smaller dataset with a drone and GoogleMap images. For more details in this context, refer to [55, 37].

(*v*) **Other downstreaming tasks.** SeaDronesSee [60] is curated for single and multi-object tracking, specifically people, floating in water. DroneSURF [28] is for person identification, especially facial recognition, in an urban environment, while [67] works on object detection, tracking, and counting. P-DESTRE [30] is a dataset designed to test pedestrian detection, tracking, re-identification, and search methods. VIRAT [49] is a video dataset from surveillance cameras, designed for testing on real-world environments and challenges.

(*vi*) **Benchmarking and evaluation.** The UAV Benchmark [24] and [36] present datasets that maximize their breadth of usability, and provide extensive comparisons, including camera motion estimation. Finally, in [68], Wu et al. provides challenges and statistics of existing DL based methods for UAV-based object detection and tracking.

B Addendum to the Dataset

In this section, we provide some extra insights on the structuring and statistics of the DVD. Additionally, we discuss about the `CVAT` annotation tool in Section B.1, and provide an analysis of color distribution of different drone based datasets and contrast them with DVD; see Section B.2.

B.1 `CVAT` annotation tool

`CVAT` is an industry-standard, open-source, cutting-edge, interactive annotation tool that produces professional-level image and video annotations for diverse computer vision tasks [3]. `CVAT` is equipped with an in-built tracker that can track an object consecutively for a few frames and results in an easier and faster annotation. Annotating in `CVAT` is done by annotating category by category.

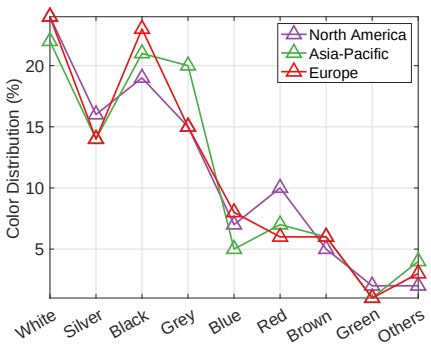


Figure 7: Car color popularity surveys conducted by American paint manufacturer DuPont for the year 2012. Source: [1].

Table 6: Details of the recording devices.

Drone/UAV	DJI Phantom 4, DJI mini 2
ISO Range	100-3200
Lens	FOV 94° 20 mm, FOV 83° 20 mm
GoPro	GoPro HERO4, HERO 6
ISO range	100-800
iphone	11, 13-Pro (when UAV not used)
FOV	120°
Resolution (GoPro, Drone)	2.7K (2704x1520) 30fps
Filetype video	.mp4 (.mov)
Filetype image	.png

Table 7: Summary of annotations in both views of DVD.

View	Train set annotations	Test set annotations	Validation set annotations	Total annotations	Total Frames	Annotations per frame
Aerial	655,608	120,517	42,927	819,052	11,024	74.23
Ground	226,461	42,440	14,651	283,552	11,024	25.72
Combined	882,069	162,957	57,578	1,102,604	22,048	50.01

This can either be done frame by frame or within an interval of frames relying on the built-in tracker for the frames in between. Figure 9 presents one such instance of annotation interface using `CVAT`.

B.2 Color distributions of different datasets—An experimental analysis

The color content of different geographies on the earth is quite diverse. Many recent studies show that the latitude influences the solar elevation, and hence the population density [31, 6] of different parts of the world. These factors have a direct effect on *color-content of the scenes*. In this scope, we analyze the color content of sample video frames from different datasets based on two key points: (*i*) color distribution in the sample frames of different datasets based on `RGB` color channels, and (*ii*) dominant color distributions in the sample frames of the datasets.

Color distribution of different datasets based on `RGB` color channels. We show the color distributions of sample frames from different datasets in Figure 12. For each dataset, we randomly sample 1000 images. All images are resized to 600×337 and an *average image* is computed. Then, a color histogram is computed for each color channel of the *average image*, and the area under each curve representing each color channel is calculated. Except for UAV123, the area under the green channel for all other datasets is about 1.5-2$\times$ lower than the DVD aerial view. However, the blue color channel of DVD is the most dominant in the aerial view. Additionally, the distribution of the blue and green channels in the ground view of the DVD are doubly-peaked, covering almost similar areas under them.

Dominant colors in DVD and other datasets. We use the `Python` tool `extract-colors-py`, which groups colors based on their visual similarities by using the `CIE76` standard [2]. The tool, `extract-colors-py` uses two hyperparameters: (*i*) the tolerance, ϵ, that determines how two colors can be grouped (default $\epsilon = 32$), and (*ii*) color limit, that is the upper limit of extracted colors in the output. We set both the ϵ and the color limit to 12 and plot the grouped colors with their percentages. In Figure 13, we analyze the most dominant colors in DVD in different sample scenes (aerial and

Figure 8: Different sample scenes (with annotation) from our dataset; the first row is the aerial-view, second row presents the same scenes from a ground camera. Similarly, the third row is the aerial-view, and the fourth row presents the same scenes from a ground camera. Some scenes have a dense object annotations, while some scenes have very few object annotations. This high variance in object distribution across different scenes in DVD is complementary to datasets like VisDrone where object detection is relatively straightforward due to their biased object distribution (dense), reflecting its demographic characteristics.

ground), while Figure 14 shows the dominant colors in other datasets. Indeed, the dominance of different spectra of blue, yellow, and green colors in DVD in both views as shown in Figure 13 directly supports our findings in Figure 12, and make DVD a stand-alone video dataset compared to the other large-scale, drone-based datasets such as VisDrone [77], UAV123 [46], Campus [54].

C Addendum to the Baseline and Evaluation

This section highlights the implementation details of our baseline DNN models; see Table 8 and 9. In Section C.2, we provide additional benchmarking results complementing Section 4 in the main paper.

C.1 Implementation details

We train all object detectors for 39 epochs on 600×337 scaled images, except DETR. DETR is a compute-heavy model and requires more than 39 training epochs [17, 80] for an optimal performance. For supervised benchmarking, we train DETR with 100 object queries, and 10 classes (9 object class, 1 background class) for 300 epochs. For D-DETR, we used 900 queries and 20 classes. We adhere to the original training methodologies of the respective methods in order to train the object detectors specifically for the DVD dataset.

Computing environment. For prototyping, we use a local testbed with an AMD EPYC 7501 32-Core Processor with 2.0GHz speed, 16 GB memory, and 1 Nvidia Tesla V100 GPU with 32 GB on-board

Figure 9: **A sample annotation using `CVAT` [3] interface.** `CVAT` has an in-built tracker that tracks an object through multiple frames. The inbuilt tracker speeds up the annotation part — once a particular frame is annotated, around 10 frames after that require minimal human supervision — leveraging the tracker. This property makes `CVAT` an attractive annotation tool.

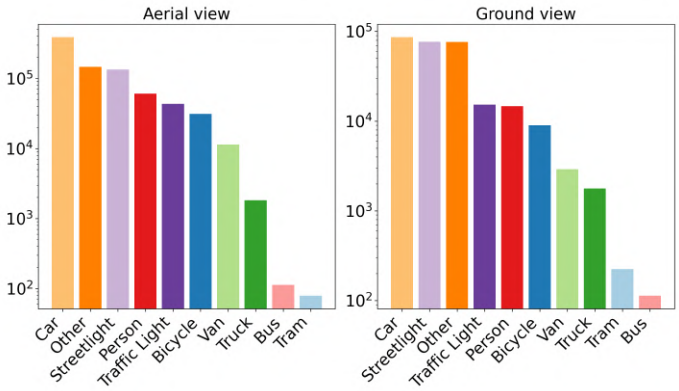


Figure 10: Total numbers of objects in each category in the aerial and ground view.

memory. For training all the supervised baselines, we use two HPC nodes: (*i*) Node-1: 2x Intel(R) Xeon(R) Gold 6230 CPU with 2.10 GHz processing speed, 32 virtual cores, 192 GB memory, and 8 NVIDIA V100 GPU each with 32 GB on-board memory; (*ii*) Node-2: AMD EPYC 7F72 CPU with 3.2 GHz processing speed, 96 virtual cores, 2048 GB memory, and 8 NVIDIA A100 GPU each with 40 GB on-board memory. For training the semi-supervised baselines, we use a server with AMD EPYC 7662 CPU, 1024GB memory, 8 RTX A5000 GPU.

C.2 Additional baseline results

In Table 10, we provide the supervised benchmark results on the test of the aerial-view of DVD by using D-DETR and YoloV7. Except a few minor discrepancies, overall our observation in the main paper holds on DVD test set results — We demonstrate that the inclusion of ground-view samples substantially improves the object detection performance.

C.2.1 Benchmarking with mix-up across views

We use the mix-up strategy to naturally augment and combine the dual views of our data.

Why mix-up? Previously, we demonstrated that jointly training the aerial-view samples with ground-view samples substantially improves object detection from an aerial perspective; see Section 4.1. Nevertheless, a natural question could be—Can a *data-augmentation strategy* be able to improve the

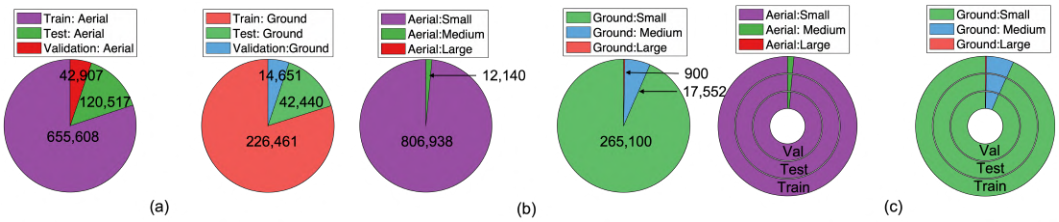


Figure 11: (a) Total number of annotations in train, test, and validation sets of aerial and ground view; (b) number of objects based on their sizes in aerial and ground view, aerial view has no *large* object annotation; (c) percentage of small, medium, and large objects in train, test, and validation sets of aerial and ground view.

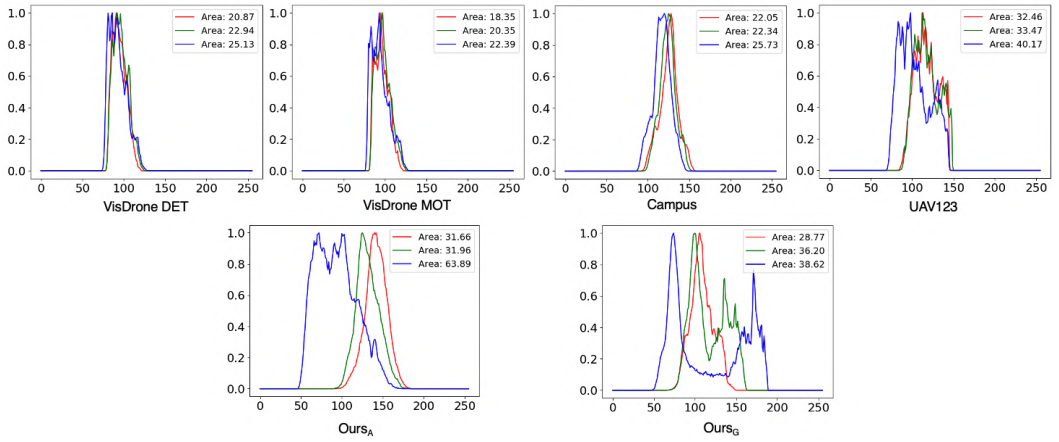
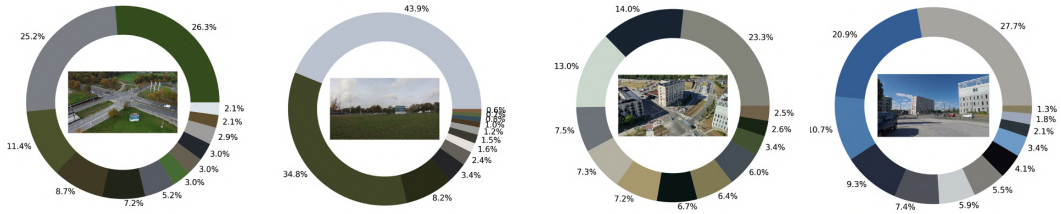


Figure 12: **Color distribution of different datasets.** In the top row, we show the color distribution of VisDrone DET and MOT, the Campus dataset, and the UAV123 dataset. VisDrone represents south-east Asian geographies (collected in 14 cities across China) [77]; the Campus dataset represents North American geographies, collected in Stanford University campus [54]; UAV123 represents the Middle East, collected primarily in King Abdullah University of Science and Technology's campus and its surroundings (Kingdom of Saudi Arabia) [46]. In the bottom row, we show the ground and aerial view color distribution of DVD.

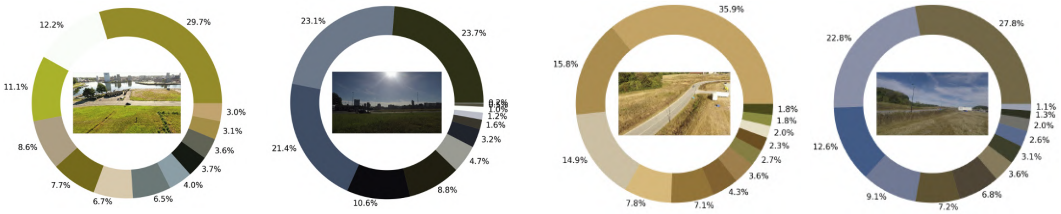
aerial-visual perception while aerial-view images are *augmented* with corresponding ground-view images? This motivates us to use mix-up [74] as an augmentation strategy that can combine these two views.

The mix-up is a data augmentation technique that creates a convex combination of the input data pair and their labels and reduces the inductive bias [74]. For input pair, (x_A, x_G), and their corresponding labels, (y_A, y_G), mix-up creates new input, $x_m = \lambda x_A + (1-\lambda)x_G$, and label, $y_m = \lambda y_A + (1-\lambda)y_G$, where $\lambda \in [0, 1]$ is the mixing parameter sampled from a $\beta_{\alpha,\beta}$-distribution with $\alpha = \beta = 1$. Thus, we apply mix-up to the 8605 pairs of aerial and ground-view samples in the input space, while the testing perspective remains the aerial view. Note that our approach to mix-up differs from the original concept. We consistently apply mix-up across the views for the same samples, as opposed to performing mix-up among random samples within a batch.

D-DETR and YoloV7 training results with mix-up. Each sample, S, consists of a pair of ground and aerial images, (x_G, x_A) of the same scene. During training, we sample the mixing parameter, $\lambda \sim \beta_{1,1}$ such that $\lambda > 0.5$, resulting in A as the dominant image. The best mAP corresponds to $\lambda \in [0.75, 1]$ for D-DETR on DVD; see Table 11 for ablation study for the optimal λ. For YoloV7, we use the best λ from the mix-up D-DETR experiments. The results in Table 11 suggest that D-DETR with mix-up parameter $\lambda > 0.5$ renders a better performance than vanilla D-DETR trained only on aerial view images; see Table 2 in Section 4. YoloV7 with mix-up parameter, $\lambda \in [0.75, 1]$

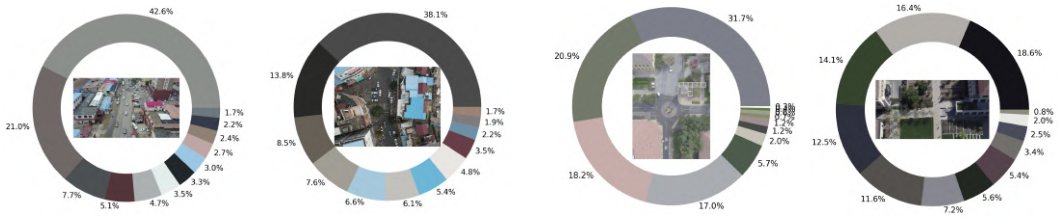


(a) Scene 1 Aerial (b) Scene 1 Ground (e) Scene 2 Aerial (f) Scene 2 Ground

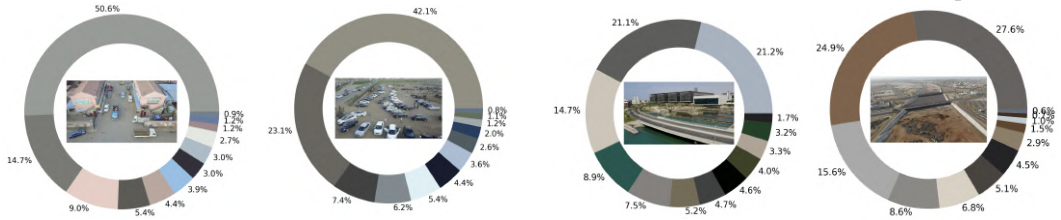


(c) Scene 6 Aerial (d) Scene 6 Ground (g) Scene 8 Aerial (h) Scene 8 Ground

Figure 13: Dominant colors in different scenes of DVD containing both views.



(a) VisDrone DET (b) VisDrone DET (e) Campus (f) Campus



(c) VisDrone MOT (d) VisDrone MOT (g) UAV123 (h) UAV123

Figure 14: Most dominant colors in the sample frames of VisDrone DET and MOT [77], the Campus dataset [54], and the UAV123 dataset [46].

performs better than the mix-up D-DETR. Overall, we can conclude that mix-up D-DETR is better than the vanilla D-DETR model trained only on aerial images; for YoloV7, the performance is almost similar. In our experiments, mix-up technique uses 17,210 images (8,605 pairs of ground and aerial view images), while only *a fraction of the 8,605 ground view images* jointly trained with 8,605 aerial images can surpass its performance as evident from Tables 3 and 10. In conclusion, although our cross-view mix-up technique enhances object detection performance, the superior strategy for improving aerial detection performance is to train aerial-view samples together with ground-view samples. Future work will explore combining both the strategies (joint training and mix-up) to improve the performance of downstream tasks in aerial perspective.

D Reproducibility, Privacy, and Broader Impact

This paper introduces a large-scale, high-definition ground and aerial-view video dataset, **DVD**, and performs extensive benchmarking on the data. The dataset is open-source, fully curated, prepared, and and we plan to release our dataset via an academic website for research, academic, and commercial

Table 8: DNN models used for benchmarking. Note that $1M = 10^6$.

Type	Model	Task	Dataset	Parameters	Optimizer	Platform	Metric
CNN	YoloV7 [62]	Detection	DVD	36.5M	SGD-M [48]	PyTorch	mAP
NAS	Yolo-NAS (L) [11]	Detection	DVD	51.1M	Adam [29]	PyTorch	mAP
Transformer	DETR [17]	Detection	DVD	41M	Adam [29]	PyTorch	mAP
	D-DETR [80]	Detection	DVD and VisDrone	41M	Adam [29]	PyTorch	mAP
	OMNI-DETR [64]	Detection	DVD and VisDrone	41M	Adam [29]	PyTorch	mAP

Table 9: Hyperparameters used for training each DNN model.

Model	Backbone	Learning Rate	Batch Size	Weight Decay	Queries	Attention Heads	Epochs
YoloV7 [62]	E-ELAN	$1, 10^{-5}, 10^{-1}$	32	5×10^{-4}	NA	NA	39
Yolo-NAS (L) [11]	QA-RepVGG	$10^{-6}, 5 \times 10^{-4}$	16	10^{-4}	NA	NA	39
DETR [17]	ResNet50 [23]	10^{-4}	2	10^{-4}	100	16	300
D-DETR [80]	ResNet50	2×10^{-4}	2	10^{-4}	900	16	39
OMNI-DETR [64]	ResNet50	10^{-4}	2	10^{-4}	900	16	39

use. The dataset is protected under the CC-BY license of creative commons, which allows the users to distribute, remix, adapt, and build upon the material in any medium or format, as long as the creator is attributed. The license allows DVD for commercial use. As the authors of this manuscript and collectors of this dataset, we reserve the right to distribute the data. Additionally, we provide the code, data, and instructions needed to reproduce the main experimental baseline results, and the statistics pertinent to the dataset. We specify all the training details (e.g., data splits, hyperparameters, model-specific implementation details, compute resources used, etc.).

There are human subjects present in the data, although there are no personal data that can resemble shreds of evidence, reveal identification, or show offensive content. Therefore, DVD is not subject to IRB (for North America) or GDPR (for Europe) compliance as it has no privacy concerns. We thoroughly discussed and validated this issue with appropriate legal experts.

The dataset can be used by multiple domain experts. Its application includes but is not only limited to surveillance, autonomous driving [44, 14], robotics and instructional videos [68], environmental monitoring [50], heavy industrial infrastructure inspection [12], developing livable and safe communities [26, 5, 76], and a few to mention. Although we do not find any foreseeable harms that the dataset can pose to human society, it is always possible that some individual or an organization can use this idea to devise a *technique* that can appear harmful to society and can have evil consequences. However, as authors, we are absolutely against any detrimental usage of this dataset, regardless by an individual or an organization, under profit or non-profitable motivation, and pledge not to support any detrimental endeavors concerning our data or the idea therein.

Table 10: Supervised benchmark on aerial view of DVD (Test Set). The first column indicates percentage of infused ground-view samples with the aerial-view train set. The last column indicates the relative change in mAP compared to the baseline model that is trained exclusively on aerial-view training set from DVD.

Extra ground view samples	AP	AP_{50}	AP_S	AP_M	Relative(↑↓) change
12.5%	39.8	68.6	39.9	55.8	286.4% ↑
25%	**44.8**	**71.5**	**42.9**	**72.4**	335.0% ↑
37%	41.1	69.1	39.7	61.6	299.0% ↑
50%	36.0	65.8	33.0	54.1	249.5% ↑
75%	28.7	56.6	26.6	62.8	178.6% ↑
100%	39.9	65.8	32.5	70.6	287.4% ↑

(a) D-DETR

Extra ground view samples	AP	AP_{50}	AP_S	AP_M	Relative(↑↓) change
12.5%	29.5	55.6	28.8	64.6	5.6% ↓
25%	30.1	56.2	29.5	64.1	3.8% ↓
37%	33.1	63.3	30.4	70.0	5.8% ↑
50%	29.6	59.0	29.2	66.1	5.4% ↓
75%	40.5	74.6	36.7	74.7	29.4% ↑
100%	**45.5**	**76.1**	**43.8**	**81.6**	45.4% ↑

(b) YoloV7

Table 11: Mix-up benchmarks after 39 epochs; the test perspective is the aerial view.

Model	Mix-up parameter	Validation Set				Test Set			
		AP	AP_{50}	AP_S	AP_M	AP	AP_{50}	AP_S	AP_M
D-DETR	[0.65, 1.0]	22.8	44.0	22.6	49.8	22.3	42.4	22.0	50.1
	[0.75, 1.0]	**33.4**	56.0	**31.2**	56.1	**29.1**	49.6	27.0	47.7
	[0.85, 1.0]	28.2	50.1	25.5	55.9	23.5	44.9	22.0	44.9
	0.9	25.8	41.6	28.3	46.4	23.3	41.3	25.0	42.3
	[0.0, 1.0]	6.4	12.5	8.7	9.1	10.4	17.7	12.9	13.3
YoloV7	[0.75, 1.0]	30.3	**58.6**	29.8	**60.7**	28.5	**55.3**	**27.9**	**57.9**

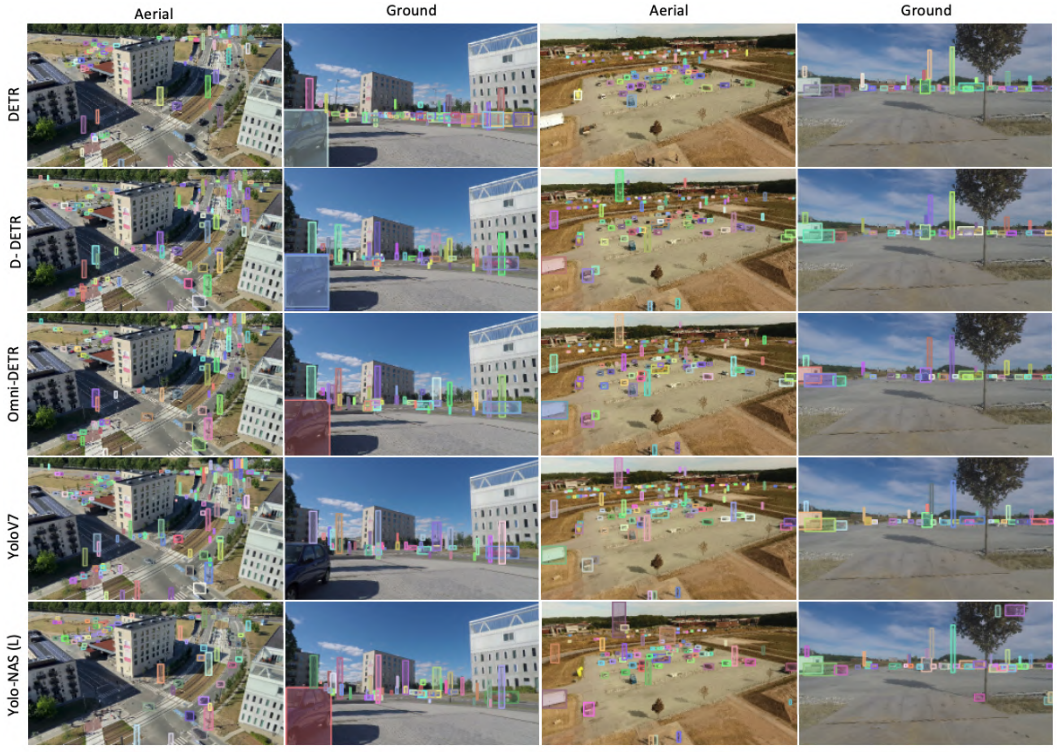


Figure 15: Qualitative inference results of different DNN models on the test set of DVD.