Appendix

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 [78].

(*i*) **Human, vehicle, and drone trajectory tracking.** PNNL 1 and 2 [3] 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. The highway-drone dataset [35] is a large-scale dataset collected from 6 different locations on German highways, crafted for the safety validation of automated vehicles. The dataset consists of more than 110,500 vehicle annotations, recorded over 147 hours, and offers each vehicle's trajectory, including type, size, and maneuvers. Among others, *UVSD* [85] is a small-scale (5,874 images), multi-view, aerial dataset for vehicle detection and segmentation. *DroneVehicle* [67] (thermal infra-red+RGB) and *BIRDSAI* [16] (thermal infra-red) are small-scale, low-resolution datasets used for detection, tracking, and counting.

MVDTD [42] is a collection of datasets to estimate 3D drone trajectories from multiple unsynchronized cameras. *UAVSwarm* [71] detects and tracks UAVs. [41] proposes drone-to-drone detection and tracking from a single drone-camera. *EyeTrackUAV2* [60] tracks drones from a ground perspective, specifically, from a *binocular* viewpoint.

(*ii*) **Action detection from aerial viewpoints.** UCF-ARG [55] 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 [14] is an aerial dataset consisting of 77,365 annotated frames, designed for 12 concurrent human action detection.

(*iii*) **Counting and 3D reconstruction.** CARPK [29] 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 [61] is a multi-view dataset consisting of 53,000 fully-annotated 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 [69] 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, [50] assembled a much smaller dataset with a drone and GoogleMap images. For more details in this context, refer to [45, 65].

(*v*) **Other downstreaming tasks.** SeaDronesSee [70] is curated for single and multi-object tracking, specifically people, floating in water. DroneSURF [32] is for person identification, especially facial recognition, in an urban environment, while [77] works on object detection, tracking, and counting. P-DESTRE [36] is a dataset designed to test pedestrian detection, tracking, re-identification, and search methods. VIRAT [58] is a video dataset from surveillance cameras, designed for testing on real-world environments and challenges.

(*vi*) **Benchmarking and evaluation.** The UAV Benchmark [28] and [43] present datasets that maximize their breadth of usability, and provide extensive comparisons, including camera motion estimation. Finally, in [78], 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 MAVREC. 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 MAVREC; 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 [2]. `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. 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 10 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 [8, 37] 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.

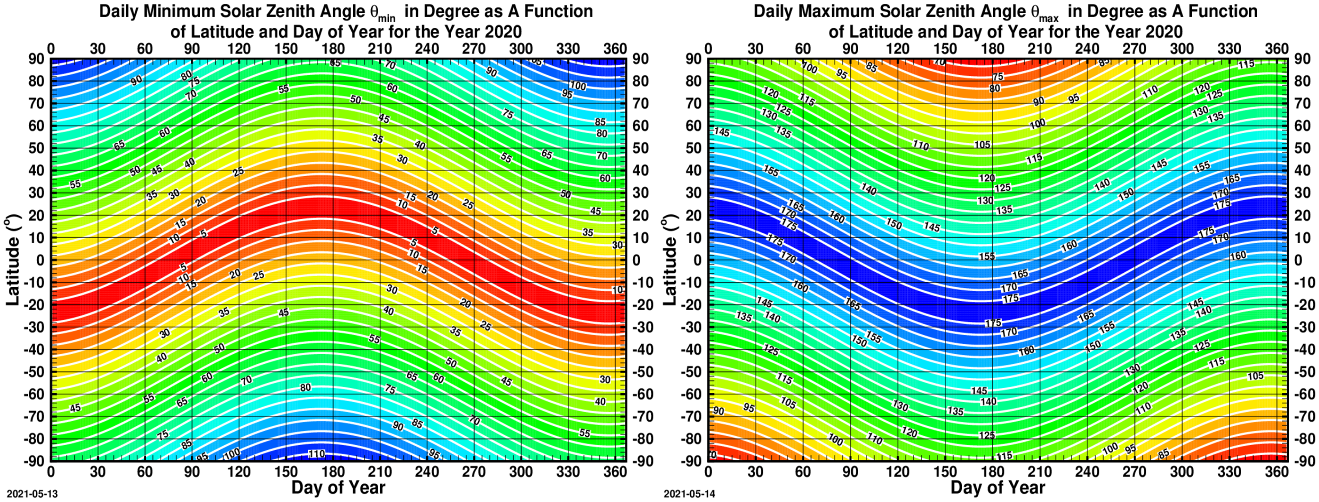


Figure 7. The daily minimum and maximum of the solar zenith angle as a function of latitude and day of year for the year 2020. In the Earth-Centered Earth-Fixed (ECEF) geocentric Cartesian coordinate system, let (ϕ_s, λ_s) and (ϕ_o, λ_o) be the latitudes and longitudes of the subsolar point and the observer's 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: [10].

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 5. Details of the recording devices.

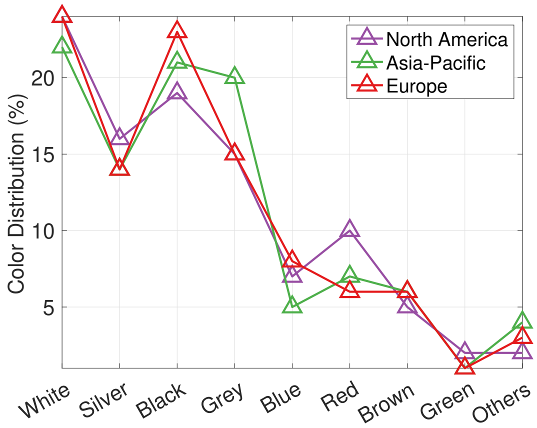


Figure 8. Car color popularity surveys conducted by American paint manufacturer DuPont for the year 2012. Source: [9].

Color distribution of different datasets based on `RGB` **color channels.** We show the color distributions of sample frames from different datasets in Figure 13. 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× lower than the MAVREC aerial view. However, the blue color channel of MAVREC is the most dominant in the aerial view. Additionally, the distribution of the blue and green channels in the ground view of the MAVREC are doubly-peaked, covering almost similar areas under them.

Dominant colors in MAVREC and other datasets. We use the `Python` tool `extract-colors-py`, which groups colors based on their visual similarities by using the `CIE76` standard [1]. 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

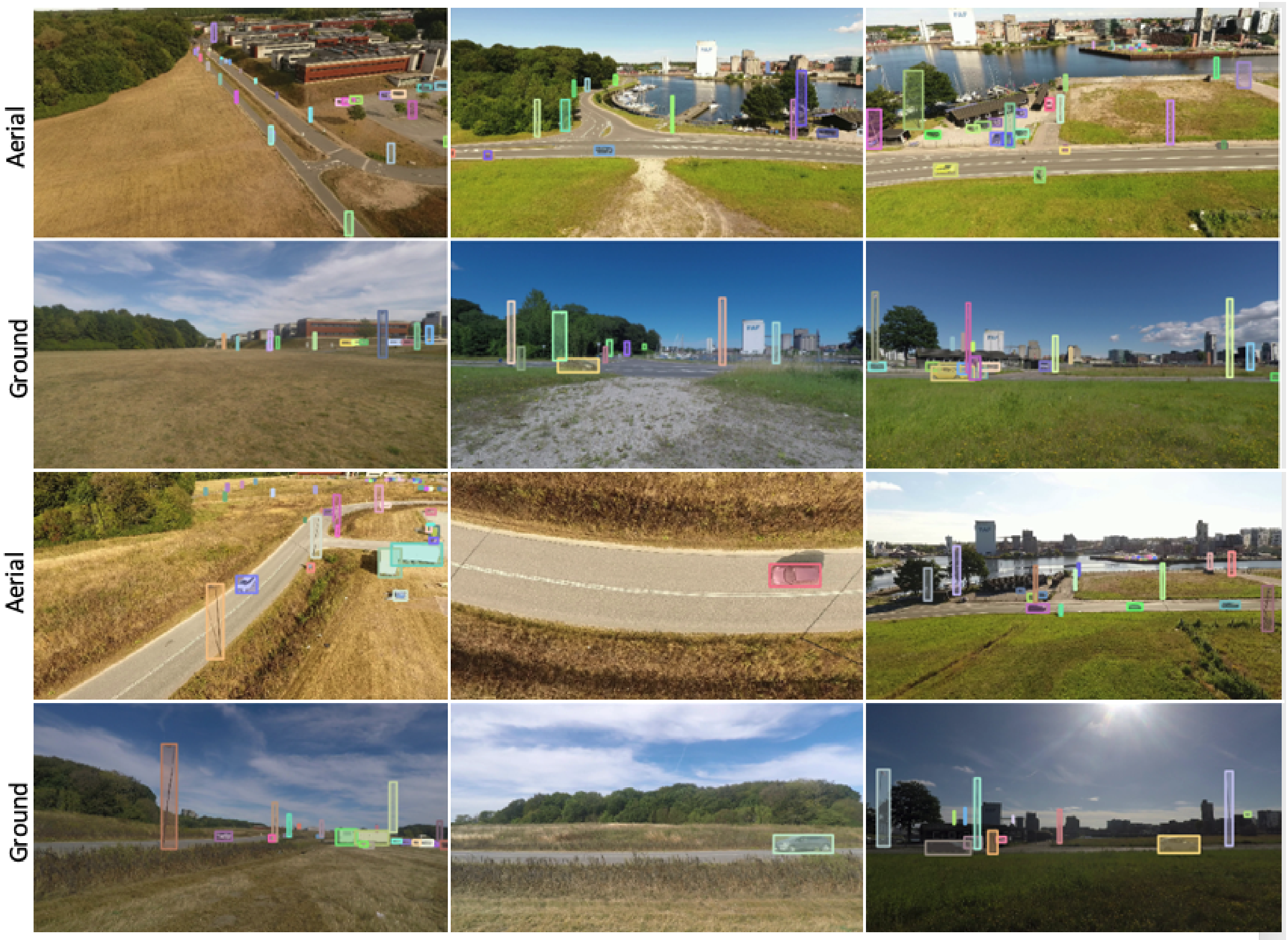


Figure 9. Different sample scenes (with annotation) from our dataset; the first row is the aerial-view, second row presents the same scenes from a ground camera. Similarly, the third row is the aerial-view, and the fourth row presents the same scenes from a ground camera. Some scenes have a dense object annotations, while some scenes have very few object annotations. This high variance in object distribution across different scenes in MAVREC is complementary to datasets like VisDrone [88] where object detection is relatively straightforward due to their biased object distribution (dense), reflecting its demographic characteristics.

Table 6. Summary of annotations in both views of MAVREC.

View	**Train set annotations**	**Test set annotations**	**Validation set annotations**	**Total annotations**	**Total annotated 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

grouped colors with their percentages. In Figure 14, we analyze the most dominant colors in MAVREC in different sample scenes (aerial and ground), while Figure 15 shows the dominant colors in other datasets. Indeed, the dominance of different spectra of blue, yellow, and green colors in MAVREC in both views as shown in Figure 14 directly supports our findings in Figure 13, and make MAVREC a stand-alone video dataset compared to the other large-scale,

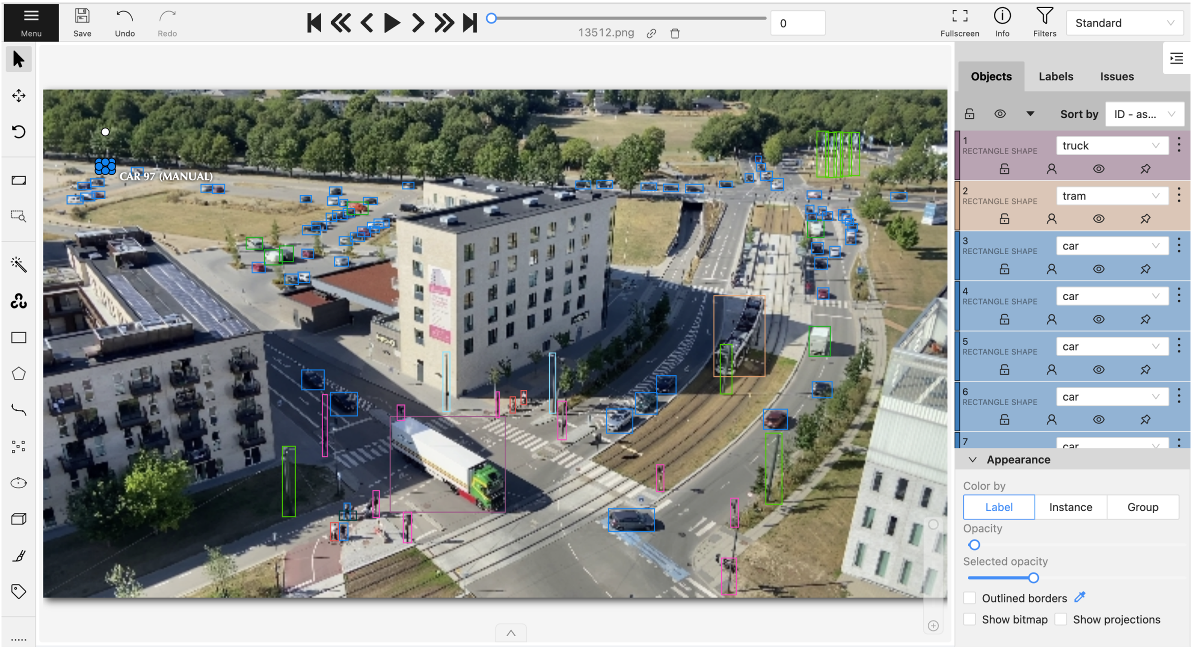


Figure 10. **A sample annotation using** `CVAT` **[2] interface.** `CVAT` has an in-built tracker that tracks an object through multiple frames. The inbuilt tracker speeds up the annotation part — once a particular frame is annotated, around 10 frames after that require minimal human supervision — leveraging the tracker. This property makes `CVAT` an attractive annotation tool.

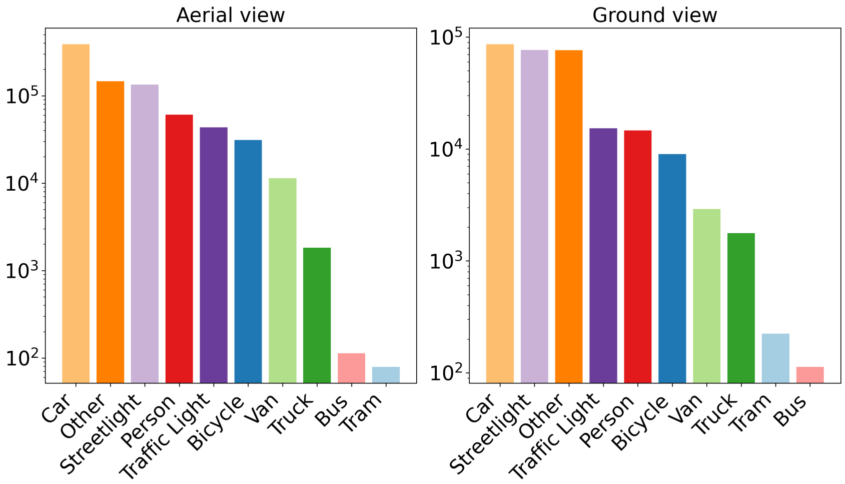


Figure 11. Total numbers of objects in each category in the aerial and ground view.

drone-based datasets such as VisDrone [88], UAV123 [54], Campus [63].

C. Addendum to the baseline and evaluation

This section highlights the implementation details of our baseline DNN models; see Table 7 and 8. In Section C.3, 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 [18, 89] 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

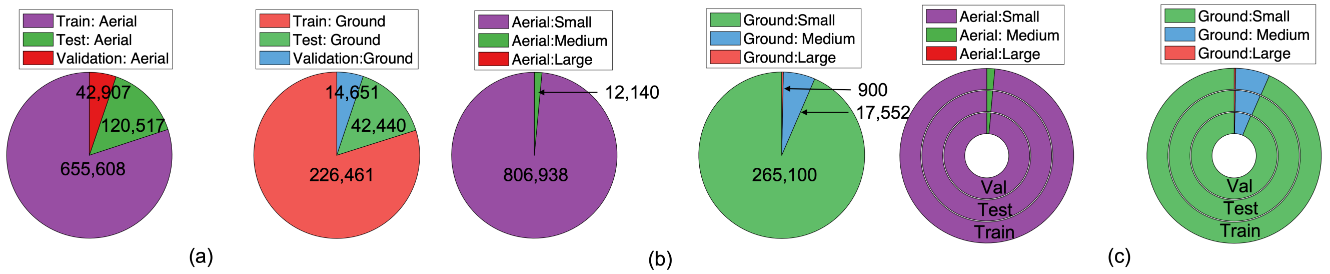


Figure 12. (a) Total number of annotations in train, test, and validation sets of aerial and ground view; (b) number of objects based on their sizes in aerial and ground view, aerial view has no *large* object annotation; (c) percentage of small, medium, and large objects in train, test, and validation sets of aerial and ground view.

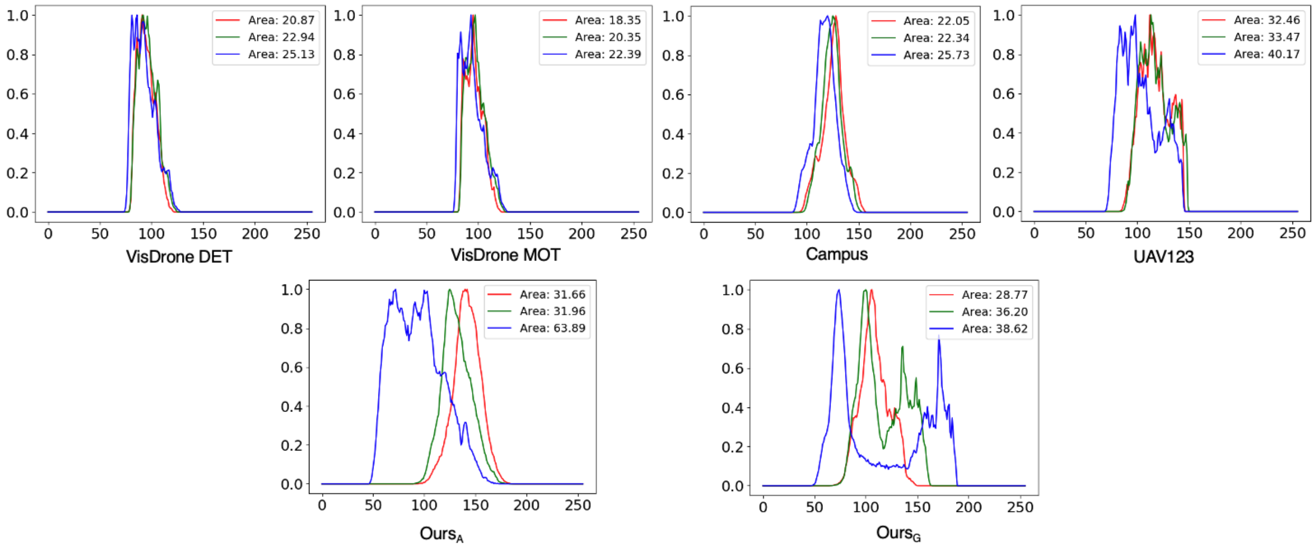


Figure 13. **Color distribution of different datasets.** In the top row, we show the color distribution of VisDrone [88] DET and MOT, the Campus dataset [63], and the UAV123 dataset [21]. VisDrone represents south-east Asian geographies (collected in 14 cities across China) [88]; the Campus dataset represents North American geographies, collected in Stanford University campus [63]; UAV123 represents the Middle East, collected primarily in King Abdullah University of Science and Technology's campus and its surroundings (Kingdom of Saudi Arabia) [54]. In the bottom row, we show the ground and aerial view color distribution of MAVREC.

to train the object detectors specifically for the MAVREC dataset.

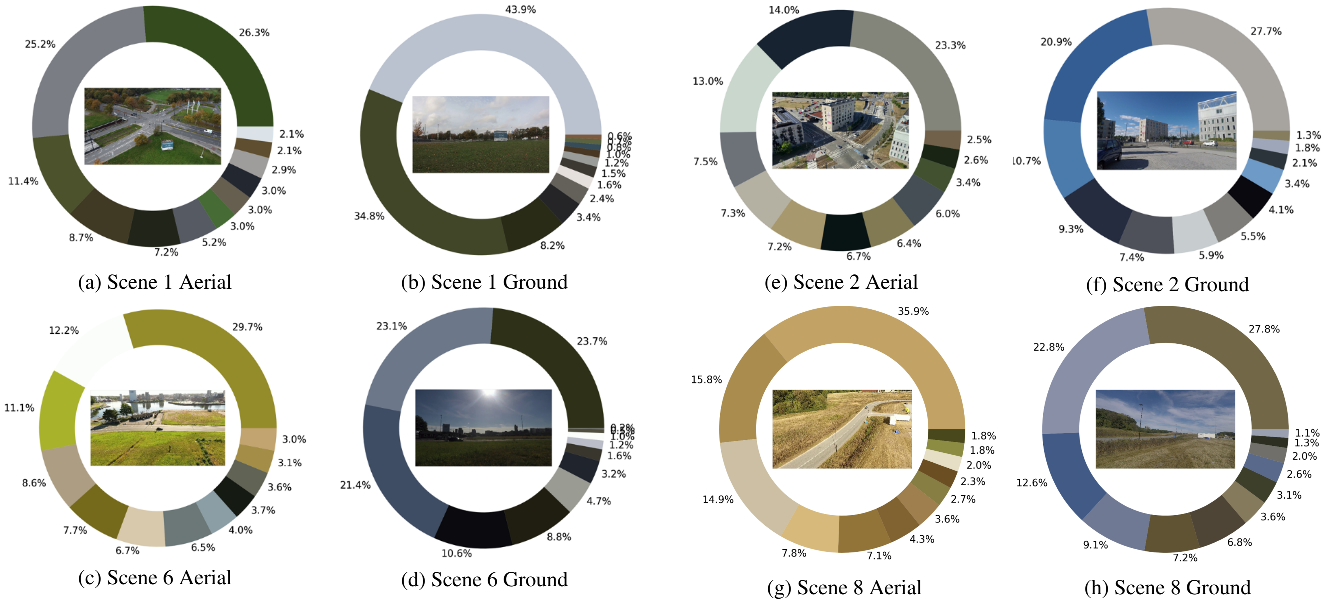
Computing environment. For prototyping, we use a local testbed with an AMD EPYC 7501 32-Core Processor with 2.0GHz speed, 16 GB memory, and 1 Nvidia Tesla V100 GPU with 32 GB on-board memory. For training all the supervised baselines, we use two HPC nodes: (*i*) Node-1: 2x Intel(R) Xeon(R) Gold 6230 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. Evaluation metric

In this section we give brief description of the metric used in our experiments.

C.2.1 Average precision (AP)

Average precision (AP) is a standard metric for information retrieval tasks and is used for object detection and instance segmentation in computer vision. We pause here, and first explain the precision and recall of a model's performance in general. For a given test of predictions (of a model) and the corresponding ground-truth labels, the precision represents the proportion of correct class labels among all predicted positives. The recall represents the proportion of correct positive predictions among all actual positives. For an user-



(a) Scene 1 Aerial

(b) Scene 1 Ground

(e) Scene 2 Aerial

(f) Scene 2 Ground

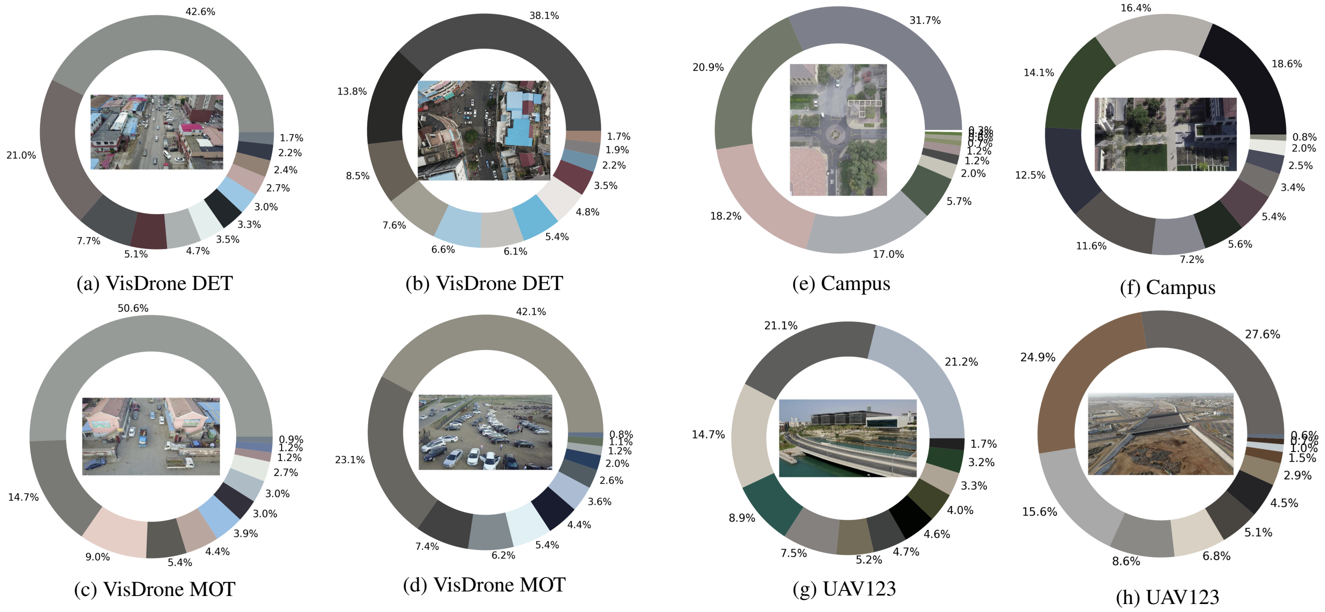
(c) Scene 6 Aerial

(d) Scene 6 Ground

(g) Scene 8 Aerial

(h) Scene 8 Ground

Figure 14. Dominant colors in different sample frames of MAVREC containing both views.



(a) VisDrone DET

(b) VisDrone DET

(e) Campus

(f) Campus

(c) VisDrone MOT

(d) VisDrone MOT

(g) UAV123

(h) UAV123

Figure 15. Most dominant colors in the sample frames of VisDrone DET and MOT [88], the Campus dataset [63], and the UAV123 dataset [54].

defined threshold, $t \in (0, 1]$, denote precision as $P(t)$ and recall as $R(t)$ and are given as follows:

$$P(t) = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{and} \quad R(t) = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where TP, FP, and FN denote true positive, false positive, and false negative, respectively. The accuracy of the model's predictions is quantified by calculating the area under the precision-recall (PR) curve.

In the context of object detection, next, we explain the intersection over union (IoU) metric. IoU describes the closeness of two bounding boxes (predicted and the ground truth) and is given as the ratio of the area of intersection between the predicted box ($A_{\text{Predicted box}}$) and ground truth

Table 7. DNN models used for benchmarking. Note that $1M = 10^6$.

Type	Model	Task	Dataset	Parameters	Optimizer	Platform	Metric
CNN	YoloV7 [73]	Detection	MAVREC	36.5M	SGD-M [57]	PyTorch	mAP
NAS	Yolo-NAS (L) [11]	Detection	MAVREC	51.1M	Adam [33]	PyTorch	mAP
Transformer	DETR [18]	Detection	MAVREC	41M	Adam [33]	PyTorch	mAP
	D-DETR [89]	Detection	MAVREC and VisDrone	41M	Adam [33]	PyTorch	mAP
	OMNI-DETR [74]	Detection	MAVREC and VisDrone	41M	Adam [33]	PyTorch	mAP

Table 8. Hyperparameters used for training each DNN model.

Model	Backbone	Learning Rate	Batch Size	Weight Decay	Queries	Attention Heads	Epochs
YoloV7 [73]	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 [18]	ResNet50 [27]	10^{-4}	2	10^{-4}	100	16	300
D-DETR [89]	ResNet50	2×10^{-4}	2	10^{-4}	900	16	39
OMNI-DETR [74]	ResNet50	10^{-4}	2	10^{-4}	900	16	39

box ($A_{\text{Ground-truth box}}$) to that of their union:

$$\text{IoU} = \frac{A_{\text{Predicted box}} \cap A_{\text{Ground-truth box}}}{A_{\text{Predicted box}} \cup A_{\text{Ground-truth box}}}.$$

Naturally, IoU falls between 0 and 1, where 1 indicates a complete overlap between the two boxes and hence, perfect detection. While 0 indicates no overlap and hence, no detection. A detection box is assigned TP, FP, and FN based on the predicted label compared to the ground truth label and the IoU between the two boxes. In multi-class classification, the model outputs the conditional probability that the bounding box belongs to a certain object class. For a probability confidence threshold, $t \in (0, 1]$, in general, the higher the number of detection, the lower the chances that the missed ground-truth labels, resulting in a higher recall. In contrast, the higher the confidence threshold, the more confident the model is its predictions, and this results in a higher precision. One can generate a PR curve based on different threshold values $t \in (0, 1]$. Finally, the average precision (AP) is defined as the area under the PR curve:

$$AP = \int_{t=0}^{1} p(t)dt.$$

In practice, numerical integration methods are used to approximately calculate this area.

Mean average precision (mAP) is the average AP across all object classes and is defined as follows:

$$\text{mAP} := \tfrac{1}{|C|} \textstyle\sum_{c \in C} AP_c,$$

where C is the set of all classes, $|C|$ is its the cardinality, and AP_c be the AP for a class $c \in C$.

C.2.2 COCO mAP [44]

Our results reported with the COCO mAP which is a cumulative sum of the average of multiple AP calculated at different IoU-thresholds ranging from 0.5 to 0.95 with an increment of 0.05. COCO mAP is the average over 10 IoU levels on all classes.

C.3. Additional baseline results

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

C.3.1 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 aerial-visual perception while aerial-view images are *augmented* with corresponding ground-view images? This motivates us to use mix-up [83] 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

Table 9. Supervised benchmark on aerial view of MAVREC (Validation 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 MAVREC. The top row represents training exclusively on aerial-view samples.

Extra ground view samples	AP	AP$_{50}$	AP$_S$	AP$_M$	**Relative(↑↓) change**
0%	24.9	39.7	27.6	45.3	–
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**
0%	31.3	57.7	34.2	61.2	–
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 10. Supervised benchmark on aerial view of MAVREC (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 MAVREC.

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

labels and reduces the inductive bias [83]. 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 MAVREC; 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, $\lambda \in [0.75, 1]$ 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 9 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, safety, and broader impact

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

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**

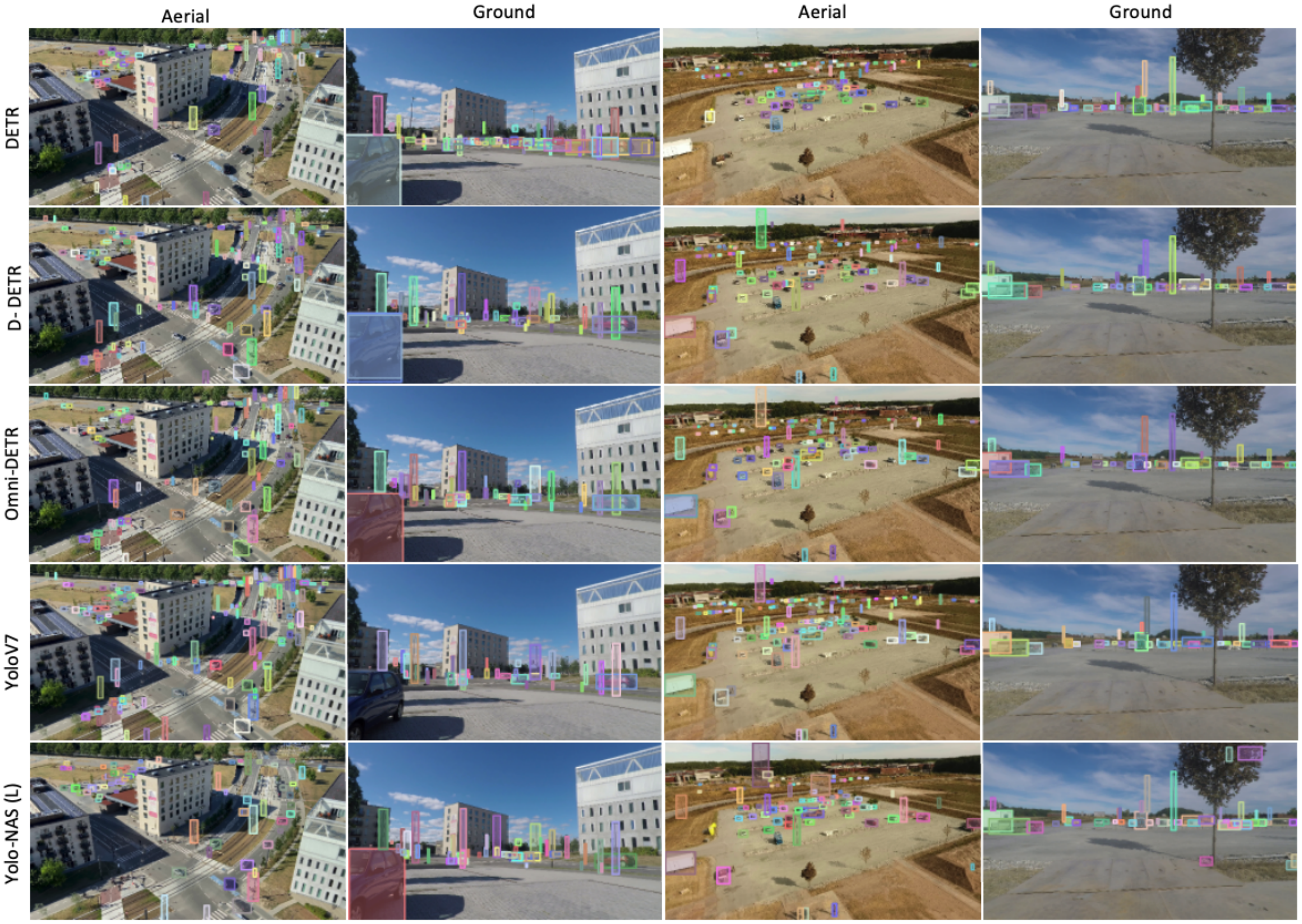


Figure 16. Qualitative inference results of different DNN models on the test set of MAVREC.

any medium or format, as long as the creator is attributed. The license allows MAVREC 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.).

We conduct the recording in public spaces in compliance with the European Union's drone regulations. In Scandinavian countries, video recording falls under surveillance if the recording lasts continuously over 6 hours; our recorded clips are only a few minutes long. Moreover, in crowded intersections, to adhere to the drone-safety protocols, we did not operate drones, instead, we used user-grade handheld cameras from a high riser. As our recordings follow these protocols, the university's legal team confirmed that we do not need additional permissions for our data collection process or publication.

MAVREC is a traffic-centric dataset, with repetitive human activities limited to bicycling, stopping at red traffic lights, and occasionally walking by. The position and distance of the ground and drone cameras do not allow any explicit human recognition. There are many human subjects present in the data, although there are no personal data that can resemble shreds of evidence, reveal identification, or show offensive content. By watching the video clips from the MAVREC, the university's legal experts have concluded that the MAVREC does not have recognizable human subjects and hence does not interfere with privacy. Therefore, MAVREC 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 [15, 52], robotics and instructional videos [78], environmental monitoring [59], heavy industrial infrastructure inspection [13], developing livable and safe communities [6, 30, 86], 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.

D.1. Maintenance plan

The authors are responsible for maintenance and continuous hosting of the dataset on the web. The project lead will assign a research assistant for this purpose. For any queries regarding corrections, annotations and learning algorithm the user can reach the maintenance team at `MAVRECdataset@gmail.com`.

The authors will release the subsequent versions of the dataset to address any reported errors and incorporate proper corrections. The authors will also add annotations if any and delete faulty annotations. The authors will determine the necessity for these updates annually, and subsequently, the latest version will be published on the website along with all previous versions. Retaining access to earlier versions of the dataset would allow the users for reference during their evaluations and verify their results with the proper versions. To differentiate between the versions, each version will be assigned a unique number.

E. Motivation for research challenges on MAVREC dataset

We offer the research community object detection challenges to investigate through a synchronized multi-view dataset. We also encourage the researchers to exploit how a multi-view dataset (with partial annotation) can provide the basis for developing techniques to improve performance in aerial object detection. We highlight a few challenges below:

1. Utilizing the synchronized views and the temporal dimension not provides implicit information and offers a resource-efficient way to enhance performance using unsupervised and semi-supervised techniques. Resource-heavy recording setup or annotations is not required to accomplish this. An advancement in this direction would bring a new era of research in an area increasingly driven by large amounts of data.
2. We underline the need for future research in sampling optimally aerial and ground views. This extends not only to MAVREC but also to other datasets from different domains and modalities. The insights gained from such research could serve as a cornerstone for comprehending more optimal dataset constituents that contribute to DNN's perception. Further, the research community can discover ways to identify samples that foster this understanding and those that hinder it.
3. Recovering objects from one view using the other has multiple motivations: (*i*) training a model on one of the views encourages us to develop techniques that can act as a backup to sensor failure in another view. This can have multiple practical use cases in surveillance and robotics. (*ii*) Recovering objects from an easier learned view can aid learning of a much more difficult view by information transfer between these two views. Encouraging such algorithms would further promote mapping between the views without sophisticated systems such as global navigation satellite/inertial navigation systems (GNSS/INS).