SeaDronesSee: A Maritime Benchmark for Detecting Humans in Open Water

Leon Amadeus Varga,* Benjamin Kiefer,* Martin Messmer,* Andreas Zell

Cognitive Systems Group
University of Tuebingen
Tuebingen, Germany

Email: leon.varga@uni-tuebingen.de, benjamin.kiefer@uni-tuebingen.de,
martin.messmer@uni-tuebingen.de, andreas.zell@uni-tuebingen.de

Abstract

Unmanned Aerial Vehicles (UAVs) are of crucial importance in search and rescue missions in maritime environments due to their flexible and fast operation capabilities. Modern computer vision algorithms are of great interest in aiding such missions. However, they are dependent on large amounts of real-case training data from UAVs, which is only available for traffic scenarios on land. Moreover, current object detection and tracking data sets only provide limited environmental information or none at all, neglecting a valuable source of information. Therefore, this paper introduces a large-scaled visual object detection and tracking benchmark (SeaDronesSee) aiming to bridge the gap from land-based vision systems to sea-based ones. We collect and annotate over 54,000 frames with 400,000 instances captured from various altitudes and viewing angles ranging from 5 to 260 meters and 0 to 90° degrees while providing the respective meta information for altitude, viewing angle and other meta data. We evaluate multiple state-of-the-art computer vision algorithms on this newly established benchmark serving as baselines. We provide an evaluation server where researchers can upload their prediction and compare their results on a central leaderboard [1].
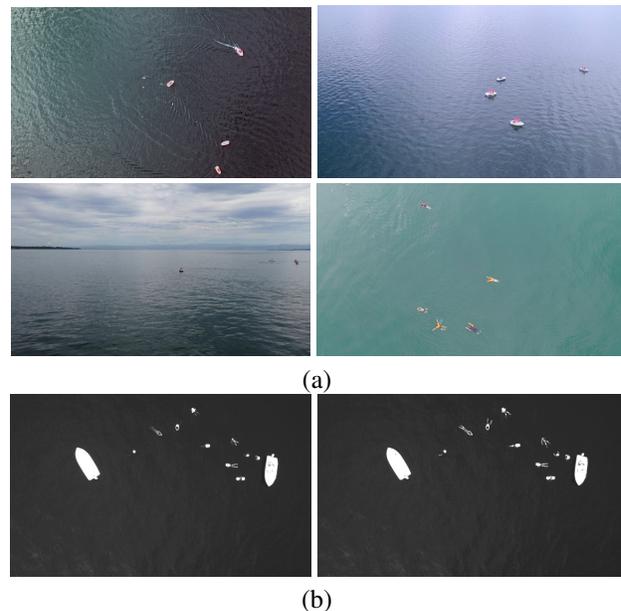
Figure 1. (a) Typical image examples with varying altitudes and angles of view: 250 m, 90°; 50 m, 30°; 10 m, 0° and 20 m, 90° (from top left to bottom right). (b) Examples of the Red Edge (717 nm, left) and Near Infrared (842 nm, right) light spectra of an image captured by the MicaSense RedEdge-MX. Note the glowing appearance of the swimmers.

1. Introduction

Unmanned Aerial Vehicles (UAVs) equipped with cameras have grown into an important asset in a wide range of fields, such as agriculture, delivery, surveillance, and search and rescue (SAR) missions [5, 48, 21]. In particular, UAVs are capable of assisting in SAR missions due to their fast and versatile applicability while providing an overview over the scene [38, 26, 6]. Especially in maritime scenarios, where wide areas need to be quickly overseen and searched, the efficient use of autonomous UAVs is crucial [54]. Among the most challenging issues in this application scenario is the detection, localization, and tracking of people in open water [20, 41]. The small size of people relative to search radii and the variability in viewing angles and altitudes require robust vision-based systems.

Currently, these systems are implemented via data-driven methods such as deep neural networks. These methods depend on large-scale data sets portraying real-case scenarios to obtain realistic imagery statistics. However, there is a great lack of large-scale data sets in maritime environ-

---

*These authors contributed equally to this work. The order of names is determined by coin flipping

[1]The leaderboard, the data set and the code to reproduce our results are available at https://seadronessee.cs.uni-tuebingen.de.

ments. Most data sets captured from UAVs are land-based, often focusing on traffic environments, such as VisDrone [58] and UAVDT [16]. Many of the few data sets that are captured in maritime environments fall in the category of remote sensing, often leveraging satellite-based synthetic aperture radar [12]. All of these are only valuable for ship detection [11] as they don't provide the resolution needed for SAR missions. Furthermore, satellite-based imagery is susceptible to clouds and only provides top-down views. Finally, many current approaches in the maritime setting rely on classical machine learning methods, incapable of dealing with the large number of influencing variables and calling for more elaborate models [44].

This work aims to close the gap between large-scale land-based data sets captured from UAVs to maritime-based data sets. We introduce a large-scale data set of people in open water, called SeaDronesSee. We captured videos and images of swimming probands in open water with various UAVs and cameras. As it is especially critical in SAR missions to detect and track objects from a large distance, we captured the RGB footage with 3840×2160 px to 5456×3632 px resolution. We carefully annotated ground-truth bounding box labels for objects of interest including swimmer, floater (swimmer with life jacket), life jacket, swimmer[†] (person on boat not wearing a life jacket), floater[†] (person on boat wearing a life jacket), and boat.

Moreover, we note that current data sets captured from UAVs only provide very coarse or no meta information at all. We argue that this is a major impediment in the development of multi-modal systems, which take these additional information into account to improve accuracy or speed. Recently, methods that rely on these meta data were proposed. However, they note the lack of large-scaled publicly available data set in that regime (see *e.g.* [27, 51, 36]). Therefore, we provide precise meta information for every frame and image including altitude, camera angle, speed, time, and others.

In maritime settings, the use of multi-spectral cameras with Near Infrared channels to detect humans can be advantageous [20]. For that reason, we also captured multi-spectral images using a MicaSense RedEdge. This enables the development of detectors taking into account the non-visible light spectra Near Infrared (842 nm) and Red Edge (717 nm).

Finally, we provide detailed statistics of the data set and conduct extensive experiments using state-of-the-art models and hereby establish baseline models. These serve as a starting point for our SeaDronesSee benchmark. We release the training and validation sets with complete bounding box ground truth but only the test set's videos/images. The ground truth of the test set is used by the benchmark server to calculate the generalization power of the models. We set up an evaluation web page, where researchers can upload their predictions and opt to publish their results on a central leader board such that transparent comparisons are possible. The benchmark focuses on three tasks: (i) object detection, (ii) single-object tracking and (iii) multi-object tracking, which will be explained in more detail in the subsequent sections. Our main contributions are as follows:

- To the best of our knowledge, SeaDronesSee is the first large annotated UAV-based data set of swimmers in open water. It can be used to further develop detectors and trackers for SAR missions.

- We provide full environmental meta information for every frame making SeaDroneSee the first UAV-based data set of that nature.

- We provide an evaluation server to prevent researches from overfitting and allow for fair comparisons.

- We perform extensive experiments on state-of-the-art object detectors and trackers on our data set.

2. Related Work

In this section, we review major labeled data sets in the field of computer vision from UAVs and in maritime scenarios which are usable for supervised learning models.

2.1. Labeled Data Sets Captured from UAVs

Over the last few years, quite a few data sets captured from UAVs have been published. The most prominent are these that depict traffic situations, such as VisDrone [58] and UAVDT [16]. Both data sets focus on object detection and object tracking in unconstrained environments. Pei *et al.* [43] collect videos (Stanford Drone Dataset) showing traffic participants on campuses (mostly people) for human trajectory prediction usable for object detection. UAV123 [39] is a single-object tracking data set consisting of 123 video sequences with corresponding labels. The clips mainly show traffic scenarios and common objects. Both, Hsieh *et al.* [24] and Mundhenk *et al.* [40] capture a data set showing parking lots for car counting tasks and constrained object detection. Li *et al.* [31] provide a single-object tracking data set showing traffic, wild life and sports scenarios. Collins *et al.* capture a single-object tracking data set showing vehicles on streets in rural areas. Krajewski *et al.* [28] show vehicles on freeways.

Another active area of research focuses on drone-based wildlife detection. Van *et al.* [50] release a data set for the tasks of low-altitude detection and counting of cattle. Ofli *et al.* [42] release the African Savanna data set as part of their crowd-sourced disaster response project.

Object detection	Env.	Platform	Image widths	Altitude	Range	Angle	Range	Other meta
DOTA [52]	cities	satellite	800-20,000	–	–	✗	90°	✗
UAVDT [16]	traffic	UAV	1,024	✗	5-200 m*	✗	$0-90°*$	✗
VisDrone [58]	traffic	UAV	960-2,000	✗	5-200 m*	✗	$0-90°*$	✗
Airbus Ship [2]	maritime	satellite	768	–	–	✗	90°	✗
AU-AIR [10]	traffic	UAV	1,920	✓	5-30 m	✗	$45-90°$	✓
SeaDronesSee	maritime	UAV	3,840-5,456	✓	5-260 m	✓	$0-90°$	✓

Single-object tracking	Env.	#Clips	Frame widths	Altitude	Range	Angle	Range	Other meta
UAV123 [39]	traffic	123	1,280	✗	5-50 m*	✗	$0-90°*$	✓
DTB70 [31]	sports	70	1,280	✗	0-10 m*	✗	$0-90°*$	✗
UAVDT-SOT [16]	traffic	50	1,024	✗	5-200 m*	✗	$0-90°*$	✓
VisDrone [58]	traffic	167	960-2,000	✗	5-200 m*	✗	$0-90°*$	✓
SeaDronesSee	maritime	208	3,840	✓	5-150 m	✓	$0-90°$	✓

Multi-object tracking	Env.	#Frames	Frame widths	Altitude	Range	Angle	Range	Other meta
UAVDT-MOT [16]	traffic	40.7 k	1,024	✗	5-200 m*	✗	$0-90°*$	✓
VisDrone [58]	traffic	40 k	960-2,000	✗	5-200 m*	✗	$0-90°*$	✓
SeaDronesSee	maritime	54 k	3,840	✓	5-150 m	✓	$0-90°$	✓

Table 1. Comparison with the most prominent annotated aerial data sets. 'Altitude' and 'Angle' indicate whether or not there are precise altitude and angle view information available. 'Other meta' refers to time stamps, GPS, and IMU data and in the case of object tracking can also mean attribute information about the sequences. The values with stars have been estimated based on ground truth bounding box sizes and corresponding real world object sizes (for altitude) and qualitative estimation of sample images (for angle). For DOTA and Airbus Ship the range of altitudes is not available because these are satellite-based data sets.

2.2. Labeled Data Sets in Maritime Environments

Many data sets in maritime environments are captured from satellite-based synthetic aperture radar and therefore fall into the remote sensing category. In this category, the airbus ship data set [2] is prominent, featuring 40k images from synthetic aperture radars with instance segmentation labels. Li *et al*. [30] provide a data set of ships with images mainly taken from Google Earth, but also a few UAV-based images. In [52], the authors provide satellite-based images from natural scenes, mainly land-based but also harbors. The most similar to our work is [34]. They also consider the problem of human detection in open water. However, their data mostly contains images close to shores and of swimming pools. Furthermore, it is not publicly available.

2.3. Multi-Modal Data Sets Captured from UAVs

UAVDT [16] provides coarse meta data for their object detection and tracking data: every frame is labeled with altitude information (low, medium, high), angle of view (front-view, side-view, bird-view) and light conditions (day, night, foggy). Wu *et al*. [51] manually label VisDrone after its release with the same annotation information for the object detection track. Mid-Air [19] is a synthetic multi-modal data set with images in nature containing precise altitude, GPS, time, and velocity data but without annotated objects. Blackbird [7] is a real-data indoor data set for agile perception also featuring these meta information. In [35],

street-view images with the same meta data are captured to benchmark appearance-based localization. Bozcan *et al*. [10] release a low-altitude (< 30 m) object detection data set containing images showing a traffic circle and provide meta data such as altitude, GPS, and velocity but exclude the import camera angle information.

Tracking data sets often provide meta data (or attribute information) for the clips. However, in many cases these do not refer to the environmental state in which the image was captured. Instead, they abstractly describe the way in which a clip was captured: UAV123 [39] label their clips with information such as aspect ratio change, background clutter, and fast motion, but do not provide frame-by-frame meta data. The same observation can be made for the tracking track of VisDrone [18]. See Table 1 for an overview of annotated aerial data sets.

3. Data Set Generation

We gathered the footage on several days to obtain variance in light conditions. Taking into account safety and environmental regulations, we asked over 20 test subjects to be recorded in open water. Boats transported the subjects to the area of interest, where quadcopters were launched at a safe distance from the swimmers. At the same time, the fixed-wing UAV Trinity F90+ was launched from the shore. We used waypoints to ensure a strict flight schedule to maximize data collection efficiency. Care was taken to maintain

Camera	Resolution	Video
Hasselblad L1D-20c	3,840×2,160	30 fps
MicaSense RedEdge-MX	1,280× 960	✕
Sony UMC-R10C	5,456×3,632	✕
Zenmuse X5	3,840×2,160	30 fps
Zenmuse XT2	3,840×2,160	30 fps

Table 2. Overview of used cameras.

Data	Unit	Min. value	Max.value
Time since start	ms	0	∞
Date and Time	ISO 8601	–	–
Latitude	degrees	−90	+90
Longitude	degrees	−90	+90
Altitude	meters	0	∞
Gimbal pitch	degrees	0	90
UAV roll	degrees	−90	+90
UAV pitch	degrees	−90	+90
UAV yaw	degrees	−180	+180
x-axis speed	m/s	0	∞
y-axis speed	m/s	0	∞
z-axis speed	m/s	0	∞

Table 3. Meta data that comes with every image/frame.

a strict vertical separation at all times. Subjects were free to wear life jackets, of which we provided several differently colored pieces (see also Figure 2).

To diminish the effect of camera biases within the data set, we used multiple cameras, as listed in Table 2, mounted to the following drones: DJI Matrice 100, DJI Matrice 210, DJI Mavic 2 Pro, and a Quantum Systems Trinity F90+. With the video cameras, we captured videos at 30 fps. For the object detection task, we extract at most three frames per second of these videos to avoid having redundant occurrences of frames. See Section 4 for information on the distribution of images with respect to different cameras.

Lastly, we captured top-down looking multi-spectral imagery at 1 fps. We used a MicaSense RedEdge-MX, which records five wavelengths (475 nm, 560 nm, 668 nm, 717 nm, 842 nm). Therefore, in addition to the RGB channels, the recordings also contain a RedEdge and a Near Infrared channel. The camera was referenced with a white reference before each flight. As the RedEdge-MX captures every band individually, we merge the bands using the development kit provided by MicaSense.

3.1. Meta Data Collection

Accompanied with every frame there is a meta stamp, that is logged at 10 hertz. To align the video data (30 fps) and the time stamps, a nearest neighbor method was performed. The data in Table 3 is logged and provided for every image/frame read from the onboard clock, barometer,

IMU and GPS sensor, and the gimbal, respectively.

Note that $\alpha = 90°$ corresponds to a top-down view, and $\alpha = 0°$ to a horizontally facing camera. The date format is given in the extended form of ISO 8601. Furthermore, note that the UAV roll/pitch/yaw-angles are of minor importance for meta-data-aware vision-based methods as the onboard gimbal filters out movement by the drone such that the camera pitch angle is roughly constant if it is not intentionally changed [25]. Note that the gimbal yaw angle is not included, as we fix it to coincide with the UAV's yaw angle.

We need to emphasize that the meta values lie within the error thresholds introduced by the different sensors, but an extended analysis is beyond the scope of this paper (see *e.g.* [61, 1, 29] for an overview).

3.2. Annotation Method

Using the non-commercial labeling tool DarkLabel [3], we manually and carefully annotated all provided images and frames with the categories swimmer (person in water without life jacket), floater (person in water with life jacket), life jacket, swimmer[†] (person on boat without life jacket), floater[†] (person on boat with life jacket), and boats. We note that it is not sufficient to infer the class floater by the location from swimmer and life jacket as this can be highly ambiguous. Subsequently, all annotations were checked by experts in aerial vision. We choose these classes as they are the hardest and most critical to detect in SAR missions. Furthermore, we annotated regions with other objects as ignored regions, such as boats on land. Moreover, the data set also covers unlabeled objects, which may not be of interest, like driftwood, birds or the coast such that detectors can be robust to distinguish from those objects. Our guidelines for the annotation are described in the appendix. See Figure 2 for examples of objects.

3.3. Data Set Split

Object Detection

To ensure that the training, validation, and testing set have similar statistics, we roughly balance them such that the respective subsets have similar distributions with respect to altitude and angle of view, two of the most important factors of appearance changes. Of the individual images, we randomly select $4/7$ and add it to the training set, add $1/7$ to the validation set and another $2/7$ to the testing set. In addition to the individual images, we randomly cut every video into three parts of length $4/7$, $1/7$, and $2/7$ of the original length and add every 10-th frame of the respective parts to the training, validation, and testing set. This is done to avoid having subsequent frames in the training and testing set such that a realistic evaluation is possible. We release the training and validation set with all annotations and the testing set's images, but withhold its annotations. Evalu-
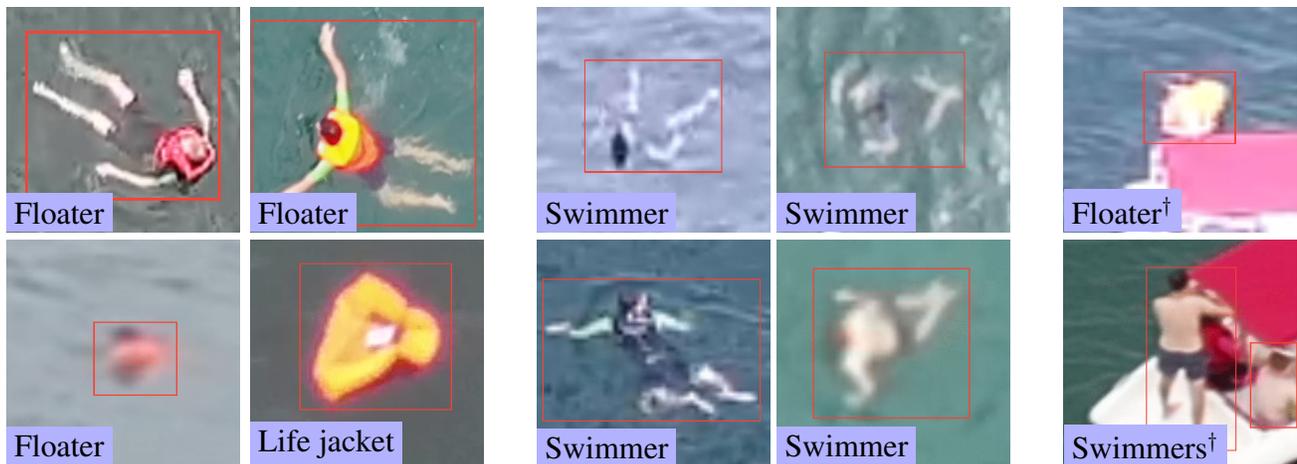
Figure 2. Examples of objects. Note that these examples are crops from high-resolution images. However, as the objects are small and the images taken from high altitudes, they appear blurry.

ation will be available via an evaluation server, where the predictions on the test set can be uploaded.

Object Tracking

Similarly, we take $4/7$ of our recorded clips as the training clips, $1/7$ as the validation clips and $2/7$ as the testing clips. As for the object detection task, we withhold the annotations for the testing set and provide an evaluation server.

4. Data Set Tasks

There are many works on UAV-based maritime SAR missions, focusing on unified frameworks describing the process of how to search and rescue people [38, 20, 33, 34, 45, 47, 22]. These works answer questions corresponding to path planning, autonomous navigation and efficient signal transmission. Most of them rely on RGB sensors and detection and tracking algorithms to actually find people of interest. This commonality motivates us to extract the specific tasks of object detection and tracking, which pose some of the most challenging issues in this application scenario.

Maritime environments from a UAV's perspective are difficult for a variety of reasons: Reflective regions and shadows resulting from different cardinal points (such as in Fig. 1) that could lead to false positives or negatives; people may be hardly visible or occluded by waves or sea foam (see Supplementary material); typically large areas are overseen such that objects are particularly small [38]. We note that these factors are on top of general UAV-related detection difficulties.

Now, we proceed to describe the specific tasks.

4.1. Object Detection

There are 5,630 images (training: 2,975; validation: 859; testing: 1,796). See Figure 3 for the distribution of images/frames with respect to cameras and the class distribu-

tion. We recorded most of the images with the L1D-20c and UMC-R10C, having the highest resolution. Having the lowest resolution, we recorded only 432 images with the RedEdge-MX. Note, for the Object Detection Task only the RGB-channels of the multi-spectral images are used to support a uniform data structure.

Furthermore, the class distribution is slightly skewed towards the class 'boat', since safety precautions require boats to be nearby. We emphasize that this bias can easily be diminished by blackening the respective regions, as is common for areas which are not of interest or undesired (such as boats here; see *e.g.* [16]). Right after that, swimmers with life jacket are the most common objects. We argue that this scenario is very often encountered in SAR missions. This type of class often is easier to detect than just swimmer as life jackets mostly are of contrasting color, such as red or orange (see Fig. 2 and Table 4). However, as it is also a likely scenario to search for swimmers without life jacket, we included a considerable amount. There are also several different manifestations/visual appearances of that class which is why we recorded and annotated swimmers with and without adequate swimwear (such as wet suit). To be able to discriminate between humans in water and humans on boats, we also annotated humans on boats (with and without life jackets). Lastly, we annotated a small amount of life jackets only. However, we note that the discrimination between life jackets and humans in life jackets can become visually ambiguous, especially in higher altitudes. See also Fig. 2.

Figure 4 shows the distribution of images with respect to the altitude and viewing angle they were captured at. Roughly 50% of the images were recorded below 50 m because lower altitudes allow for the whole range of available viewing angles $(0 - 90°)$. That is, to cover all viewing angles, more images at these altitudes had to be taken. On the other hand, there are many images facing downwards $(90°)$, because images taken at greater altitudes tend to face down-
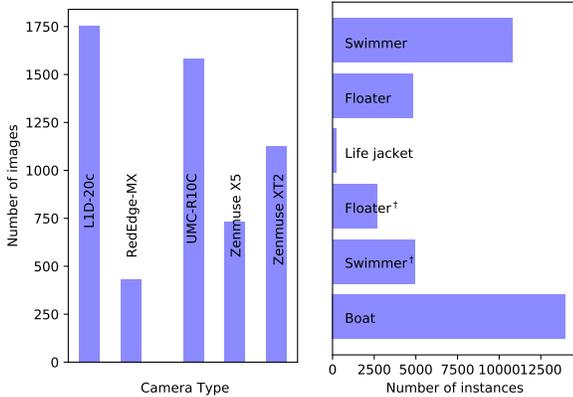
Figure 3. Distribution of training images over camera types (left) and distribution of objects over classes (right).
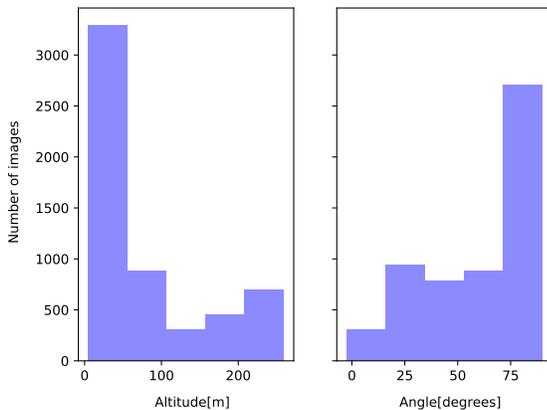

Figure 4. Distribution of images over altitudes (left) and angles (right), respectively.

wards since acute angles yield image areas with tiny pixel density, which is unsuitable for object detection. Nevertheless, every altitude and angle interval is sufficiently represented.

4.2. Single-Object Tracking

We provide 208 short clips (>4 seconds) with a total of 393,295 frames (counting the duplicates), including all available objects labeled. We randomly split the sequences into 58 training, 70 validation and 80 testing sequences. We do not support long-term tracking. The altitude and angle distributions are similar to these in the object detection section since the origin of the images of the object detection task is the same.

4.3. Multi-Object Tracking

We provide 22 clips with a total of 54,105 frames and 403,192 annotated instances, the average consists of 2,460 frames. We differentiate between two use-cases. In the first task, only the persons in water (floaters and swimmers) are tracked, it is called MOT-Swimmer. In the second task, all objects in water are tracked (also the boats, but not people

on boats), called MOT-All-Objects-In-Water. In both tasks, all objects are grouped into one class. The data set split is performed as described in section 3.3.

4.4. Multi-Spectral Footage

Along with the data for the three tasks, we provide multi-spectral images. We supply annotations for all channels of these recordings, but only the RGB-channels are currently part of the Object Detection Task. There are 432 images with 1,901 instances. See Figure 1 for an example of the individual bands.

5. Evaluations

We evaluate current state-of-the-art object detectors and object trackers on SeaDronesSee. All experiments can be reproduced by using our provided code available on the evaluation server. Furthermore, we refer the reader to the Supplementary Material for the exact form and uploading requirements.

5.1. Object Detection

The used detectors can be split into two groups. The first group consists of two-stage detectors, which are mainly built on Faster R-CNN [23] and its improvements. Built for optimal accuracy, these models often lack the inference speed needed for real-time employment, especially on embedded hardware, which can be a vital use-case in UAV-based SAR missions. For that reason, we also evaluate on one-stage detectors. In particular, we perform experiments with the best performing single-model (no ensemble) from the workshop report [60]: a Faster R-CNN with a ResNeXt-101 64-4d [53] backbone with P6 removed. For large one-stage detectors, we take the recent CenterNet [57]. To further test an object detector in real-time scenarios, we choose the current best model family on the COCO test-dev according to [4], i.e. EfficientDet [49], and take the smallest model, $D0$, which can run in real-time on embedded hardware, such as the Nvidia Xavier [27]. We refer the reader to the appendix for the exact parameter configurations and training configurations of the individual models.

Similar to the VisDrone benchmark [58], we evaluate detectors according to the COCO json-format [32], i.e. average precision at certain intersection-over-union-thresholds. More specifically, we use $AP=AP^{IoU=0.5:0.05:0.95}$, $AP_{50} = AP^{IoU=0.5}$ and $AP_{75} = AP^{IoU=0.75}$. Furthermore, we evaluate the maximum recalls for at most 1 and 10 given detections, respectively, denoted $AR_1 = AR^{max=1}$, and $AR_{10} = AR^{max=10}$. All these metrics are averaged over all categories (except for "ignored region"). We furthermore provide the class-wise average precisions. Moreover, similar to [27], we report AP_{50}-results on different equidistant levels of altitudes 'low' = 5-56 m (L), 'low-medium' = 55-106 m (LM), 'medium' = 106-157 m (M), 'medium-high'

Model	AP	AP_{50}	AP_{75}	AR_1	AR_{10}	S	F	$S^\dagger$	$F^\dagger$	B	LJ	FPS
F. ResNeXt-101-FPN [53]	30.4	54.7	29.7	18.6	42.6	78.1	82.4	25.9	44.3	96.7	0.6	2
F. ResNet-50-FPN [23]	14.2	30.1	7.2	6.4	17.7	24.6	54.1	4.9	7.5	89.2	0.3	14
CenterNet-Hourglass104 [57]	25.6	50.3	22.2	17.7	40.1	65.1	73.6	19.1	48.1	95.8	0.3	6
CenterNet-ResNet101 [57]	15.1	36.4	10.8	9.6	21.4	16.8	39.8	0.8	1.7	74.3	0	22
CenterNet-ResNet18 [57]	9.9	21.8	9.0	7.2	19.7	20.9	21.9	2.6	3.3	81.9	0.4	78
EfficientDet–$D0$ [49]	20.8	37.1	20.6	11.5	29.1	65.3	55.1	3.1	3.3	95.5	0.1	26

Table 4. Average precision results for several baseline models. The right part contains AP_{50}–values for each class individually. All reported FPS numbers are obtained on a single Nvidia RTX 2080 Ti. The abbreviation 'F.' stands for Faster R-CNN. For visualization purposes, the classes are abbreviated as swimmer(†) → S(†), floater(†) → F(†), boat → B, life jacket → LJ.

= 157-208 m (MH), and 'high' = 208-259 m (H). To measure the universal cross-domain performance, we report the average over these domains, denoted AP_{50}^{avg}. Similarly, we report AP_{50}-results for different angles of view: 'acute' = 7-23° (A), 'acute-medium' = 23-40° (AM), 'medium' = 40-56° (M), 'medium-right' = 56-73° (MR), and 'right' = 73-90° (R). Ultimately, it is the goal to have robust detectors across all domains uniformly, which is better measured by the latter metrics.

Table 4 shows the results for all object detection models. As expected, the large Faster R-CNN with ResNeXt-101 64-4d backbone performs best, closely followed by CenterNet-Hourglass104. Medium-sized networks, such as the ResNet-50-FPN, and fast networks, such as CenterNet-ResNet18 and EfficientDet-$D0$, expectedly perform worse. However, the latter can run in real-time on an Nvidia Xavier [27]. Swimmers are detected significantly worse than floaters by most detectors. Notably, life jackets are very hard to detect since from a far distance these are easily confused with swimmers† (see Fig. 2). Since there is a heavy class imbalance with many fewer life jackets, detectors are biased towards floaters.

Table 5 and 6 show the performances for different altitudes and angles, respectively. These evaluations help assess the strength and weaknesses of individual models. For example, although ResNeXt-101-FPN performs overall better than Hourglass104 in AP_{50} (54.7 vs. 50.3), the latter is better in the domain of medium angles (45.2 vs. 49.7). Furthermore, the great performance discrepancy between CenterNet-ResNet101 and CenterNet-ResNet18 in AP_{50} (36.4 vs. 21.8) vanishes when averaged over angle domains (23.8 vs. 23.1 AP_{50}^{avg}) possibly indicating ResNet101's bias towards specific angle domains.

5.2. Single-Object Tracking

Like VisDrone [59], we provide the success and precision curves for single-object tracking and compare models based on a single number, the success score. As comparison trackers, we choose the DiMP family (DiMP50, DiMP18, PrDiMP50, PrDiMP18) [9, 14] and Atom [13] because they were the foundation of many of the submitted trackers to the last VisDrone workshop [18].

Model	L	LM	M	MH	H	AP_{50}^{avg}
ResNeXt-101-FPN	56.8	54.6	49.2	65	78.3	60.8
ResNet-50-FPN	32.8	29.8	23.5	40.5	48.9	35.1
Hourglass104	50.6	52.0	47.5	64.9	73.2	57.6
ResNet101	20.2	30.4	24.1	35.1	38.0	29.6
ResNet18	23.8	20.3	19.2	29.3	31.9	24.9
$D0$	39.6	38.0	30.4	42.5	54.5	41.0

Table 5. Results on different altitude-domains. E.g. ResNeXt's AP_{50} performance in low-medium (LM) altitudes is 54.6 AP_{50}.

Model	A	AM	M	MR	R	AP_{50}^{avg}
ResNeXt101-FPN	68.3	55.1	45.2	63.6	51.5	56.7
ResNet50-FPN	32.8	35.5	32.7	35.7	27.6	32.9
Hourglass104	66.4	42.1	49.7	58.7	46.9	52.76
ResNet101	7.4	35.8	20.5	33.6	21.7	23.8
ResNet18	9.6	29.5	26.3	27.9	22.1	23.1
$D0$	26.9	47.0	40.5	40.3	36.8	38.3

Table 6. Results on different angle-domains. For example, ResNeXt's AP_{50} performance in medium-right (MR) angles (57-73°) is 63.6 AP_{50}.
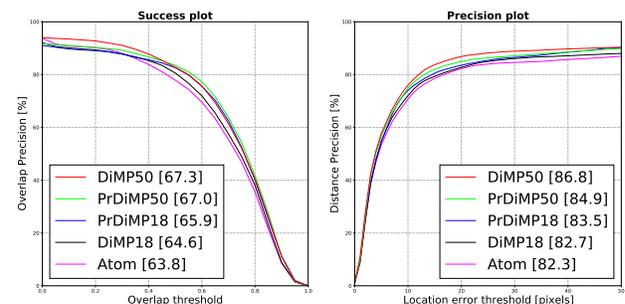

Figure 5. Success and precision plots for single-object tracking task (best viewed in color).

Figure 5 shows that the PrDiMP- and DiMP-family expectedly outperform the older Atom tracker in both, success and precision. Surprisingly, PrDiMP50 slightly trails the accuracy of its predecessor DiMP50. Furthermore, all trackers' performances on SeaDronesSee are similar or worse than on UAV123 (*e.g.* Atom with 65.0 success) [9, 14, 13], for which they were heavily optimized. We argue that in SeaDronesSee there is still room for improvement, especially considering that the clips feature precise meta information that may be helpful for tracking. Furthermore, in our experiments, the faster trackers DiMP18 and Atom run at approximately 27.1 fps on an Nvidia RTX 2080 Ti. How-

Model	MOTA	IDF1	MOTP	MT	ML	FP	FN	Recall	Prcn	ID Sw.	Frag
FairMOT-D34 [56]	39.0	44.8	23.6	17	17	3,604	9,445	57.2	77.8	307	1,687
FairMOT-R34 [56]	15.2	27.6	33.7	6	37	2,502	12,592	30.1	68.4	181	807
Tracktor++ [8]	55.0	69.6	25.6	62	4	7,271	3,550	85.5	74.2	165	347

Table 7. Multi-Object Tracking evaluation results for the **Swimmer** task.

Model	MOTA	IDF1	MOTP	MT	ML	FP	FN	Recall	Prcn	ID Sw.	Frag
FairMOT-D34 [56]	36.5	43.8	20.9	28	49	3,788	20,867	47.2	83.1	447	1,599
FairMOT-R34 [56]	30.5	40.8	27.3	29	127	4,401	28,999	40.2	81.6	285	1,588
Tracktor++ [8]	71.9	80.5	20.1	123	5	7,741	5,496	88.5	84.5	192	438

Table 8. Multi-Object Tracking evaluation results for the **All-Objects-In-Water** task.

Model	L	LM	M	MH	H	AP_{50}^{avg}
F. ResNet-50-FPN	**32.8**	29.8	23.5	40.5	**48.9**	35.1
5×Altitude@3[27]	**32.8**	**29.9**	**26.2**	41.5	**48.9**	**35.9**
Model	A	AM	M	MR	R	AP_{50}^{avg}
F. ResNet-50-FPN	32.8	35.5	32.7	35.7	27.6	32.9
5×Angle@3[27]	**42.0**	**35.5**	**39.3**	**35.7**	**27.7**	**36.0**

Table 9. Results on different altitude- and angle-domains.

ever, we note that they are not capable of running in real-time on embedded hardware, a use-case especially important for UAV-based SAR missions.

5.3. Multi-Object Tracking

We use a similar evaluation protocol as the MOT benchmark [37]. That is, we report results for Multiple Object Tracking Accuracy (MOTA), Identification F1 Score (IDF1), Multiple Object Tracking Precision (MOTP), number of false positives (FP), number of false negatives (FN), recall (R), precision (P), ID switches (ID sw.), fragmentation occurrences (Frag). We refer the reader to [46] or the appendix for a thorough description of the metrics.
We train and evaluate FairMOT [56], a popular tracker, which is the base of many trackers submitted to the challenge [17]. FairMOT-D34 employs a DLA34 [55] as its backbone while FairMOT-R34 makes use of a ResNet34. Another SOTA tracker is Tracktor++ [8], which we also use for our experiments. It performed well on the MOT20 [15] challenge and is conceptually simple.
Surprisingly, Tracktor++ was better than FairMOT in both tasks. One reason for this may be the used detector. Tracktor++ utilizes a Faster-R-CNN with a ResNet50 backbone. In contrast, FairMOT is using a CenterNet with a DLA34 and a ResNet34 backbone, respectively.

5.4. Meta-Data-Aware Object Detector

Developing meta-data-aware object detectors is difficult since there are no large-scale data sets to evaluate their performances. However, some works provide promising preliminary results using this metadata [51, 36, 27]. We provide an initial baseline from [27] incorporating the meta data. We evaluate the performances of 5×Altitude@3- and

5×Angle@3-experts, which are constructed on top of a Faster R-CNN with ResNet-50-FPN, respectively. Essentially, these experts make use of meta-data by allowing the features to adapt to their responsible specific environmental domains.
As Table 9 shows, meta data can enhance the accuracy of an object detector considerably. For example, 5×Angle@3 outperforms its ResNet-50-FPN baseline by 3.1 AP_{50}^{avg} while running at the same inference speed. The improvements are especially significant for underrepresented domains, such as +9.2 and +6.4 AP_{50}^{avg} for the acute angle (A) and the medium angle (M), respectively, which are underrepresented as can be seen from Fig. 4.

6. Conclusions

This work serves as an introductory benchmark in UAV-based computer vision problems in maritime scenarios. We build the first large scaled-data set for detecting and tracking humans in open water. Furthermore, it is the first large-scaled benchmark providing full environmental information for every frame, offering great opportunities in the so-far restricted area of multi-modal object detection and tracking. We offer three challenges, object detection, single-object tracking, and multi-object tracking by providing an evaluation server. We hope that the development of meta-data-aware object detectors and trackers can be accelerated by means of this benchmark. Moreover, we provide multi-spectral imagery for detecting humans in open water. These images are very promising in maritime scenarios, having the ability to capture wavelengths, which set apart objects from the water background.

References

[1] Aerial data accuracy – an experiment comparing 4 drone approaches. `https://www.sitemark.com/blog/accuracy`. Accessed: 2021-03-01.

[2] Airbus Ship Detection Challenge. `https://www.kaggle.com/c/airbus-ship-detection`. Accessed: 2021-03-01.

[3] Darklabel video/image labeling and annotation tool. `https://github.com/darkpgmr/DarkLabel`. Accessed: 2020-08-31.

[4] Object Detection on COCO test-dev. `https://paperswithcode.com/sota/object-detection-on-coco`. Accessed: 2021-03-01.

[5] Telmo Adão, Jonáš Hruška, Luís Pádua, José Bessa, Emanuel Peres, Raul Morais, and Joaquim Joao Sousa. Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry. *Remote Sensing*, 9(11):1110, 2017.

[6] Antonio Albanese, Vincenzo Sciancalepore, and Xavier Costa-Pérez. Sardo: An automated search-and-rescue drone-based solution for victims localization. *arXiv preprint arXiv:2003.05819*, 2020.

[7] Amado Antonini, Winter Guerra, Varun Murali, Thomas Sayre-McCord, and Sertac Karaman. The blackbird dataset: A large-scale dataset for uav perception in aggressive flight. In *International Symposium on Experimental Robotics*, pages 130–139. Springer, 2018.

[8] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[9] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6182–6191, 2019.

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

[11] Christina Corbane, Laurent Najman, Emilien Pecoul, Laurent Demagistri, and Michel Petit. A complete processing chain for ship detection using optical satellite imagery. *International Journal of Remote Sensing*, 31(22):5837–5854, 2010.

[12] DJ Crisp. The state-of-the-art in ship detection in synthetic aperture radar imagery. defence science and technology organization (dsto). *Information Science Laboratory, Research Report No. DSTO-RR-0272*, 2004.

[13] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4660–4669, 2019.

[14] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2020.

[15] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, Laura Leal-Taixé, and Taix´ Taixé. MOT20: A benchmark for multi object tracking in crowded scenes. Technical report.

[16] 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 (ECCV)*, pages 370–386, 2018.

[17] Heng Fan, Dawei Du, Longyin Wen, Pengfei Zhu, Qinghua Hu, Haibin Ling, Mubarak Shah, Junwen Pan, Arne Schumann, Bin Dong, et al. Visdrone-mot2020: The vision meets drone multiple object tracking challenge results. In *European Conference on Computer Vision*, pages 713–727. Springer, 2020.

[18] Heng Fan, Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Haibin Ling, Mubarak Shah, Biao Wang, Bin Dong, Di Yuan, et al. Visdrone-sot2020: The vision meets drone single object tracking challenge results. In *European Conference on Computer Vision*, pages 728–749. Springer, 2020.

[19] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, June 2019.

[20] Antonio-Javier Gallego, Antonio Pertusa, Pablo Gil, and Robert B Fisher. Detection of bodies in maritime rescue operations using unmanned aerial vehicles with multispectral cameras. *Journal of Field Robotics*, 36(4):782–796, 2019.

[21] Ruben Geraldes, Artur Goncalves, Tin Lai, Mathias Villerabel, Wenlong Deng, Ana Salta, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Uav-based situational awareness system using deep learning. *IEEE Access*, 7:122583–122594, 2019.

[22] Siti Nur Alidda Mohd Ghazali, Hardy Azmir Anuar, Syed Nasir Alsagoff Syed Zakaria, and Zaharin Yusoff. Determining position of target subjects in maritime search and rescue (msar) operations using rotary wing unmanned aerial vehicles (uavs). In *2016 International Conference on Information and Communication Technology (ICICTM)*, pages 1–4. IEEE, 2016.

[23] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

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

[25] Karol Jedrasiak, Damian Bereska, and Aleksander Nawrat. The prototype of gyro-stabilized uav gimbal for day-night surveillance. In *Advanced technologies for intelligent systems of national border security*, pages 107–115. Springer, 2013.

[26] Yunus Karaca, Mustafa Cicek, Ozgur Tatli, Aynur Sahin, Sinan Pasli, Muhammed Fatih Beser, and Suleyman Turedi.

The potential use of unmanned aircraft systems (drones) in mountain search and rescue operations. *The American journal of emergency medicine*, 36(4):583–588, 2018.

[27] Benjamin Kiefer, Martin Messmer, and Andreas Zell. Leveraging domain labels for object detection from uavs. *arXiv preprint arXiv:2101.12677*, 2021.

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

[29] David L Kulhavy, I Hung, Daniel Unger, Yanli Zhang, et al. Accuracy assessment on drone measured heights at different height levels. 2017.

[30] Qingpeng Li, Lichao Mou, Qingjie Liu, Yunhong Wang, and Xiao Xiang Zhu. Hsf-net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(12):7147–7161, 2018.

[31] 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.

[32] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[33] Eleftherios Lvsouras and Antonios Gasteratos. A new method to combine detection and tracking algorithms for fast and accurate human localization in uav-based sar operations. In *2020 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 1688–1696. IEEE, 2020.

[34] Eleftherios Lygouras, Nicholas Santavas, Anastasios Taitzoglou, Konstantinos Tarchanidis, Athanasios Mitropoulos, and Antonios Gasteratos. Unsupervised human detection with an embedded vision system on a fully autonomous uav for search and rescue operations. *Sensors*, 19(16):3542, 2019.

[35] András L Majdik, Charles Till, and Davide Scaramuzza. The zurich urban micro aerial vehicle dataset. *The International Journal of Robotics Research*, 36(3):269–273, 2017.

[36] Martin Messmer, Benjamin Kiefer, and Andreas Zell. Gaining scale invariance in uav bird's eye view object detection by adaptive resizing. *arXiv preprint arXiv:2101.12694*, 2021.

[37] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[38] Balmukund Mishra, Deepak Garg, Pratik Narang, and Vipul Mishra. Drone-surveillance for search and rescue in natural disaster. *Computer Communications*, 156:1–10, 2020.

[39] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *European conference on computer vision*, pages 445–461. Springer, 2016.

[40] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European Conference on Computer Vision*, pages 785–800. Springer, 2016.

[41] Imen Nasr, Meriem Chekir, and Hichem Besbes. Shipwrecked victims localization and tracking using uavs. In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pages 1344–1348. IEEE, 2019.

[42] Ferda Ofli, Patrick Meier, Muhammad Imran, Carlos Castillo, Devis Tuia, Nicolas Rey, Julien Briant, Pauline Millet, Friedrich Reinhard, Matthew Parkan, et al. Combining human computing and machine learning to make sense of big (aerial) data for disaster response. *Big data*, 4(1):47–59, 2016.

[43] Zhao Pei, Xiaoning Qi, Yanning Zhang, Miao Ma, and Yee-Hong Yang. Human trajectory prediction in crowded scene using social-affinity long short-term memory. *Pattern Recognition*, 93:273–282, 2019.

[44] Dilip K Prasad, Huixu Dong, Deepu Rajan, and Chai Quek. Are object detection assessment criteria ready for maritime computer vision? *IEEE Transactions on Intelligent Transportation Systems*, 21(12):5295–5304, 2019.

[45] Jorge Peña Queralta, Jenni Raitoharju, Tuan Nguyen Gia, Nikolaos Passalis, and Tomi Westerlund. Autosos: Towards multi-uav systems supporting maritime search and rescue with lightweight ai and edge computing. *arXiv preprint arXiv:2005.03409*, 2020.

[46] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016.

[47] William Roberts, Kelly Griendling, Anthony Gray, and D Mavris. Unmanned vehicle collaboration research environment for maritime search and rescue. In *30th Congress of the International Council of the Aeronautical Sciences*. International Council of the Aeronautical Sciences (ICAS) Bonn, Germany, 2016.

[48] Khin Thida San, Sun Ju Mun, Yeong Hun Choe, and Yoon Seok Chang. Uav delivery monitoring system. In *MATEC Web of Conferences*, volume 151, page 04011. EDP Sciences, 2018.

[49] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[50] Jan C van Gemert, Camiel R Verschoor, Pascal Mettes, Kitso Epema, Lian Pin Koh, and Serge Wich. Nature conservation drones for automatic localization and counting of animals. In *European Conference on Computer Vision*, pages 255–270. Springer, 2014.

[51] Zhenyu Wu, Karthik Suresh, Priya Narayanan, Hongyu Xu, Heesung Kwon, and Zhangyang Wang. Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1201–1210, 2019.

[52] 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 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018.

[53] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[54] SP Yeong, LM King, and SS Dol. A review on marine search and rescue operations using unmanned aerial vehicles. *International Journal of Marine and Environmental Sciences*, 9(2):396–399, 2015.

[55] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018.

[56] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv e-prints*, pages arXiv–2004, 2020.

[57] Xingyi Zhou, Dequan Wang, and Philipp Kr"ahenb"uhl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019.

[58] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.

[59] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, and Haibin Ling. Vision meets drones: Past, present and future. *arXiv preprint arXiv:2001.06303*, 2020.

[60] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Haibin Ling, Qinghua Hu, Qinqin Nie, Hao Cheng, Chenfeng Liu, Xiaoyu Liu, et al. Visdrone-det2018: The vision meets drone object detection in image challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[61] F Zimmermann, C Eling, L Klingbeil, and H Kuhlmann. Precise positioning of uavs-dealing with challenging rtk-gps measurement conditions during automated uav flights. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2017.