MOBDrone: A Drone Video Dataset for Man OverBoard Rescue

Donato Cafarelli[1], Luca Ciampi[1], Lucia Vadicamo[1](✉), Claudio Gennaro[1],
Andrea Berton[3], Marco Paterni[2], Chiara Benvenuti[2], Mirko Passera[2],
and Fabrizio Falchi[1]

[1] Institute of Information Science and Technologies, CNR, Pisa, Italy
lucia.vadicamo@isti.cnr.it
[2] Institute of Clinical Physiology, CNR, Pisa, Italy
[3] Institute of Geosciences and Earth Resources, CNR,
Via G. Moruzzi 1, 56124 Pisa, Italy

Abstract. Modern Unmanned Aerial Vehicles (UAV) equipped with cameras can play an essential role in speeding up the identification and rescue of people who have fallen overboard, i.e., man overboard (MOB). To this end, Artificial Intelligence techniques can be leveraged for the automatic understanding of visual data acquired from drones. However, detecting people at sea in aerial imagery is challenging primarily due to the lack of specialized annotated datasets for training and testing detectors for this task. To fill this gap, we introduce and publicly release the MOBDrone benchmark, a collection of more than 125K drone-view images in a marine environment under several conditions, such as different altitudes, camera shooting angles, and illumination. We manually annotated more than 180K objects, of which about 113K man overboard, precisely localizing them with bounding boxes. Moreover, we conduct a thorough performance analysis of several state-of-the-art object detectors on the MOBDrone data, serving as baselines for further research.

Keywords: Man overboard · Object detection · Unmanned Aerial Vehicles · Drone · Benchmark

1 Introduction

The 2021 Annual Overview of Marine Casualties and Incidents [9] reported that 22,532 marine casualties and incidents were occurred between 2014 and 2020 in the waters of EU Member States or involving EU ships. 7,051 of these events involved people, with 550 lives lost and 6,921 injured. The main events that resulted in fatalities were ship collisions and people slipping/falling into the

Supported by NAUSICAA - "NAUtical Safety by means of Integrated Computer-Assistance Appliances 4.0", a project funded by the Tuscany region (CUP D44E20003410009).

D. Cafarelli, L. Ciampi and L. Vadicamo—Co-first authors.

S. Sclaroff et al. (Eds.): ICIAP 2022, LNCS 13232, pp. 633–644, 2022.
https://doi.org/10.1007/978-3-031-06430-2_53

water. Of the falls, 9.8% were falling overboard, resulting in 84 lives lost. Survival chances in a Man Overboard (MOB) incident depend on many variables, including the height of the fall, the water temperature, the sea state, and the weather conditions, along with the rescue operation time, the person's state of consciousness and ability to swim, to name but a few. Unfortunately, in most cases (estimated between 85–90%), it ends in death [12]. Indeed, the rescue operations are usually long and complicated. If the person falls overboard while the boat is navigating (e.g., at a speed of 18 knots), the time that elapses from when the alarm is given to when the boat can slow down and turn 180° to return to the MOB point is several minutes. Not to mention that, for safety reasons, the boat cannot turn back rapidly as it would risk running over the victim. Moreover, since the exact rescue point is not always detectable due to sea currents and alarm delays, it is clear that the MOB scenario is very critical and dangerous.

Quick and effective search and rescue operations (SAR) are crucial to increasing the victim's chances of survival. To this end, it is essential to determine a limited search area and plan paths for rescue boats [20]. Unmanned Aerial Vehicles (UAVs) equipped with thermal and/or video cameras can be profitably used to localize and track people overboard, thus expediting rescue operations and increasing their probability of success. In this regard, the "NAUtical Safety by means of Integrated Computer-Assistance Appliances 4.0" (NAUSICAA) project aims at creating a system for medium and large boats in which the conventional control, propulsion, and thrust systems are integrated with a series of latest generation sensors (including lidar systems, cameras, radar, drones) for assistance during the navigation and mooring phases. Specifically, within the project, we will use commercial aerial drones (equipped with a video camera) and Artificial Intelligence (AI) techniques to search for people overboard automatically.

Many AI techniques have achieved outstanding results in localizing and recognizing people and objects in images and video frames in recent years [21,22,26]. However, evaluating these approaches (or developing new ones) in a MOB scenario is difficult due to the lack of labeled data. Although many annotated datasets containing people and objects in everyday scenarios are publicly available, to the best of our knowledge, the same cannot be said for the case of aerial footage of people and objects in marine environments. To fill this gap, we collected and publicly released [3] a large-scale dataset of aerial footage of people who, being in the water, simulated the need to be rescued. Our dataset, named *MOBDrone*, contains 66 video clips with 126, 170 frames manually annotated with more than 180K bounding boxes (of which more than 113K belonging to the *person* category). The videos were gathered from one UAV flying at an altitude of 10 to 60 m above the mean sea level.

This paper introduces our dataset and describes the data collection and annotation processes. Moreover, it presents an in-depth experimental analysis of the performance of several state-of-the-art object detectors on this newly established MOB scenario, serving as baselines. We hope that this benchmark and the preliminary results may become a reference point for the scientific community concerning the localization of MOBs from UAV imagery.

Evaluation code and all other resources for reproducing the results are available at http://aimh.isti.cnr.it/dataset/MOBDrone.

2 Related Work

In the last years, many annotated datasets have been released for supporting the supervised learning of modern detectors based on deep neural networks [2,6,7,17]. However, only a few include images or videos taken from UAVs, and most are not focused on the marine environment. This section briefly reviews some of these drone-view datasets suitable for object detection.

VisDrone [27] is the largest object detection and tracking dataset in this category. It consists of 179,264 frames extracted from 263 video clips captured by various drone-mounted cameras covering different urban and suburban areas under various weather and lighting condition. Frames are manually annotated with more than 2.6 million bounding boxes localizing targets such as pedestrians, vehicles, and bicycles. Another remarkable dataset is UAVTD [8], suitable for vehicle detection. It consists of 80K images gathered from a UAV platform in different urban scenarios and contains 2,700 vehicles annotated with bounding boxes. Another annotated dataset for car detection is the MOR-UAV [19], which comprises more than 10K drone-view images. Finally, CAPRK [15] is a view drone dataset exploited for detecting and counting parked vehicles [1].

Few works have been done to date on creating datasets of images taken by drones of people in marine environments. Lygouras et al. [18] addressed the problem of open water human detection by conducting real-time recognition onboard a rescue hexacopter. They gathered a swimmers dataset composed of images collected from the internet and frames recorded from a drone. In total, the dataset consists of just 4,500 full HD images. Recently, Varga et al. [24] released the SeaDroneDataset that contains over 54K annotated frames captured from various altitudes and viewing angles. The dataset mainly contains people swimming in open water, and the frames are annotated using six classes: swimmer, floater (swimmer with life jacket), swimmer† (a person on a boat not wearing a life jacket), floater† (a person on a boat wearing a life jacket), life jacket, and boat. The main difference with our dataset is that we focus on people at sea without a life jacket (since in the fall into the water from a large ship it is unlikely that the person was previously wearing a life jacket), and we also consider different scenarios of the person's state of consciousness. Nevertheless, the SeaDronesSee dataset is an excellent reference for the task of human detection and tracking in the marine environment that we plan to use in the future, at least for some classes, in conjunction with our MOBDrone for training and testing deep neural networks. Finally, Ferau et al. in [11] faced the problem of assisting SAR operations in MOB incidents using autonomous UAV-based systems. Unlike our work, they aim to locate people in the water by analyzing images recorded with thermal instead of video cameras. The automatic detection and classification of objects in water from thermal images acquired using UAVs was also explored in [16].

3 The MOBDrone Dataset

Our *MOBDrone Dataset*, which we publicly released at [3], aims to overcome the lack of large public datasets of drone-based imagery for overboard human detection. Its realization required nearly 80 h of work between data acquisition, post-processing, and annotation, involving, among others, a certified pilot of the Fly&Sense Service of the CNR of Pisa for UAV flight operations and two professional divers for in-water activities. In the following, we detail the processes of data collection and curation.

Data Collection. We carried out the drone shooting activities in the Gombo beach of the Migliarino, San Rossore, and Massaciuccoli Park (Pisa, Italy). This choice was dictated for privacy reasons and to ensure compliance with the safety protocols of UAV flight operations. Indeed, the Gombo beach is a segregated area that can be accessed only after obtaining the appropriate authorization from the Park Authority.

To guarantee variability of data, we identified several dimensions of interest, including (i) *subjects/objects to be filmed* (people, lifebuoys, boats, rocks, pieces of wood, parts of the land, and whatever else there is naturally), (ii) *person's state of consciousness* (conscious, semiconscious or unconscious), (iii) *person's visual appearance* (man, woman, persons in light, dark or colored clothing, persons in a bathing suit, etc.), (iv) *light changes* (different shooting times), (v) *altitude and camera directions* (e.g., try to fly at high altitudes to see a more significant portion of the sea, and at lower altitudes better to see the objects and a possible man overboard, also changing the camera shooting angle).

We gathered a total of 49 videos at high resolution (4K) exploiting the DJI FC6310 camera of the Phantom 4 Pro V2 drone. The camera angle was perpendicular to the water (90°), except for a small set of shots where a 45° angle was used. Two professional divers (one male and one female) simulated various scenarios of a person overboard, including a conscious person (swimming, floating, or waving their arms to attract attention) and an unconscious person (floating body in a supine or prone position, or partially floating, i.e., part of the body is below the water surface). Some videos incidentally captured people close to the portion of the sea where our divers were positioned. We split these videos into multiple video clips to remove portions where people were identifiable for privacy concerns. The final dataset contains 66 videos that we post-processed, as described in the following section.

Data Curation. First, we converted the 66 video clips captured in the data acquisition campaign from 4K to 1080p resolution. Then, we extracted the frames from the videos at a rate of 30 FPS, obtaining a total of 126,170 images (see Table 1 for summary statistics). Finally, a human expert annotator manually annotated them. Specifically, the annotation process took approximately 60 h, and the Computer Vision Annotation Tool (CVAT) [23] was used. Although our work focuses on localizing and recognizing people, we also annotated other

Table 1. Dataset details. The MOBDrone benchmark contains 126,170 drone-view images at six different heights in MOB scenarios.

Altitude	# Images	# Video clips
10 m	958	1
20 m	10,053	6
30 m	29,404	15
40 m	33,046	13
50 m	29,183	16
60 m	23,526	15
tot	**126,170**	**66**

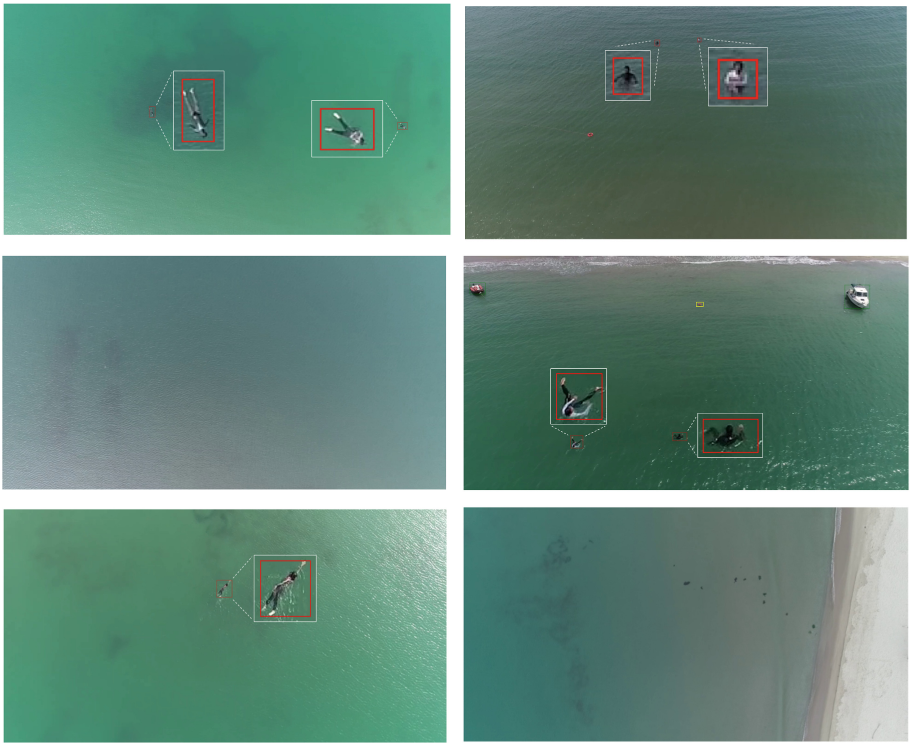
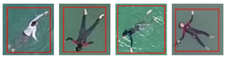
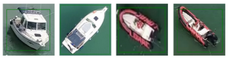
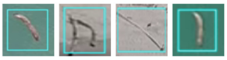
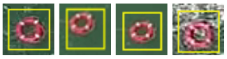
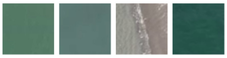


Fig. 1. Samples of the MOBDrone Dataset. Examples of images captured at different altitudes, light conditions, and camera directions. The bounding box annotations localizing the labeled objects are also shown. Objects belonging to the *person* category, which is the one of paramount interest in MOB scenarios, are outlined with red bounding boxes and zoomed. Note that 27.72% of the images do not contain objects (i.e., images of clear water) and that interfering objects in the background, such as rocks, often trigger false positive detections. (Color figure online)

Table 2. Annotation statistics. We labeled with bounding boxes 181,689 objects belonging to 5 categories.

Class	#Annotations	#Images	Samples
person	$113,408$	$77,365$	
boat	$39,967$	$31,238$	
wood	$15,980$	$9,040$	
life buoy	$10,401$	$10,386$	
surfboard	$1,933$	$1,933$	
`no object`		$34,976$	
total	181,689		

objects present in the scenes. In particular, we considered a total of 5 classes (*person*, *boat*, *surfboard*, *wood*, *life_buoy*). We provide a bounding box precisely localizing each instance of the objects of interest. The total number of annotations is $181,689$, of which the ones related to the *person* class, which is of primary interest in the MOB scenario, is $113,408$. However, note that about 27.72% of the images do not contain any objects (i.e., images of clear water). We report some statistics concerning the annotations in Table 2, while we show some samples of our dataset in Fig. 1.

4 Detection Performance Analysis

In this section, we evaluate several state-of-the-art object detectors on our MOB-Drone dataset[1], focusing on the detection of the overboard people, i.e., on the localization of the object instances belonging to the class *person*. In the first part of our performance analysis, we compare 9 of the most popular and performing object detectors present in the literature. Then, we look upon the best three ones, performing a more in-depth analysis of the obtained results.

The detection methods considered in our analysis can be roughly grouped into three categories, i.e., anchor-based Convolutional Neural Network (CNN) methods, anchor-free CNN methods, and Transformer-based methods. We briefly summarize them below. We refer the reader to the papers describing the specific detectors for more details.

[1] Although in this work we exploited the whole dataset as a test benchmark, in [3] we provide training and test splits.

Table 3. Comparison of the considered detectors. mAP@[0.50:0.95] is the AP averaged over 10 IoU thresholds in the range [0.50 : 0.95] with a step size of 0.05, while AP50 is the AP computed at the single IoU threshold value of 0.50.

Method	AP50 ↑	mAP@[0.50:0.95] ↑
VarifocalNet [25]	**0.378**	**0.144**
TOOD [10]	0.314	0.116
Deformable DETR [28]	0.199	0.075
YOLOX [13]	0.126	0.049
Faster R-CNN [22]	0.126	0.041
CenterNet [26]	0.124	0.041
DETR [4]	0.128	0.040
Mask R-CNN [14]	0.109	0.033
YOLOv3 [21]	0.011	0.009

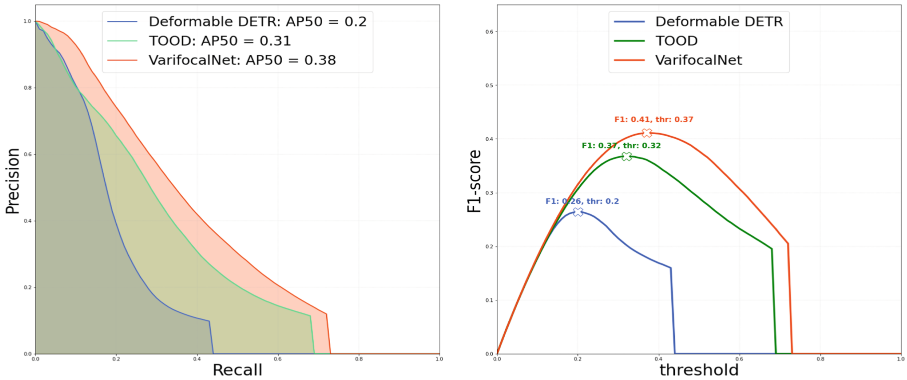
Anchor-based CNN methods compute bounding box locations and class labels of object instances exploiting CNN-based architectures that rely on anchors, i.e., prior bounding boxes with various scales and aspect ratios. They can be divided into two groups: i) the two-stage approach, where a first module is responsible for generating a sparse set of object proposals and a second module is in charge of refining these predictions and classifying the objects; and ii) the one-stage approach that directly regresses to bounding boxes by sampling over regular and dense locations, skipping the region proposal stage. Here, we use Faster R-CNN [22] and Mask R-CNN [14] regarding the first group, and YOLOv3 [21], TOOD [10] and VarifocalNet (VfNet) [25] concerning the second one.

Anchor-free CNN methods rely on the prediction of key-points, such as corner or center points, to predict the objects, instead of using anchor boxes and their inherent limitations. In this work, we exploit CenterNet [26] and YOLOX [13].

Transformer-based methods rely on the recently introduced Transformer attention modules in processing image feature maps, removing the need for hand-designed components like a non-maximum suppression procedure or anchor generation. In this paper, we consider DEtection TRansformer (DETR) [4] and one of its evolution, Deformable DETR [28].

We evaluate and compare the above-described detectors over our MOBDrone dataset following the golden standard Average Precision (AP), i.e., the average precision value for recall values over 0 to 1. Specifically, we consider the MS COCO mAP@[0.50:0.95] [17], i.e., the AP averaged over 10 IoU thresholds in the range $[0.50, 0.95]$ with a step size of 0.05, and the AP50, i.e., the AP computed at the single IoU threshold value of 0.50. We refer the reader to [17] for more details. All the detection techniques that we employed were pre-trained[2] on the COCO dataset [17], a popular collection of images in everyday contexts compris-

[2] Pre-trained models are available, e.g., in the model zoo of MMDetection project [5].



(a) Precision vs. Recall curves (IoU=0.5). Areas under curves correspond to AP50.

(b) F_1-score vs. detection threshold curves (IoU=0.5).

Fig. 2. Comparison of the three best detectors. We report Precision-Recall (a), and F_1-detection threshold (b) curves of the three best models (VfNet, TOOD, and Deformable DETR). VfNet shows best performances.

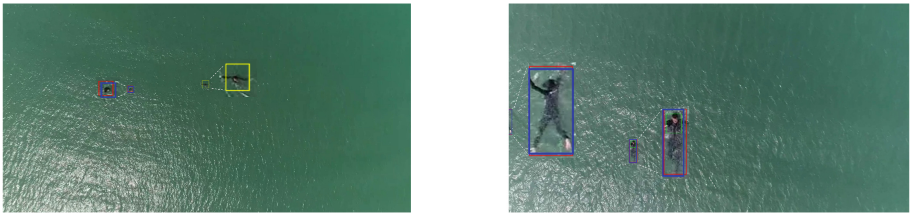


Fig. 3. Detections produced by VarifocalNet. We indicate false positives in green, false negatives in yellow, true positive in blue and gt in red. (Color figure online)

ing objects belonging to 80 different categories, of which is present the *person* class. To evaluate the performance of the detectors, we filtered the obtained detections considering only the ones classified as *person*. We report the obtained results in Table 3. The model which turns out to be the most performing is VarifocalNet, considering both the metrics, followed by TOOD and Deformable DETR. However, in general, all the detectors exhibit moderate performance, indicating the difficulties in localizing persons in this challenging scenario. We deem that the most significant metrics in our scenario is the AP50, since i) the dataset is manually labeled by humans and therefore is accurate in terms of classification and inaccurate in terms of boundaries, ii) it is *not* crucial to precisely localize instances, i.e., it is critical to detect overboard persons but the quality of the localization is less important. With this in mind, in the following, we show an in-depth analysis of the three AP50 best models, i.e., VarifocalNet, TOOD, and Deformable DETR.

Table 4. Comparison of the three best detectors at different altitudes. AP50 and F_1 are the AP and the F_1-score computed with IoU set to 0.50.

	VarifocalNet [25]		TOOD [10]		Deformable DETR [28]	
Altitude	AP50 ↑	F_1 ↑	AP50 ↑	F_1 ↑	AP50 ↑	F_1 ↑
10 m	0.973	0.444	**0.989**	0.363	0.959	**0.636**
20 m	**0.771**	**0.318**	0.681	0.308	0.514	0.279
30 m	0.400	0.199	**0.407**	**0.223**	0.240	0.210
40 m	**0.540**	**0.226**	0.406	0.203	0.314	0.209
50 m	**0.241**	**0.161**	0.187	0.140	0.107	0.08
60 m	**0.205**	**0.223**	0.171	0.196	0.063	0.131

In Fig. 2a, we report the Precision-Recall curves, i.e., precision and recall values for different detection confidence thresholds, of these three best detectors while setting the IoU threshold at 0.50. Areas under these curves correspond to AP50 values. As can be seen, the VarifocalNet detector exhibits the best performance at all confidence thresholds. The same trend is confirmed in Fig. 2b, where we show F_1-score values (where $F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$) at different detection confidence thresholds, again setting the IoU threshold at 0.50. Still, VarifocalNet shows superior performance compared to the other two detectors. Please note that the maximum values of these curves indicate the detection confidence score that may be used by potential users, enclosing a trade-off between the resulting Precision and Recall values. Figure 3 shows some qualitative outputs produced by VarifocalNet using this confidence score.

In Table 4, we show a comparison of the best three detectors at different altitudes in terms of AP50 and F_1-score. As expected, in general, performances decrease with increasing altitude. However, it is interesting to note that TOOD and Deformable DETR particularly struggle to detect small objects, i.e., when the altitude is above 40 m, while achieving comparable or even better results than VarifocalNet at altitudes below 30 m.

Finally, in Table 5, we report a classwise analysis of the obtained detections, i.e., we consider the detections belonging to all the 80 classes and not only the detections classified as *person*. Specifically, we take into account also errors due to misclassified objects, i.e., detected objects that matched with *person* annotations but that were classified as objects belonging to another category. We define the *True Positive Rate* (TPR) as the ratio between the number of correctly detected and classified person instances (TP) and the total number of person instances in the ground-truth (P). On the other hand, we define $\mathrm{dTPR}(c) = \frac{\mathrm{dTP(c)}}{\mathrm{P}}$ as the *detection True Positive Rate* for the output class c with respect to the target *person* class, that is the number dTP(c) of person instances detected correctly (i.e., considering only the IoU of predicted and target bounding boxes) but classified with category c divided by the total number of person instances in the ground-truth. In other words, dTPR(c) gives us the proportion of person

Table 5. Classwise Analysis. We consider the detections of all 80 COCO classes, accounting for errors due to misclassified objects, i.e., detected objects that matched with *person* annotations but that were classified as objects of another category. TPR is the *True Positive Rate* with respect to the target *person* class. dTPR is the ratio of person instances correctly detected but misclassified. The overall *detection Recall* (dR) is the proportion of detected person instances considering also misclassified objects; the overall *detection Miss Rate* is the proportion of person instances that were not detected at all. We set IoU to 0.5.

	Person	Bird	Airplane	Kite	Other	Overall	
Method	TPR ↑	dTPR ↑	dTPR ↑	dTPR ↑	dTPR ↑	dR ↑	dMR ↓
VfNet [25]	0.285	0.266	0.190	0.067	0.012	**0.818**	**0.182**
TOOD [10]	**0.326**	0.118	0.212	0.125	0.017	0.799	0.201
Def. DETR [28]	0.206	0.311	0.072	0.041	0.026	0.657	0.343

instances that were detected correctly but misclassified with category c. The sum of the TPR and all the dTPR(c) gives the overall *detection Recall* (dR), i.e., the ratio of person instances detected correctly without considering the output classification. Similarly, the overall *detection Miss Rate*, defined as dMR=1-dR, is the portion of person instances that were not detected at all. For example, from Table 5, we can observe that the pre-trained VarifocalNet correctly detected 81.8% of the ground-truth person instances even if, in most cases, it misclassified them. This may suggest that the same model fine-tuned on MOB data may have room for growth in localizing *person* instances.

5 Conclusion and Future Directions

This paper presents the MOBDrone benchmark, a large-scale drone-view dataset suitable for detecting persons overboard. It is part of the NAUSICAA project aiming at creating a control system that, for the first time, uses aerial and marine drones and augmented and virtual reality to provide increased safety to medium and large vessels. Specifically, we collected more than 125K images, and we manually annotated more than 180K objects in a marine environment under several conditions, like different altitudes and camera shooting angles. Furthermore, we report an in-depth experimental evaluation of several state-of-the-art object detectors, serving as baselines for further research on this topic. Our analysis shows that detectors pre-trained on standard datasets of everyday objects exhibit moderate performance in localizing and recognizing people at sea in aerial images acquired at mid-high altitudes. The classification stage is the primary source of error for the best of the tested models, i.e., VarifocalNet, as about 82% of the ground-truth persons were correctly detected but misclassified, thus suggesting that the same model fine-tuned on MOB data may have room for growth. To this end, as a future direction, we plan to extend our dataset with additional annotated data for the supervised training procedure, also considering synthetic images coming from virtual worlds.

References

1. Amato, G., Ciampi, L., Falchi, F., Gennaro, C.: Counting vehicles with deep learning in onboard UAV imagery. In: 2019 IEEE Symposium on Computers and Communications (ISCC). IEEE, June 2019. https://doi.org/10.1109/iscc47284.2019.8969620
2. Amato, G., Ciampi, L., Falchi, F., Gennaro, C., Messina, N.: Learning pedestrian detection from virtual worlds. In: Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds.) ICIAP 2019. LNCS, vol. 11751, pp. 302–312. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30642-7_27
3. Cafarelli, D., et al.: MOBDrone: a large-scale drone-view dataset for man overboard detection, February 2022. https://doi.org/10.5281/zenodo.5996890
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
5. Chen, K., et al.: MMDetection: Open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
6. Ciampi, L., Messina, N., Falchi, F., Gennaro, C., Amato, G.: Virtual to real adaptation of pedestrian detectors. Sensors **20**(18), 5250 (2020). https://doi.org/10.3390/s20185250
7. Ciampi, L., Santiago, C., Costeira, J., Gennaro, C., Amato, G.: Domain adaptation for traffic density estimation. In: Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SCITEPRESS - Science and Technology Publications (2021). https://doi.org/10.5220/0010303401850195
8. Du, D., et al.: The unmanned aerial vehicle benchmark: object detection and tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 375–391. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_23
9. European Maritime Safety Agency: Annual overview of marine casualties and incidents 2021 (2021)
10. Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: TOOD: task-aligned one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3510–3519, October 2021
11. Feraru, V.A., Andersen, R.E., Boukas, E.: Towards an autonomous UAV-based system to assist search and rescue operations in man overboard incidents. In: 2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), pp. 57–64. IEEE (2020). https://doi.org/10.1109/SSRR50563.2020.9292632
12. Garay, E.: What Happens When Someone Falls Off a Cruise Ship (2017). https://www.cntraveler.com/story/what-happens-when-someone-falls-off-a-cruise-ship. Accessed 25 Jan 2022
13. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430 (2021)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: IEEE International Conference on Computer Vision, ICCV 2017, pp. 2980–2988. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.322
15. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4145–4153 (2017)

16. Leira, F.S., Johansen, T.A., Fossen, T.I.: Automatic detection, classification and tracking of objects in the ocean surface from UAVs using a thermal camera. In: 2015 IEEE Aerospace Conference, pp. 1–10. IEEE (2015). https://doi.org/10.1109/AERO.2015.7119238
17. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
18. Lygouras, E., Santavas, N., Taitzoglou, A., Tarchanidis, K., Mitropoulos, A., Gasteratos, A.: Unsupervised human detection with an embedded vision system on a fully autonomous UAV for search and rescue operations. Sensors **19**(16), 3542 (2019). https://doi.org/10.3390/s19163542
19. Mandal, M., Kumar, L.K., Vipparthi, S.K.: MOR-UAV: a benchmark dataset and baselines for moving object recognition in UAV videos. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2626–2635 (2020). https://doi.org/10.1145/3394171.3413934
20. Mou, J., Hu, T., Chen, P., Chen, L.: Cooperative MASS path planning for marine man overboard search. Ocean Eng. **235**, 109376 (2021). https://doi.org/10.1016/j.oceaneng.2021.109376
21. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017). https://doi.org/10.1109/tpami.2016.2577031
23. Sekachev, B. et al.: Computer Vision Annotation Tool (CVAT) (2020). https://github.com/openvinotoolkit/cvat
24. Varga, L.A., Kiefer, B., Messmer, M., Zell, A.: SeaDronesSee: a maritime benchmark for detecting humans in open water. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2260–2270, January 2022
25. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: VarifocalNet: an IoU-aware dense object detector. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2021. https://doi.org/10.1109/cvpr46437.2021.00841
26. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
27. Zhu, P., et al.: Detection and tracking meet drones challenge. IEEE Trans. Pattern Anal. Mach. Intell. **01**, 1 (2021). https://doi.org/10.1109/TPAMI.2021.3119563
28. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: 9th International Conference on Learning Representations, ICLR 2021. OpenReview.net (2021)