Person Detection in Drone Imagery

1st Sasa Sambolek
High school Tina Ujevica
Kutina, Croatia
sasa.sambolek@gmail.com

2nd Marina Ivasic-Kos
Department of Informatics
University in Rijeka, Rijeka, Croatia
marinai@uniri.hr

***Abstract*— The use of drones in search and rescue operations has become standard almost everywhere in the world. A special challenge in the search and rescue operation is the automatic detection of persons in different terrains, in different situations and body positions, in different weather conditions, and from different shooting heights during a drone flight. This paper investigates the accuracy of people detection in drone images on existing VisDrone, Okutama - Action datasets, and on a custom SARD image dataset built to simulate search and rescue scenes. A Faster R-CNN with FPN as the backbone, pre-trained on the COCO data set, was used as a detector. The person detector is additionally trained on the SARD data set containing 1,981 images and on the subset of the VisDrone set. After transfer learning, a significant improvement in the detection results of persons in the images taken by the drone was achieved concerning mAP and precision and recall.**

Keywords—Faster RCNN, drone imagery, search and rescue

I. INTRODUCTION

In the case of searching for a missing person, it is of great importance to find the person in the shortest possible time as this increases the likelihood of survival.

In the past few years, unmanned aerial vehicles (drones) have been included in the search and rescue operations in addition to existing resources such as search dogs, human resources, helicopters. During a drone flight, the operator must simultaneously operate the drone and search for the missing person, who, due to the distance, is generally small in size, very often in a lying or crouching position, in inaccessible terrain, obscured by vegetation, which further complicates the detection of missing persons. Ground forces can check the terrain well, but they progress very slowly and have a small view field, especially in the case of dense vegetation so the assistance of the aircraft is necessary.

An ideal search and rescue system would be one that would include drones that could autonomously fly and detect objects of interest in real-time, and then alarms ground teams and forwards them the location and the image of detected objects. At lower altitudes, the drone can capture more details about objects of interest, while at higher altitudes it covers a larger area but the objects are extremely small on them.

The drone footage is being analyzed by video analysts today. In [1] is described that the human video analyst was able to detect the victim within 25 seconds in the drone recording (4K image, with target size 5 – 50 pixels), focusing on the small part of the image that, according to previous experience, is the most likely to be the person being sought. High concentration is required for that task and the help of an automated detector can be of great benefit.

In recent years, considerable progress has been made in automatic object detection in images using deep learning (convolutional neural networks). However, it has been shown that popular detectors such as SSDs [2], YOLO [3], and RetinaNet [4] do not achieve equally good detection results from a bird's eye view or on images captured by drones [5].

Automatic detection of objects on drone imagery poses greater challenges than the same task on stationary camera images. One reason is the change in shooting height, which causes a significant change in the size of the object, a change in the shooting angle and the position of the object towards the camera, and a change in perspective. In the case of a search and rescue operation, the visibility of the object is also affected by changes in lighting (daytime, nighttime) and weather conditions (sunny, cloudy, foggy or rainy). With all of the above, the challenge of detecting an object captured by a drone is very often very small object size that is hard to see in a cluttered background with frequent occlusions.

In this paper, the performance of a popular state-of-the-art object detector, a Faster R-CNN for detecting persons in drone-captured images was investigated.

Two publicly available sets of images taken with a drone and prepared for deep learning tasks, the VisDrone and Okutama – Action datasets have been selected. Each of these datasets includes scenes designed for a specific task and tailored to a specific problem. For a specific problem, such as a search and rescue operation, they do not have proper scenarios with people lying in the grass, crouching behind a stone, leaning against a tree or other atypical poses for urban scenes, so our own custom set of images called SARD (Search And Rescue Dataset) was created.

The rest of the paper is organized as follows: in Section II. an overview of the drone-related research is given with an emphasis on image datasets and commonly used detection methods. Section III. describes experiment and training of the Faster RCNN model for person detection on custom dataset SARD containing typical scenes for the rescue operation and two public datasets of drone imaginary. Obtained results and discussion are given in Section IV. The paper ends with a conclusion and a proposal for future research.

II. RELATED WORK

The detection of persons in drone images and videos is of increasing relevance and has a significant role to play in the safety of persons and the surveillance in urban and non-urban areas.

A. Datasets

A prerequisite for the use of models in various applications, and so is in the field of UAV imaging, is the preparation of appropriate image databases used for supervised model learning. Publicly available datasets that have contributed to the development of computer vision research in the field of drone images are Campus [6], UAV123 [7], CARPK [8], Okutama – Action [9], UAVDT [10], VisDrone [11].

Each of the image databases is intended for a specific purpose and is tailored to a specific problem. They usually contain different classes taken from a bird's eye view that are present in urban scenes such as pedestrians and skateboards on the streets or squares, cyclists, cars, buses, and trucks on

roads, crossings, or parking lots [6]. There are also examples containing non-urban landscapes such as fields and beaches with objects such as boats and bathers [11]. In some cases, activities of the people such as walking, running, reading, hugging and the like are also indicated [9].

In this work, VisDrone, and Okutama – Action datasets have been used, so they were described in more detail.

1) VisDrone

The VisDrone dataset contains 288 videos and 10,209 images captured on different drone platforms (DJI Mavic, DJI Phantom Series 3, 3A, 3SE, 3P, 4, 4A, 4P) in 14 different cities in China. The set covers different weather and light conditions of maximum video resolution (3840 x 2160 px) and images (2000 x 1500 px). Within the set are 10 categories of objects (pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle).

2) Okutama-Action

The dataset contains 43 drone-recorded video clips for training and testing models to detect multiple simultaneous actions within different categories, human to human interaction: handshaking, hugging, human to object interaction: reading, drinking, pushing/pulling, carrying, calling, non-interaction: running, walking, lying, sitting, standing. Using the open-source tool to annotate objects VATIC [12], they manually annotate every tenth frame, and the tags were linearly interpolated to 30 fps.

The videos were shot using two DJI Phantom 4 drones on baseball court in 4K resolution with 30 fps, at a height of 10 m to 45 m, with a camera angle of 45 or 90 degrees. The data set for each video contains metadata such as camera angle, speed, and height. The shots were taken with two different lighting conditions (sunny and cloudy).

Analyzing the available databases of drone images, the conclusion was that there is still no publicly available dataset containing scenes captured by a drone during search and rescue operations, so in this paper, our dataset for this purpose has been created.

B. Methods used to detect persons in rescue operations

In recent years, drones have been increasingly used, and methods for the automatic detection of drone imaging objects have been increasingly developed. We are particularly interested in detection methods used to detect persons in search and rescue operations. One of the earlier works is [13] where drones are used to find injured persons in search and rescue operations using HOG descriptors [14].

The advantages of using deep learning for computer vision tasks using drones are presented in [15], where authors have analyzed three models (SSD, Faster R-CNN, and RetinaNet) and showed that RetinaNet is faster and more accurate model than others analyzed, in object detection task on drones imagery.

In [16] multi-spectral and visible-spectrum cameras are used, with modified MobileNet architecture to detect and localize bodies in the sea. The upgraded version of the MobileNetv2 model and the Okutama-Action dataset is used in [17] for person detection. In [18] for detection of persons in the water, a Tiny YOLO V3 Architecture integrated on NVIDIA Jetson TX1 computer is used. The model was trained on a COCO dataset and a custom swimmer's dataset recorded with an unmanned aerial vehicle.

The use of drones to detect avalanche casualties is described in the [19], where the Inception model with the Support Vector Machine classifier is used for detection.

The YOLO detector was used to detect aircraft in real-time on videos obtained from the UAV during the flight [20] while the aircraft were grounded. The YOLO detector has also proven to be a good solution for people detection from a bird's eye view in quite demanding shooting conditions [21] with a large number of objects on the scene [22], with occlusion among people and indoors [23].

In [24] image segmentation, contrast enhancement, and convolution neural networks are applied for the detection of persons (range 5 to 50 px) on drone imagery. They have also used ARMA3 a 3D game editor to generate synthetic search and rescue datasets and data augmentation (flip, rotation, zoom in/out). An approach that increases a relatively modest set of real-world data with synthesized images has also been applied in [25] to influence the improved performance of object detectors. The size and position of the persons or object in general in the synthesized images should be adjusted to the actual situations, e.g. in these works persons was set on 5-30 px.

For search and rescue operations to be carried out even when there is no more daylight, the use of IR light should also be considered. A Yolo detector was used to detect humans on thermal images recorded at night in [26] and in [27] to recognize humans while sneaking through the woods and animals during bad weather. In [28] an infrared camera was mounted on an unmanned aerial vehicle to detect poachers and control animal movements using Faster R-CNN.

The [29] describes applying multiple object-based visual tracking to aerial imagery for search and rescue purposes. Person detection was based on color and depth information and the use of the Human Shape Validation Filter that uses the locations of the human joints detected by the Convolutional Pose Machine [30] to avoid false detections. During the tracking of persons, the method used must be invariant for the scale, movement, and rotation of the object and also that has the ability to re-identify persons. For that purpose, in [31] a DeepSort method was used to track people on the sports field. When monitoring objects, especially when the objects are very far from the cameras and often in occlusion, as is the case with drone imagery, satisfactory results are not yet achieved. Something that certainly goes in favor of solving this challenge is more precise object detection.

III. EXPERIMENT SETUP

The experiment aimed to detect people in scenes appropriate to search and rescue cases.

For detection, Faster R-CNN [32] was decided to use, which has become the de facto standard after proving to be a multi-purpose detector that enables high accuracy of detecting small and large objects [33]. The original implementation of the Faster R-CNN model was used with Feature Pyramid Network - FPN [34] as a backbone without changing the hyperparameters of the model. The model was trained on the COCO [35] dataset. According to results reported in [36], average precision (AP) of the faster_rcnn_R_50_FPN_3x model for person detection on COCO (val2017) dataset was 54.46%.

Fig. 1. Some example of drone images from VisDrone dataset [11] (top), Okutama-Action [9] (middle) and SARD dataset (bottom)

Our goal was to apply the knowledge from the pre-trained model and features and weights learned on a COCO-dataset for person detection to the new but related problem of person detection on images captured by drones. The goal was to use transfer learning to overcome the isolated learning paradigm for only one task and to avoid learning models from scratch.

The key motivation was that learning a deep learning model for a complex task requires a large amount of data that is not easy to collect and can be very time-consuming and arduous to label and prepare data for supervised learning. An additional motivation to use transfer learning was to learn a model that goes beyond specific tasks and tries to use knowledge from pre-trained models to solve new problems and to avoid the bias problems the most models have, that can be successfully used only on the specific domain for which they were specialized.

Three datasets have been used in this experiment: VisDrone, Okutama - Action, and our dataset SARD.

From the VisDrone dataset, 2,000 images containing person class (Fig 1, top row) were selected. Objects that represent a person are labeled either as pedestrians or as persons in the VisDrone dataset. The set was divided into two subsets, a training set containing 1,598 images with 29,797 labeled persons, and a test set containing 402 images with 7,329 person objects. A model trained on images from the VisDrone dataset is called a CV model.

A custom dataset has been built and prepared, referred to as SARD, containing images recorded by the DJI Phantom 4A drone in the area of Moslavacka Gora, Croatia (Fig. 1, bottom row). The footages were taken in a non-urban area along the road, lake, meadow, quarry, forest. The flight altitude of drones during the shooting was 5 m to 50 m, with a camera angle of 45° to 90° and lens FOV 84°. Different people were recorded while performing various actions such as walking, running, sitting, lying down according to scenarios depicting the injured person. The aim was to capture different situations in which the people being searched may find themselves.

The dataset was obtained from 8 videos in 1920px x 1080px resolution, 50fps with a total of 115,767 frames, by selecting 1,981 images and manually tagging the person on them. The set was divided into two subsets, a training set containing 1,579 images with 5,160 tagged individuals and a test set of 402 images containing 1,317 tagged persons. To prepare ground truth data, the boxes to each person in the images using the LabelImg tool was ticked.

A model trained on the SARD dataset is called CS.

Besides, the data from the VisDrone dataset and the SARD dataset have been merged to train the model that is referred to as CVS.

The modes were trained on a laptop with an i5-7300HQ CPU and GeForce GTX 1050Ti 4GB GPU on Ubuntu 18.04.4 64-bit. Detectron2, the open-source object detection system from Facebook AI Research, was used as the software. The CV model was trained in 36,000 iterations for 5 hours on the VisDrone subset, and the CS model 5.5 hours on the SARD dataset. The CV model was additionally trained for 5.5 hours on the SARD dataset (CVS label).

For additional testing of the generality of CV, CS, and CVS models, images from the Okutama - Action dataset have been used (Fig.1, middle row). The set consists of 290 selected frames with 2,066 persons that were manually labeled. The image resolution was reduced to 1280px x 720px for this experiment.

IV. Results and Discussion

The model performance was compared concerning average precision (AP). The detections are considered true

Fig. 2. Visual representation of intersection over union (IoU) criteria equal to or greater than 50% [21]

positive when the intersection over union (IoU) of the detected bounding box and the ground truth box exceed the threshold

of 0.5. The IoU is defined as the ratio of the intersection of the detected bounding box and the ground truth (GT) bounding box and their union (Fig. 2)

First, we have tested all the models on the SARD dataset. The original model trained on the COCO dataset, with no additional training (referred to as COCO model) achieved AP of 36.84%, much lower than reported on the COCO dataset. The CV model, re-trained on images from the VisDrone dataset achieved 35.88% AP for person detection on SARD data. That is even lower than the original model and represents a negative knowledge transfer probably because images in the VisDrone dataset used for re-training were taken at higher altitudes than in the SARD dataset on which the model was tested.

The CS and CVS models were both re-trained using images from the SARD training dataset, and were more successful, achieving 95,84% AP and 96.40% AP, respectively. The huge difference in detection results is due to the large difference in training sets compared to the test set. In the COCO dataset, there are no images from a bird's eye view and in the VisDrone dataset, the distance of the person from the camera is much greater. On the other hand, images from the SARD training set had an important impact on more accurate adjustment of feature maps and better detection results, since were shot under the same conditions, at the same distance, and from the same perspective as in the case of the SARD test dataset. The graphical representation of the results on the SARD dataset is shown in Fig. 3 (blue columns).

In the next case, we have tested all the models on the VisDrone dataset. The person detection results on the VisDrone test set are shown in Fig. 3 (green columns) and are as follows: COCO has an AP of 17.37%, CS: 9%, CV: 40.3% and CSV: 12.88%.

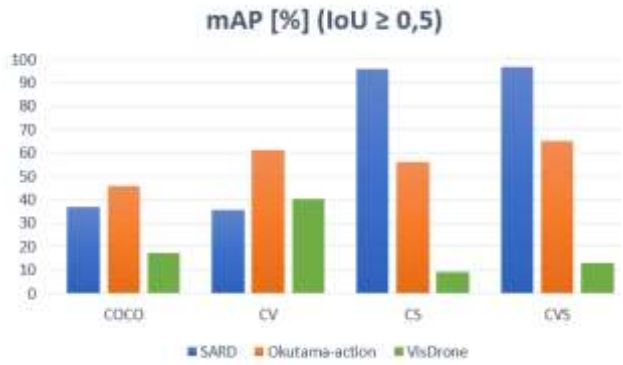

Fig. 3. Person detection mAP results of COCO, CV, CS, and CVS models on SARD, Okutama-action and VisDrone test datasets

All the models not trained on the images of the VisDrone training set (COCO, CS, CSV), achieve significantly worse results than in the first case. The probable reason is that the images in the VisDrone dataset were taken from a much higher shooting height and the objects are tiny, so models that did not have such examples in the learning set cannot detect them.

Finally, all the models were tested on selected images from the Okutama - Action database. This dataset is not used for the training of any of the models. The results are shown in Fig. 3 (orange columns) and are as follows: COCO: 45.97%, CV: 61.31%, CS: 56.12%, and CVS: 65.33%. The best results were achieved by models that had images from the VisDrone database in the training set. The CV model achieved more than 15% better accuracy, and CVS almost 20% better accuracy than the base COCO model. This shows that the initial model

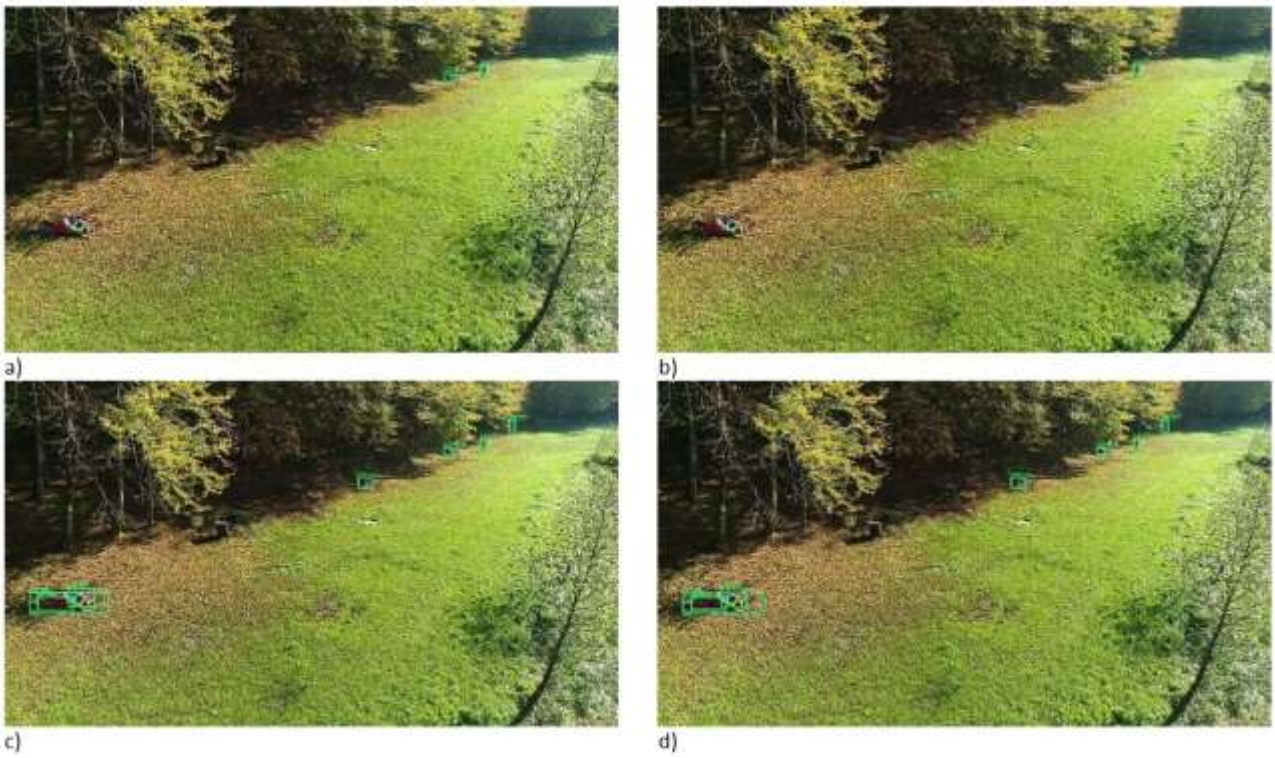

Fig. 4. Detection results of: a) COCO, b) CV, c) CS, d) CVS models

is much better trained for detecting persons in drone images after transfer learning on drone datasets.

The performances of the COCO, CV, CS, and CVS models on different test datasets in terms of the average precision and recall are shown in Fig. 5.

Overall, the highest precision and recall of over 90% is achieved by the CS and CVS models on the SARD dataset. That provides a promising base ground for further research

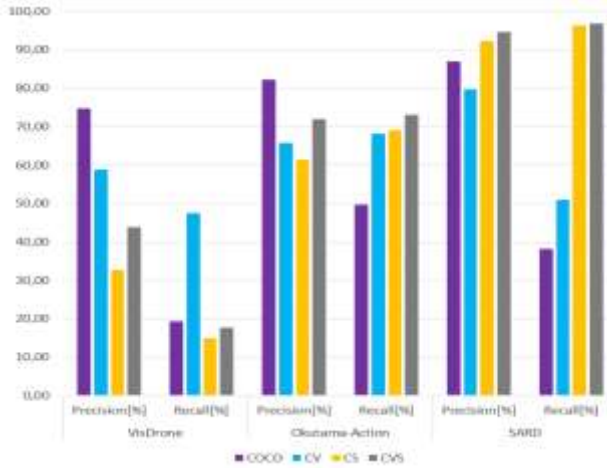

Fig. 5. Average precision and recall of COCO, CV, CS and CVS models on different test datasets

when we can investigate the results of precision and recall in the case when the IoU decreases because the goal is to find the lost person and not to detect it completely. On Okutama – Action dataset the best result of 82.15% of precision has the COCO model that was also the best with respect of the highest precision on VisDrone dataset (74.65%) but with a rather low recall of 18%. CV models on the VisDrone dataset get a precision of 58.76%, but with the highest recall of 48%. CS and CSV performed much better on the Okutama - Action dataset in terms of both precision and recall.

A Fig. 5. shows an example of detection results for all four models. There are seven people in the scene, one standing, one running, and five lying down (three on each other - an occlusion example). COCO model has detected running kid and one person lying down, CV model only running kid while CS and the CVS models have detected all persons on the image.

In a case with a camera positioned from a bird's perspective Fig. 6., COCO and CV models have detected the same two pedestrians, while CS and CVS models have detected all persons on the image.

V. CONCLUSION

Recordings taken from the drones today are used mainly in search of missing persons, in mountain rescue, in the border control, and the like. The ability to automatically detect persons and objects on the images taken from a bird's perspective would greatly facilitate the search and rescue of the people.

In this paper, we have tested the performance of the Faster R-CNN detector for a person detection task on three datasets: SARD, custom dataset built to simulate search and rescue operations, and freely available drone datasets Okutama-action and VisDrone. In experiment we have used publicly available Faster R-CNN model implementation with corresponding weights learned on the COCO data set.

We have additionally trained the Faster R-CNN model on VisDorone and SARD datasets to fine-tune the model parameters for person detection on drone-captured images. In

Fig. 6. Detection results from bird's perspective of; a) COCO, b) CV, c) CS, d) CVS models

the experiment, we showed a positive impact of transfer learning so that the model that was re-trained on SARD images and VisDrone images achieved the best results of person detection in drone-captured images concerning both mAP precision and recall metrics.

In future work, we will expand our database with additional drone imagery and focus on changes in detector architecture to achieve even better results in object detection.

ACKNOWLEDGMENTS

This research was supported by Croatian Science Foundation under the projects IP-2016-06-8345 "Automatic recognition of actions and activities in multimedia content from the sports domain" (RAASS) and IP-2018-01-7619 "A Knowledge-based Approach to Crowd Analysis in Video Surveillance (KACAVIS) and by the University of Rijeka (project number 18-222).

REFERENCES

[1] K. Yun, L. Nguyen, T. Nguyen, D. Kim, S. Eldin, A. Huyen, E. Chow, "Small target detection for search and rescue operations using distributed deep learning and synthetic data generation," in Pattern Recognition and Tracking XXX (Vol. 10995, p. 1099507), 2019.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, "SSD: Single shot multi-box detector," in European conference on computer vision, Springer, Cham, 2016, pp. 21-37.

[3] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[4] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980-2988.

[5] D. R. Pailla, "VisDrone-DET2019: the vision meets drone object detection in image challenge results, 2019.

[6] A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in European conference on computer vision, Springer, Cham, 2016, pp. 549-565.

[7] M. Mueller, N. Smith, B. Ghanem, "A benchmark and simulator for UAV tracking," in European conference on computer vision, Springer, Cham, 2016, pp. 445-461.

[8] M. R. Hsieh, Y. L. Lin, W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in Proc. of the IEEE International Conference on Computer Vision, 2017, pp. 4145-4153.

[9] M. Barekatain, et. al. "Okutama-action: An aerial view video dataset for concurrent human action detection," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 28-35.

[10] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 370-386.

[11] P. Zhu, L. Wen, X. Bian, H. Ling, Q. Hu, "Vision meets drones: A challenge," arXiv preprint arXiv:1804.07437, 2018.

[12] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision, 101(1):184–204, 2013.

[13] M. Andriluka, P. Schnitzspan, J. Meyer, S. Kohlbrecher, K. Petersen, O. Von Stryk, B. Schiele, "Vision-based victim detection from unmanned aerial vehicles," in 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2010, pp. 1740-1747.

[14] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in International Conference on computer vision & Pattern Recognition, 2005, pp. 886-893.

[15] X. Wang, P. Cheng, X. Liu, B. Uzochukwu, "Fast and Accurate, Convolutional Neural Network Based Approach for Object Detection from UAV," in IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society, 2018, pp. 3171-3175.

[16] A. J. Gallego, A. Pertusa, P. Gil, R. B. Fisher, "Detection of bodies in maritime rescue operations using unmanned aerial vehicles with multispectral cameras," Journal of Field Robotics.

[17] R. Geraldes, A. Gonçalves, T. Lai, M. Villerabel, W. Deng, A. Salta, H. Prendinger, "UAV-based situational awareness system using deep learning," IEEE Access, 2019, 7, 122583-122594.

[18] E. Lygouras, N. Santavas, A. Taitzoglou, K. Tarchanidis, A. Mitropoulos, A. Gasteratos, "Unsupervised human detection with an embedded vision system on a fully autonomous UAV for search and rescue operations," Sensors, 2019, 19(16), 3542.

[19] M. Bejiga, A. Zeggada, A. Nouffidj, F. Melgani, "A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery," Remote Sensing, 2017, 9(2), 100

[20] M. Radovic, O. Adarkwa, Q. Wang, "Object recognition in aerial images using convolutional neural networks," Journal of Imaging, 2017.

[21] M. Pobar, M. Ivasic-Kos, "Active Player Detection in Handball Scenes Based on Activity Measures," Sensors 20 (5), 1475

[22] M. Buric, M. Pobar, M. Ivasic-Kos, Object Detection in Sports Videos. In Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018.

[23] M. Buric, M. Pobar, M. Ivasic-Kos, "Adapting YOLO network for a ball and player detection," Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019), 2019/2, pp. 845-851

[24] K. Yun, L. Nguyen, T. Nguyen, D. Kim, S. Eldin, A. Huyen, E. Chow, "Small target detection for search and rescue operations using distributed deep learning and synthetic data generation," in Pattern Recognition and Tracking XXX (Vol. 10995, p. 1099507), International Society for Optics and Photonics, 2019.

[25] M. Buric, G. Paulin, M. Ivasic-Kos, "Object Detection Using Synthesized Data," ICT Innovations 2019 Web proceedings, (14) pp. 110-124.

[26] M. Ivasic-Kos, Mate Kristo and Miran Pobar, "Human Detection in Thermal Imaging Using YOLO," in Proceedings of the 5th ACM International Conference on Computer and Technology Applications, ICCTA 2019, NY, USA, pp.20-24.

[27] M. Kristo, M. Ivasic-Kos, M. Pobar, "Thermal Object Detection in Difficult Weather Conditions Using YOLO," IEEE Access 8, 2020, 125459-125476

[28] E. Bondi, F. Fang, M. Hamilton, D. Kar, D. Dmello, J. Choi, R. Nevatia "Spot poachers in action: Augmenting conservation drones with automatic detection in near real-time," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[29] A. Al-Kaff, M. J. Gómez-Silva, F. M. Moreno, A. de la Escalera, J. M. Armingol, "An appearance-based tracking algorithm for aerial search and rescue purposes," Sensors, 2019, 19(3), 652.

[30] S. E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, "Convolutional pose machines," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724-4732.

[31] M. Buric, M. Ivasic-Kos and M. Pobar, "Player Tracking in Sports Videos," 2019 IEEE International Conference on Cloud Computing Technology and Science, Sydney, Australia, 2019, pp. 334-340.

[32] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-CNN: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, 91-99.

[33] S. Sambolek, M. Ivašić-Kos, "Detection of toy soldiers taken from a bird's perspective using convolutional neural networks," in International Conference on ICT Innovations, Springer, Cham, 2019, pp. 13-26.

[34] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117-2125.

[35] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, "Microsoft COCO: Common objects in context," in Proceedings of the European Conference on Computer Vision - ECCV, 2014.

[36] https://github.com/facebookresearch/detectron2/blob/master/MODEL_ZOO.md Accessed: 18. Mar