



Methodology for Image Analysis in Airborne Search and Rescue Operations

Francesco Ciccone^(✉) , Antonio Bacciaglia , and Alessandro Ceruti 

Department of Industrial Engineering – DIN, University of Bologna, Viale del Risorgimento 2,
40132 Bologna, Italy
francesco.ciccone2@unibo.it

Abstract. Nowadays, Search and Rescue operations can be performed using manned or unmanned Aerial Vehicles. In this latter case, compact cameras are mounted onboard and a bird's eye view is available to find the missing person. However, the analysis of the video frames can be very challenging and dull for the operators. In this context, the use of graphical methodologies can boost the searching operations and improve the process. In this study, a methodology based on the object detector Yolov5 is introduced: the performances in detecting small objects such as persons in aerial images are evaluated. These algorithms implement shallow layers of the feature extractor to increase the spatial-rich features and help the detector to find small objects. Finally, detection algorithms are tested using a video simulating a scenario for Search and Rescue operations. The filtering of frames containing false positives, is carried out using a classical graphical tool such as the Hamming distance.

Keywords: Image analysis · Graphical methodologies · Object detection · Aerial images · SAR operations

1 Introduction

The use of Unmanned Aerial Vehicles (UAV) for Search And Rescue (SAR) operations has become of paramount importance to scout a huge area in a brief time, giving the possibility to accelerate the searching operations. UAVs well suits D3 (dull, dirty, dangerous) operations. The videos filmed using UAVs are then analyzed by specialized operators to find the missing person: a Ground Control Station is equipped with a video window reproducing in real time what framed by the camera on board. Capturing people from aerial images can be very challenging and sometimes distraction or loss of attention leads to a failed detection. Human factors studies suggest that high attention can be assured by humans only for a limited time, especially when the images are without distinctive features (e.g., sea, or uniform background). Handcrafted graphical methodologies can be adopted to help operators, using a simple RGB camera. The papers [1–3] implemented a computer-assisted spectral signature software that analyzes each image in a dataset to detect user-defined spectral signatures based on the clothes worn

by the missing individual. Hyperspectral cameras can provide extra information about the spectral signature and can be used in conjunction with appropriate software. These sensors can help in the rescue of missing people [4, 5]. But the aforementioned methodologies require specialized operators and could be difficult to implement for real-time applications.

Recently, new automatic graphical methodologies have been developed thanks to Artificial Intelligence. Particularly interesting for image analysis are the Convolutional Neural Networks (CNNs): these are substantially automatic feature extractors, in which at the input image are applied filters to extract specific characteristics. With respect to classical handcrafted image analysis, Deep Learning object detection algorithms can be a different solution: low-cost, easy to use, and capable to be implemented in real-time. Object detection algorithms combine two tasks, localizing and classifying one or more objects in an image. The model processes the input first computing a bounding box around the object of interest, and then proceeding with the classification in one or more categories. There exist two main families of object detector models: two-stage and one-stage detectors. The formers are mainly composed of three modules, a Region Proposal that generates and extracts candidate bounding boxes, a CNN with the aim of features extractor, and finally a classifier. Model like R-CNN [6], Fast R-CNN [7], Faster R-CNN [8], RetinaNet [9] belong to this family. One-stage detectors are characterized by end-to-end detection (bounding boxes are computed directly on the grid cells) without any explicitly extracting object proposals: SSD model [10] and YOLO [11] are examples. Generally, two-stage detectors achieve better accuracies while one-stage detectors are much faster, giving the possibility to be implemented for real-time applications.

Referring to aerial images, the small object detection problem is particularly challenging for Deep Learning algorithms. This is because small objects need to be detected, recognised, and extracted from the background, requiring high accuracy. CNNs include some low-level abstractions of the images, combining them in final objects that must be detected. When the input image gradually passes the hidden layers, its resolution decreases making it harder to detect objects that occupy few pixels inside the image. This is the reason why CNNs are so powerful in feature extraction but also weak in detecting small objects.

In this paper, a comprehensive and in-depth analysis of person detection algorithms from aerial images for SAR operations is presented. In the following, the state-of-the-art object detector Yolov5 is trained on the Stanford Drone Dataset, and the outcome of tests carried out is evaluated.

2 Related Works

Using UAVs for SAR operations can be crucial: this is supported by the fact that UAVs can offer the possibility to cover large areas in short times with a bird's eye view. In [12] it is discussed the importance of UAV in searching missing persons for three operational paradigms: sequential, remote-led, and base-led operations, according to the ground mobility and the general scenario in which conduct the search.

Furthermore, the source [1] compares the computer-assisted approach to the manual way during SAR training. According to the study, manual interpretation techniques resulted in faster location of missing objects, but at the expense of weariness and lack of interest over time. On the other hand, the group employing computer-assisted methods became more efficient over time and better able to detect targets. However, by utilizing deep learning techniques, computer-assisted methods - the majority of which need spectral-signature software - can be replaced by autonomous object detectors.

The work [13] describes a successful algorithm to support SAR operations called SARUAV model. This CNN-based software analyses remotely the aerial images captured by UAV, detecting the possible location of the missed person. Even if the software is fast in the processing of all the images, the entire system isn't real-time: in this case about four hours passed from the UAV images acquisition to the final response of the model.

The authors of the paper [14] implemented the one-stage object detector EfficientDET [15] training it on the high-resolution aerial database HERIDAL [16], and demonstrated to reach about 93% accuracy.

In [17] a semantic segmentation model [18] called MAGI was developed: it has been trained on the custom UAV Mosaicking and Change Detection (UMCD) dataset, with 90% acquiring accuracy.

A real-time object detector called RGDNet was developed in [19]: the model uses a Faster R-CNN processing RGB aerial images and Point Cloud Data. The model was proposed for SAR operations, reaching an 88% accuracy, and a 0.9 s processing time, with 1 fps of inference.

Finally, an investigation between Faster R-CNN, Yolov4, RetinaNet, and Cascade R-CNN was conducted in [20]. The four object detectors were trained firstly on the VisDrone dataset [21], and successively on a custom dataset for SAR operations called SARD. The results showed that higher accuracy can be reached by Yolov4: this model was analysed in deep, showing how the resolution of the images influence the final accuracy of the model. The best result in terms of average precision (AP) was obtained using an image resolution of 832×832 pixels, but training on 320×320 pixels images, the model reached 10,3 fps speed, becoming a good solution for real-time applications.

2.1 The Problem of Detecting Small Objects

When discussing human detection using aerial photos, the challenge of recognizing tiny things must be considered. There exist two definitions of small objects: the first is referred to the real world, in which an object is considered small comparing its physical size with other objects. Another definition, more appropriate considering deep learning algorithms, is given in the evaluation metrics used by the MS COCO dataset [22]: following this source, "small objects" are objects occupying an area less than or equal to 32×32 pixels.

According to [23], detecting small objects is difficult for three main reasons:

- Is difficult to distinguish them from the background;
- Is required higher precision;
- The majority of detectors are tuned for large objects detection.

Moreover, CNNs act as feature extractors, computing the input image through many layers. The convolution and pooling operations filter the images extracting information but lowering the resolution. In this context, it can be said that deep layers contain semantic-rich features and coarse resolutions, thus are more suitable for the detection of larger objects. On the contrary, the first layers contain spatial-rich features and have higher resolutions being more appropriate for detecting small objects. For this reason, object detectors have been developed in order to take advantage of the different layers of the feature extractor CNN (also called backbone) concatenating the spatial-rich and semantic-rich information as in the Feature Pyramid Network (FPN) [24]. Therefore, acting on the hierarchical features of the different layers is a proper way to improve an object detector for small objects.

Another way for assisting such models in spotting tiny objects is data augmentation. It is worth noting that all the strategies for enhancing the dataset are referred to as “data augmentation”. In particular, as mentioned in [25], increasing the types and numbers of small objects samples in the dataset helps the detector accuracy. This can be obtained by training the detector on a specific dataset such as aerial images.

3 The Proposed Methodology

A preliminary study of a potential approach to enable detection of missing person from aerial photos for Search and Rescue operations is described in this research. The study aims to build a system capable to detect missing persons from UAV in an automated and robust way. The requirements of such image analysis models are:

- The possibility to be implemented and stored in an embedded system, with limited: size occupied in memory, computational performances required, current consumption, size and weight (installation requirements);
- The capability to be implemented for real-time applications, preferring an algorithm able to compute images at high fps; this is fundamental to extend these algorithms also to manned aerial platforms, where higher flying speeds could be necessary;
- The detection of persons from aerial images with high accuracy and low detection errors.

The methodology herein developed is based on the use of Yolov5, an algorithm representing the state-of-the-art in terms of object detectors. This model is available in different versions (Yolov5s, Yolov5m, Yolov5l, Yolov5x) the version “s” is the fastest and lightest, the version “x” is the most accurate and heaviest [26].

The metrics used to evaluate the models' performance are then briefly introduced:

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \end{aligned} \quad (1)$$

where TP, TN, FP, FN mean respectively *True Positive*, *True Negative*, *False Positive*, and *False Negative*. So, in this context, Precision measures how accurate is the prediction, while Recall measures how good the model finds all the possible detections. Finally, the mAP is called *Mean Average Precision* and is defined as the average (per class) of the area of the curve between the Precision and the Recall. The detector establishes a true positive when the proportion between the area of the bounding box computed by the model and the ground truth bounding box is greater than 0.5 (this is the definition of *Intersection over Unit*, *IoU*). So, it is possible to have mAP.5 (mean Average Precision computed considering $\text{IoU} = 0.5$), but also mAP.5: .05:.95 (mean Average Precision computed considering IoU from 0.5 to 0.95 with a step of 0.05).



Fig. 1. Image example of the stanford drone dataset used in [27].

The model is trained on the Stanford Drone Dataset (SDD) [27]: Fig. 1 shows an example of the dataset. SDD is a large-scale dataset with videos containing different targets (pedestrian, bikers, skateboarders, cars, buses, and golf carts) that navigate on a university campus (open dataset). The scenes are captured with a 4K camera mounted on a quadcopter with an average altitude of 80 m.

The dataset has been augmented by applying horizontal flip, random shear, and random exposure adjustment. The resolution of the images was modified to 640×640 pixels. Moreover, the targets were modified in order to replace “pedestrian”, “bikers”, and “skateboarders” in the unique class “person”. Finally, the “file.yaml” containing the labels for each image accepted by Yolov5 was generated.

Moreover, in this paper a modified version of Yolov5 (called Yolov5-P2) has been trained: in this case, the detection head has been passed also (for detection head is intended the last part of the model concerning the detection) the shallow layer (P2) in order to account for higher resolution images containing spatial-rich features. The paper [28] suggests that this expedient has led to a 30% improvement in precision.

Finally, the best model has been implemented on a test video in order to verify its efficacy for a real-world problem.

4 Results

The models trained during this study have been Yolov5x and Yolov5s. The training was conducted for 300 epochs with a batch size of 8, starting from the versions pre-trained on the MS-COCO dataset. The dataset used for the training phase was the Stanford Drone Dataset. In Fig. 2 the metrics obtained for Yolov5x are presented. The entire training phase has occupied 24 h on Google Colab using cloud computing with a Tesla T4 16 Gb GPU. Figure 3 presents an example of the detections obtained from the validation. The results show 81.3% mAP.5, with 38% mAP.5:.95.

The training of the version Yolov5s has been faster, occupying approximately 10 h for 300 epochs and a batch size of 8. Figure 4 and Fig. 5 show the relative results obtained. Yolov5s has achieved 78.7% mAP.5, with 29.9% mAP.5:.95.

In order to take account for spatial-rich features, two other versions of Yolov5x and Yolov5s have been trained, respectively, Yolov5x-P2 and Yolov5s-P2. These versions were trained for 450 epochs because the pre-training of the entire models was not available. Table 1 contains a comparison between the models. However, the training of the versions Yolov5x-P2 and Yolov5s-P2 has not yielded appreciable results. Both models show lower precision and recall values than the other versions, which is unexpected. Versions-P2 have been stopped before finishing the training (in Table 1, there is no training time available).

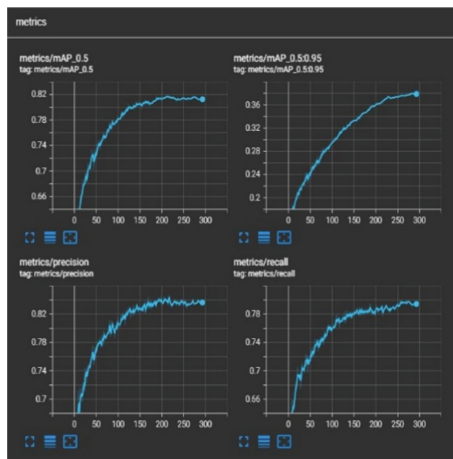


Fig. 2. Metrics obtained at the end of the training of Yolov5x on the stanford drone dataset.



Fig. 3. Example of Yolov5x for person detection.

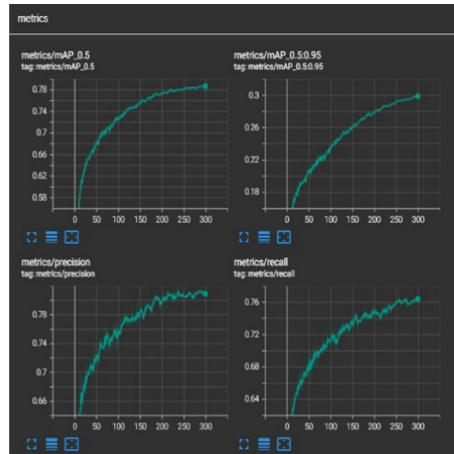


Fig. 4. Metrics obtained at the end of the training of Yolov5s on the stanford drone dataset.

Finally, the best model (Yolov5x in terms of mAP) was implemented on a test video in order to simulate a real-world problem. The scene is captured on a helicopter and shows a seaside scene in which persons are walking on a bay. The model can correctly detect the persons, but in the different frames, lots of false-positive are present. Figure 6 shows an example.



Fig. 5. Example of Yolov5s for person detection.

Table 1. Comparison of the results

Model	Dataset	mAP.5	mAP.5:95	Precision	Recall	Size (Mb)	Inference time (s)	FPS	Training time (hours)
Yolov5x	SDD	81.23%	37.9%	83.12%	79.8%	173	0.18	5	24
Yolov5s	SDD	78.7%	29.91%	81%	76.38%	14.5	0.02	50	10
Yolov5x-P2	SDD	69%	22.7%	73.8%	69.9%	360.7	0.113	8.84	/
Yolov5s-P2	SDD	73.5%	27.1%	77.8%	72.7%	360.7	0.118	8.47	/

An ideal image analysis methodology should help the SAR operators in the detection of missing persons with the highest accuracy possible and avoiding false positives to spare time and resources. During the view of the processed video, has been noted that the model correctly detected the persons for the great part of the frames, while the false positives were displayed in a randomly way. The methodology has been improved in such a way to mitigate this problem: to capture a sort of “average detection”, the video processed by the object detector model was split into frames and for each of the resulting images was generated the respective “hash file” (the image is transformed in an alphanumeric string) and then the Hamming distance [29] has been computed. The Hamming distance is a metric for comparing two binary data of equal length: it is defined as the number of bit positions in which the bits are different.

$$d_H = \sum_{i=0}^n (p_i \oplus q_i) \tag{2}$$

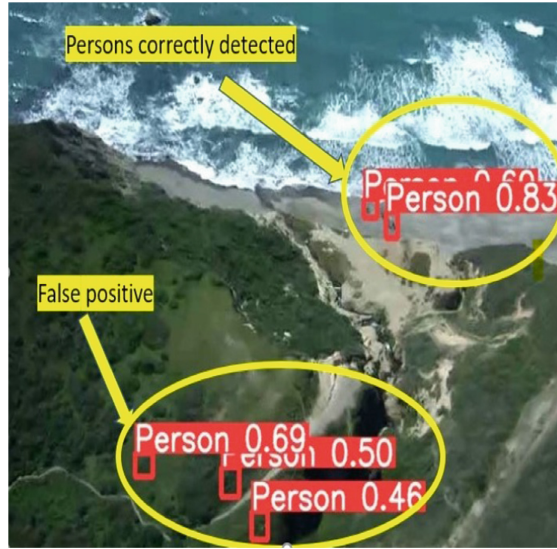


Fig. 6. An example of detection of persons with a real-world scenario performed by Yolov5x.

where the \oplus symbol denotes the logical *exclusive-or* (*XOR*) operator, p_j and q_j are the corresponding bits.

Hamming distance is typically used to find duplicates or the most similar data. The authors propose its application to the entire sequence of frames, selecting the images with a Hamming distance less or equal to 0.05. Considering the 32 characters of the hash file, this means that the frames are considered “duplicates” if they differ by only a single character in the hash file. In this way, it is possible to find the “similar” images, which means to compute a sort of “average detection”, eliminating almost all frames containing false positives. The initial video was composed of 510 frames. The application of the Hamming distance has reduced the images to a final number of 13. These final images represent the frames “repeated” during the video and since the True Positives are correctly detected for almost all the frames, the Hamming distance indirectly works as a sort of filter. Figure 7 shows an example of this result, the images are divided into *Frames with no detection*, *Duplicates obtained with Hamming* and *Original frames with detections*, where *Duplicates* indicate the images most similar to the *Original*.

Finally, in Table 2 are shown the Precision and Recall computed for the detections before and after applying the Hamming distance. The result proves the efficiency of the method. The most important parameter is the Precision since the aim is to reduce the number of false positives during the detection.

Table 2. Comparison of the results

Without hamming distance		With hamming distance	
Precision	Recall	Precision	Recall
46.3%	66%	60%	65%

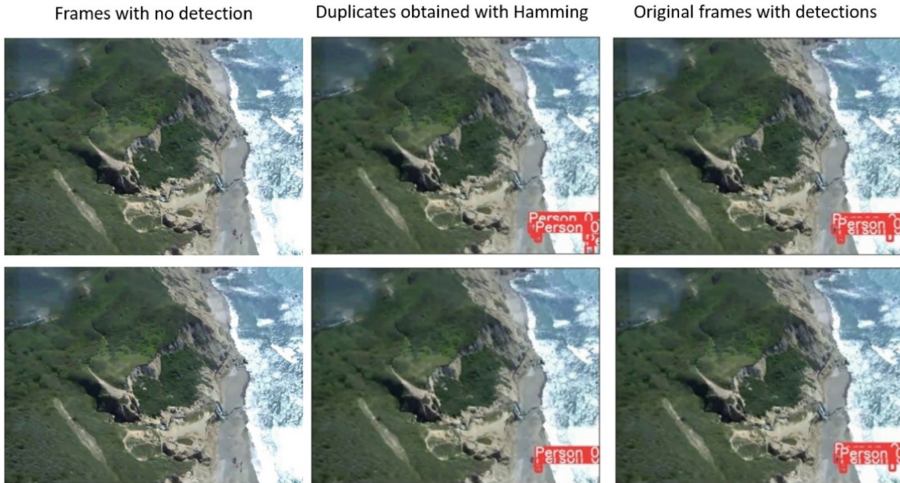


Fig. 7. Examples from the results of the application of the Hamming distance to the test video. The Hamming distance acts as a filter.

5 Conclusion and Future Works

This work should be considered a preliminary study in which authors developed a prototype of a graphical methodology to detect missing persons during Search and Rescue operations. Different versions of the Yolov5 object detector have been trained on Stanford Drone Dataset to study the capacity of these models to detect persons from aerial images. As expected, the best version with the higher accuracy is resulted to be Yolov5x with 81.23% mAP, but for real-time applications, Yolov5s reached 50 FPS. In the following, the trained models have been tested on a real-world scenario. The results showed the detection of many false positives. To overcome this problem, the authors introduced a second step in the methodology, applying the Hamming distance as a filter to select the images containing the right detections. This application has proven to be effective.

Many other studies should be conducted to deeper analyse this methodology. First, Yolov5x-P2 and Yolov5s-P2 have not given the results expected: this requires further research. Moreover, the application of such models should be investigated in real-time, using hardware such as Nvidia Jetson Tx2, Google Coral USB or Intel Neural Stick. This will help the study to analyse the performance of Yolov5 on embedded systems and evaluate if the current hardware capabilities are sufficient for effective image analysis

missions. Moreover, the effects of light, sunshine, shadows must be investigated to better understand the robustness of these algorithms.

References

1. Weldon, W.T., Hupy, J.: Investigating methods for integrating unmanned aerial systems in search and rescue operations. *Drones* **4**, 38 (2020)
2. Proft, J., Suarez, J., Murphy, R.: Spectral anomaly detection with machine learning for wilderness search and rescue. In: 2015 IEEE MIT Undergraduate Research Technology Conference (URTC). pp. 1–3. IEEE, Cambridge, MA, USA (2015)
3. Morse, B.S., Thornton, D., Goodrich, M.A.: Color anomaly detection and suggestion for wilderness search and rescue. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12, p. 455. ACM Press, Boston, Massachusetts, USA (2012)
4. Nunez, A.S., Mendenhall, M.J.: Detection of Human Skin in Near Infrared Hyperspectral Imagery. In: IGARSS 2008–2008 IEEE International Geoscience and Remote Sensing Symposium. p. II-621–II-624. IEEE, Boston, MA, USA (2008)
5. Simard, J.-R., Mathieu, P., Fournier, G.R., Larochelle, V., Babey, S.K.: Range-gated intensified spectrographic imager: an instrument for active hyperspectral imaging. Presented at the AeroSense 2000, Orlando, FL 5 Sep. 2000
6. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587 (2014). <https://doi.org/10.1109/CVPR.2014.81>
7. Girshick, R.: Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448 (2015). <https://doi.org/10.1109/ICCV.2015.169>
8. 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 (2016)
9. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327 (2020). <https://doi.org/10.1109/TPAMI.2018.2858826>
10. Liu, W., et al.: SSD: single shot MultiBox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788. IEEE, Las Vegas, NV, USA (2016)
12. Goodrich, M.A., Cooper, J.L., Adams, J.A., Humphrey, C., Zeeman, R., Buss, B.G.: Using a Mini-UAV to support wilderness search and rescue: practices for human-robot teaming. In: 2007 IEEE International Workshop on Safety, Security and Rescue Robotics. pp. 1–6. IEEE, Rome, Italy (2007)
13. Niedzielski, T., Jurecka, M., Miziński, B., Pawul, W., Motyl, T.: First successful rescue of a lost person using the human detection system: a case study from Beskid Niski (SE Poland). *Remote Sens.* **13**, 4903 (2021)
14. Dousai, N.M.K., Loncaric, S.: Detection of humans in drone images for search and rescue operations. In: 2021 3rd Asia Pacific Information Technology Conference. pp. 69–75. ACM, Bangkok Thailand (2021)
15. Tan, M., Pang, R., Le, Q.V.: EfficientDet: scalable and efficient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 10778–10787 (2020). <https://doi.org/10.1109/CVPR42600.2020.01079>

16. Gotovac, S., Zelenika, D., Marušić, Ž, Božić-Štulić, D.: Visual-based person detection for search-and-rescue with UAS: humans vs. machine learning algorithm. *Remote Sens.* **12**, 3295 (2020)
17. Avola, D., Pannone, D.: MAGI: multistream aerial segmentation of ground images with small-scale drones. *Drones* **5**, 111 (2021)
18. Liu, X., Deng, Z., Yang, Y.: Recent progress in semantic image segmentation. *Artif. Intell. Rev.* **52**(2), 1089–1106 (2018). <https://doi.org/10.1007/s10462-018-9641-3>
19. Kim, J., Cho, J.: RGDNet: efficient onboard object detection with faster R-CNN for air-to-ground surveillance. *Sensors* **21**, 1677 (2021). <https://doi.org/10.3390/s21051677>
20. Sambolek, S., Ivasic-Kos, M.: Automatic person detection in search and rescue operations using deep CNN detectors. *IEEE Access* **9**, 37905–37922 (2021)
21. Zhu, P., et al.: Detection and tracking meet drones challenge. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). <https://doi.org/10.1109/TPAMI.2021.3119563>
22. 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
23. Tong, K., Wu, Y., Zhou, F.: Recent advances in small object detection based on deep learning: a review. *Image Vis. Comput.* **97**, 103910 (2020)
24. Vo, X.-T., Tran, T.-D., Nguyen, D.-L., Jo, K.-H.: Stair-step feature pyramid networks for object detection. In: Jeong, H., Sumi, K. (eds.) *IW-FCV 2021*. CCIS, vol. 1405, pp. 168–175. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-81638-4_13
25. Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., Cho, K.: Augmentation for small object detection. In: *9th International Conference on Advances in Computing and Information Technology* (2019)
26. Jocher, G., Stoken, A., Ayush Chaurasia, Borovec, J., NanoCode012, TaoXie, Yonghye Kwon, Kalen Michael, Changyu, L., Jiacong Fang, Abhiram V, Laughing, Tkianai, YxNONG, Skalski, P., Hogan, A., Jebastin Nadar, Imyhxy, Mammana, L., AlexWang1900, Fati, C., Montes, D., Hajek, J., Diaconu, L., Minh, M.T., Marc, Albinxavi, Fatih, Oleg, Wanghaoyang0106: ultralytics/yolov5: v6.0 - YOLOv5n “Nano” models, Roboflow integration, TensorFlow export, OpenCV DNN support. Zenodo (2021)
27. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: human trajectory understanding in crowded scenes. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9912, pp. 549–565. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_33
28. Zhan, W., et al.: An improved Yolov5 real-time detection method for small objects captured by UAV. *Soft. Comput.* **26**(1), 361–373 (2021). <https://doi.org/10.1007/s00500-021-06407-8>
29. Hamming, R.W.: Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, 147–160 (1950)