


Article

Multimodel Deep Learning for Person Detection in Aerial Images

Mirela Kundid Vasić ^{1,2,*}  and Vladan Papić ²

¹ Faculty of Mechanical Engineering, Computing and Electrical Engineering, University of Mostar, 88000 Mostar, Bosnia and Herzegovina

² Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, 21000 Split, Croatia; vpapic@fesb.hr

* Correspondence: mirela.kundid.vasic@fsre.sum.ba

Received: 17 August 2020; Accepted: 4 September 2020; Published: 7 September 2020



Abstract: In this paper, we propose a novel method for person detection in aerial images of nonurban terrain gathered by an Unmanned Aerial Vehicle (UAV), which plays an important role in Search And Rescue (SAR) missions. The UAV in SAR operations contributes significantly due to the ability to survey a larger geographical area from an aerial viewpoint. Because of the high altitude of recording, the object of interest (person) covers a small part of an image (around 0.1%), which makes this task quite challenging. To address this problem, a multimodel deep learning approach is proposed. The solution consists of two different convolutional neural networks in region proposal, as well as in the classification stage. Additionally, contextual information is used in the classification stage in order to improve the detection results. Experimental results tested on the HERIDAL dataset achieved precision of 68.89% and a recall of 94.65%, which is better than current state-of-the-art methods used for person detection in similar scenarios. Consequently, it may be concluded that this approach is suitable for usage as an auxiliary method in real SAR operations.

Keywords: convolutional neural networks; aerial images; person detection; search and rescue

1. Introduction

Object detection is one of the elementary problems in computer vision. When an object of interest is small, due to its limited resolution and information, the problem becomes even more challenging. In this paper, we focus on small object detection in the specific problems of Search And Rescue (SAR) operations where the object of interest is a person. The main goal of SAR operations is to provide search, technical rescue, and provision of aid to people who are lost and possibly in danger. This operation takes many forms, but in this paper, we are focused on the SAR operations in the wilderness. SAR operations include a set of actions, where the first step, which also takes most of the time, is the search of suspicious terrain. This step typically has to be physically performed by rescuers. Considering nonurban terrain, it is not uncommon that a part of the terrain is inaccessible or hardly accessible.

In order to improve search performance and save time, it is essential to introduce new solutions based on the advantages of new technologies such as Unmanned Aerial Vehicles (UAVs), colloquially known as drones [1,2]. Currently, due to the low cost and ease of operation of drones, they are increasingly being used as a support in some industries. Behind the standard operations, e.g., for the inspection of damaged assets or monitoring crops for farmers, drones are used even for delivery services. The use of drones for supporting search functions within SAR operations could bring significant benefits due to their ability to thoroughly survey a wide geographical area from a “bird’s eye view”. As a result, rescuers no longer have to physically access all hazardous and difficult-to-reach locations. During a drone survey of a suspicious area, the ground is photographed

sequentially, which consequently results in a large number of high resolution images. Because of the high altitude, a person in these images covers very few pixels (around 0.1% of the image). All of these images must be inspected by a human in order to determine whether they contain an object of interest or not. In other words, a member of a search team has to visually check whether a lost person is present in the covered area. Due to the large number of images and extremely small size of the object of interest, this task is very demanding and time consuming, and omissions are very likely. A system that could automatically process recorded data would be of a tremendous contribution to the efficiency of SAR operations. Increasing efficiency means multiple benefits, such as optimization of human resources, reducing duration, reducing costs, reducing the risk of injuries to rescuers, facilitating search in hazardous or even inaccessible terrain, and increasing the possibility to find a lost person and to provide adequate assistance on time.

Motivated by the idea of developing a system that can be used as an auxiliary method in search and rescue operations, in this paper, we propose a method for person detection in aerial images that is more effective than current state-of-the-art methods on the HERIDAL dataset (<http://ipsar.fesb.unist.hr/HERIDAL%20database.html>). In order to demonstrate its effectiveness, several deep learning methods for this particular problem were analyzed and experimentally compared. Some of them performed well in terms of the recall measure, while others have proven to be better in terms of precision. This means that increasing true positive detections also increases false positive detections. In order to overcome this issue, it has been concluded that the architecture based on the multimodel deep learning approach could achieve noticeable improvements in terms of system accuracy. Therefore, our main contributions are as follows:

- (i) We propose a novel multimodel approach for person detection in aerial images in order to support SAR operations. The proposed model combines two different convolutional neural network architectures in the region proposal stage, as well as in the classification stage;
- (ii) We introduce the usage of contextual information contained in the surrounding area of the proposed region in order to improve the results in the classification stage;
- (iii) Our proposed approach achieves better results compared with state-of-the-art methods on the HERIDAL dataset.

The paper is organized as follows: Section 2 provides a review of the relevant literature and methods that can be used with this type of problems, while Section 3 describes the dataset and all the methods used including our multimodel approach. Section 4 contains all experimental results together with the analysis in detail, and Section 5 concludes the paper and discusses future intentions for research and potential improvements.

2. Related Work

2.1. Small Object Detection

In recent years, small object detection, as a part of computer vision, has received great attention because it is widely applied in people's lives including autonomous driving, robotics, video surveillance, and the manufacturing industry. The concept of small objects refers to those objects that are represented by a small number of pixels (less than 1% of the image area). In addition to normal object detection challenges, small object detection has additional challenges because it is hard to distinguish it from the background. Observing the object detection method's evolution throughout history, it may be stated that one of the most important milestones was in 2012 due to the appearance of convolutional neural networks [3,4]. The use of convolutional neural networks has made noteworthy improvements in object detection performance due to its ability to produce powerful feature representations [5]. Various deep learning based approaches for object detection, those performed in one stage such as Single Shot MultiBox Detector (SSD) [6] and YOLO [7], as well as those performed in two stages (region proposal and classification stage) such as Fast R-CNN [8] and

Faster R-CNN [9], use a combination of convolution and pooling layers. The feature maps of small objects become even smaller with each pooling layer. For example, objects measuring 32×32 pixels in size after five pooling layers would be represented with just one pixel and can be easily missed. This is the main reason why state-of-the-art methods for object detection still struggle in detecting small objects [10]. In order to overcome this issue, standard deep learning methods need to be modified. Modification has been performed on both two stage detector [11,12] and one stage detectors [13–17] by different authors and in different ways.

While most papers deals with the detection of the objects presented with horizontal bounding boxes, there are some approaches for multi-oriented object detection in aerial images [18,19]. Some of them are even based on predicting the axis of the object of interest instead of using predefined anchors, which reduces the computational complexity [20]. Although our objects of interest (persons) can be physically arbitrarily oriented, they cannot generally be classified as orientated object in aerial images. This is caused by the high altitude of recording, which makes the aerial footprints of persons more often squared with different heights and widths than having a significant difference in box orientation.

2.2. Search and Rescue Operations

In SAR operations in the wilderness, the primary focus is to locate a missing person. The fundamental issue is to determine the search area where the person may be located based on expert knowledge [21]. Generally, the search area is a wide-ranging area with a complex environment, and it is nearly impossible to search the entire area physically within a reasonable time. Undoubtedly, drones have profoundly improved search and rescue activities providing the potential for the automated and reliable survey of suspicious terrain. In order to use its full potential, it is essential to properly use information from the drone sensors. This implies the need for developing a system with the ability of person detection in the gathered images. Person detection in aerial images gathered with drones during search and rescue missions could be considered one of the most challenging tasks in object detection because it deals with several problems. The first problem is that the product of recording is a large number of high resolution images that need to be transmitted to the control center for processing. To address this problem, Musić et al. used the compressive sensing algorithm in order to decrease the amount of image data [22] for transmission and also reconstructed the initial image for further processing using mean shift clustering. The second problem is that the objects of interest (person) in those images are limited in pixels and resolution. The size of the object in an image actually depends on the drone altitude during recording. With increasing altitude, the level of object detail decreases, while on the other hand, the size of the observation area becomes wider, which is essential for SAR operations because it provides faster area surveying. Usually, in real SAR operations, images are taken at the altitude of about 50 m, which results in images where the person covers a very small part of an image (around 0.1%). Additionally, objects of interest vary in position, orientation, viewpoint, cloth color, scale, etc., and can be camouflaged within the environment causing the lack of details that distinguish the object from the background. That is why many published papers related to drone usage in SAR operations rely on the addition of Thermal Infrared (TIR) cameras, which allow warm bodies to be seen distinctly from their surroundings [23,24]. However, in the area where we performed our research, using thermal cameras for person detection is often not appropriate since in summer months, the ground temperature may be even higher than the human body temperature. In this case, we are limited to the detection of objects in the visible spectrum. Marušić et al. in [25] performed person detection on UAV images from the HERIDAL dataset in the visible spectrum using Faster R-CNN as a backbone. Furthermore, Božić-Štulić et al. [26] dealt with the same problem. They used a visual attention algorithm for the detection of salient objects in images in order to reduce the large-scale search space. Then, binarization and connected component labeling were performed followed by binary classification for the top 300 selected region proposals. In the end, non maxima suppression was performed for reducing false positive detections by clustering proposals by spatial closeness, and the achieved detection rate was 88.9%, while the precision was 34.8%, which make their

methods currently the state-of-the-art for this particular problem. Since our research was performed on the same dataset, it is most appropriate to compare our results with those mentioned above and presented in [25,26].

2.3. Using Contextual Information

Usually, within an object detection task, only isolated interior object features are considered essential, while environmental features are ignored. However, the human eye works quite the opposite, since surrounding contextual information facilitates object recognition. Context implies any information that can somehow contribute to the understanding of the scene and objects within the scene. It may be contained in image illumination (shadows, contrasts, etc.), geographical performance (GPS location, terrain type, elevation, etc.), semantic content (scene category, expected event, etc.), or the time frame (recording time, surrounding images, etc.) [27]. Although machines and humans have qualitatively different context representations, context can be a valuable source of information about object type even in machine learning. Different methods of introducing contextual information in segmentation and detection have been presented by Mottaghi et al. [28]. Moreover, in some examples, contextual information may be more conducive to object recognition than features of the object itself, especially in low resolution images [29]. It can be useful even for detecting some missing objects from the scene [30]. Usually, contextual information is used for improving detection accuracy by concatenating multi-scale features from different layers or by skipping pooling to extract information on multiple scales [31]. In the problem of detecting people in aerial images of nonurban terrains, there is no uniform shape of the object or location where a lost person could be found. Statistical data gathered in real SAR operations with probabilities for different types of terrain and other topological features exist [21], but using this kind of information and confirming hypothesis has some serious issues. For the use of this type of contextual information in problems of person detection in aerial images as a support of SAR operations, we should have a dataset of images collected in real SAR operations. Considering the fact that recording during SAR operations produces a large number of images that do not contain any objects of interest (in this case, persons) and simultaneously potentially only a few images that contain the object of interest, it is implied that a dataset with enough positive examples (images with a lost person) cannot be attained, at least not in a reasonable time period. Therefore, with the available dataset, our assumption is that, besides the features of the proposed region, including contextual features of surrounding regions in the classification stage could result in more reliable classification.

3. Proposed Methods

3.1. Dataset Description

Most published papers for the implementation and evaluation of small object detection systems use standard publicly available datasets such as VisDrone [32], PASCAL VOC [33], or DOTA [34]. However, the problem of detecting persons in aerial imagery for search and rescue purposes is very specific, so it is not possible to use standard image datasets. In order to solve this problem, research from the University of Split [26] published a dataset named HERIDAL, which is also used in this paper. The dataset contains 1677 images of wilderness at various locations acquired from an aerial perspective with an Unmanned Aerial Vehicle (UAV), also known as a drone, with a high definition camera. The dataset consists of three folders: Patches (image parts 81×81 in size), Test images, and Train images. There are 1583 images in the Train images and 101 images in the Test images. All of them are labeled, and the labels are in the .xml format. In the Patches folder, there are 29,050 positive patches that contain objects (person) and 39,700 negative patches without objects. Images are taken with a high resolution camera on the DJI Phantom 3, vertically at a 50 m altitude with the selected resolution of 4000×3000 px. In order to compile a dataset that would be a realistic representation of a

real search and rescue operation, the authors used statistical data and expert knowledge about SAR operations [21].

3.2. Classification Stage

All evaluated and proposed algorithms (except the SSD method) in this paper depend on a two stage approach: the region proposal stage and the classification stage. The idea behind two stage object detectors is to extract the most prominent regions in an image and classify them as person or non-person using a convolutional neural network. Using CNN for classification tasks can deliver extremely competitive results, comparable to human level performance [35]. For the region proposal stage, the following methods are used: edge boxes, mean shift, Region Proposal Network (RPN), and feature pyramid network. Then, in the second stage, binary classification is carried out on each proposed region. This is achieved by using a small convolutional neural network that we empirically designed for classification problems using patches from the HERIDAL dataset for training and testing. This network consists of five convolutional layers with max pooling layers on the first and second convolutional layer. In max pooling layers, the 3×3 filters with stride 3 are used. The ReLU activation is used in all convolutional layers, and the number of convolutional filters is 32 in the first two layers and 64 in the other layers. A stochastic gradient descent algorithm is used as an optimization function with a learning rate of 0.001. The network architecture is shown in Figure 1. Using a designed network, different patches were efficiently classified into two classes, person and non-person, with an accuracy of 99.21%.



Figure 1. Classification network architecture.

3.3. Region Proposal Stage

The first method trained and tested on the HERIDAL dataset is edge boxes, one of the state-of-the-art region proposal algorithms that generates object bounding box proposals using edges [36]. This method is based on the observation that the number of contours that are fully connected in a bounding box is indicative of the box containing the object. We used this method for the region proposal stage in a specific task of person detection in aerial images, while the proposed regions were classified using CNN shown in Figure 1. This resulted in 214 true positive detections and 581 false positive detections. Accordingly, the recall measure is 63.5%, and the precision is 26.91%. It is not very surprising that the results obtained with this method are not promising since UAV images of wilderness have an overly-complex content with many edges, resulting in a huge number of proposed regions.

Next, for comparison, an approach from [37] was adopted. This approach uses two stage segmentation for the detection of artificial materials and objects in nonurban terrain. The mentioned method is used for the region proposal stage, and the proposed regions are further filtered using a pretrained convolutional neural network explained above for a binary classification task. In order to ensure a fixed size of the network inputs, the approach proposed in [38] is adopted, and regions are rescaled using the image warping algorithm. Using this method, one-hundred seventy-two persons, out of a total of 337 contained in the HERIDAL dataset test images, were successfully detected. This corresponds to achieving a precision of 52.76% and a recall of 51.03%. These results, as well as

those obtained by the method based on the edge boxes algorithm are not very promising because almost half of the missing detections still exist.

In addition to two stage methods, we also implemented a one stage deep learning based method for object detection in images with emphasis on execution speed—the Single Shot MultiBox Detector (SSD) [6]. SSD tries to detect multiple objects using a single pass through the neural network, making it easy to train and very fast. Unfortunately, due to a low resolution feature map, this method performed badly on person detection in aerial images and resulted in an enormous number of region proposals (more than 17,000). Using the non-maxima suppression in order to discard highly overlapping bounding boxes reduced the number of proposed regions by almost 60%, but still, too many proposed regions remained (7014), which gives an unbearably low precision of 4.33%.

The fourth evaluated method was based on Region Proposal Network (RPN), proposed in [9] for extracting potential regions containing objects, followed by the classification step where the trained classification network shown in Figure 1 is used to determine whether the proposed region contains an object or not. RPN takes an image as the input, and the output is the set of rectangular proposals of images that contain an object of interest. Since images in the HERIDAL dataset are too large, in order to reduce the computational cost, images were divided into blocks. For further processing, only those blocks that contain an object are taken. For the testing phase described in this paper, blocks with dimensions of 500×500 with an overlap of 100 px in the horizontal and 200 px in the vertical direction were used. The first step of the RPN algorithm is feature extraction, which is done using the convolutional neural network VGG16 [39] architecture without fully connected layers. It is publicly available pretrained CNN for image classification tasks with the “ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC2014)” competition. Since this network consists of a total of 5 max pooling layers, the output is a feature map with dimensions $62 \times 46 \times 512$. The second step is classification, which determines the probability of a proposal having the target object, and regression, which regresses the coordinates of the proposal. This step starts with anchor determination. A total of 9 proposals for every pixel, 3 scales (8, 16, 32), and 3 aspect ratios (1:1, 1:2, 2:1) were chosen. The number of anchors in the whole input image is 25,668 ($62 \times 46 \times 9$). For the regressor training, only anchors with the Intersection over Union (IoU) (with bounding boxes in the ground truth data) equal to or greater than 0.7 were used. The output of the regression step is the bounding boxes, which have to be classified using binary classification. For this purpose, the simple convolutional neural network explained in Section 3.2 was used. The use of RPN provided the best result, if only the number of true positive detections is considered, which is in this case 322. It gives a recall of 95.54%, which is remarkable. However, precision is 41.54% because of a large number (in this case 453) of false positive detections, which in real SAR operations means directing rescuers to the wrong location.

In order to address this problem and reduce the number of false positive detections, we decided to use an approach with multiscale features. A generic solution for building feature pyramids inside deep ConvNets was presented in 2017, called the Feature Pyramid Network (FPN) [40]. The FPN works well as a backbone in small object detection [41]. Unlike standard RPN, which uses single-scale features on top of the last layer of CNN, this is replaced with FPN. FPN is a feature extractor, and RPN detection is performed over extracted multi-scale feature maps. Feature extraction is carried out in two steps: a bottom-up and a top-down pathway. While a bottom-up pathway is achieved by forward propagation through convolutional layers, top-down pathway implies upsampling high resolution features by a factor of two using the nearest neighbor. Furthermore, lateral connections are used, which means merging of bottom-up and top-down feature maps using element-wise addition. Predictions are made independently at all levels. This approach significantly reduced the number of false positive detections (88), but also the number of true positive detections (292). Accordingly, the achieved recall measure is 86.64%, while the precision measure is 76.84%.

3.4. Multimodel Approach

Observing the results obtained by the methods mentioned above, it is evident that some methods achieve better performance in terms of precision measurements, while others achieve better recall measurement results. Obviously, the recall measure is critical in searching for lost persons because every false negative detection means a person is missed. On the other hand, a large number of false positive detections (consequently a low precision measure) means that too many suspicious locations would be investigated and, therefore, additional time and resources spent. In order to optimize both measures, our central premise was that a novel method based on a multimodel approach would increase precision while keeping recall at the same level. Therefore, a multimodel approach shown in Figure 2 is proposed. Because of the computational complexity, the input image (4000×3000) is divided into blocks of 500×500 px with an overlap of 100 px in the horizontal and 200 px in the vertical direction.

The input to the region proposal algorithm is an image block, while the output is a set of bounding boxes. A bounding box ($bbox_k$) is a 4-dimensional vector containing the upper left (x_i, y_i) and lower right (x_j, y_j) coordinates of the proposed regions, as is shown in Equation (1). In the region proposal stage, we simultaneously used two models, RPN and FPN, so the output from this stage is a set of bounding boxes proposed by the RPN method named R_{RPN} , as well as the set of bounding boxes proposed by the FPN method named R_{FPN} .

$$\begin{aligned} bbox_k &= [x_i, y_i, x_j, y_j]; x_i, y_i, x_j, y_j, i, j, k \in \mathbb{N} \\ R_{FPN} &= [bbox_1, \dots, bbox_n]; \\ R_{RPN} &= [bbox_{n+1}, \dots, bbox_{n+m}]; n, m \in \mathbb{N} \end{aligned} \quad (1)$$

Next, the IoU between R_{RPN} and R_{FPN} bounding boxes is calculated with the standard formula shown in Equation (2):

$$\begin{aligned} IoU(bbox_i, bbox_j) &= \frac{|bbox_i \cap bbox_j|}{|bbox_i| + |bbox_j| - |bbox_i \cap bbox_j|} \\ &; bbox_i \in R_{RPN}, bbox_j \in R_{FPN} \end{aligned} \quad (2)$$

All regions with an IoU greater than the empirically determined threshold of 0.5 are excluded from further processing because they overlap with those regions already checked in the FPN method. Other regions are added in set $R_{RPN'}$ as shown in Equation (3).

$$R_{RPN'} = \begin{cases} bbox_i & \text{if } (IoU(bbox_i, bbox_j) < 0.5) \\ \emptyset & \text{else} \end{cases}; bbox_i \in R_{RPN}, bbox_j \in R_{FPN} \quad (3)$$

Finally, all regions from R_{FPN} and those from $R_{RPN'}$ are combined in a unique R_{TOT} set (Equation (4)) and forwarded to a binary classification network.

$$R_{TOT} = R_{RPN'} \cup R_{FPN} \quad (4)$$

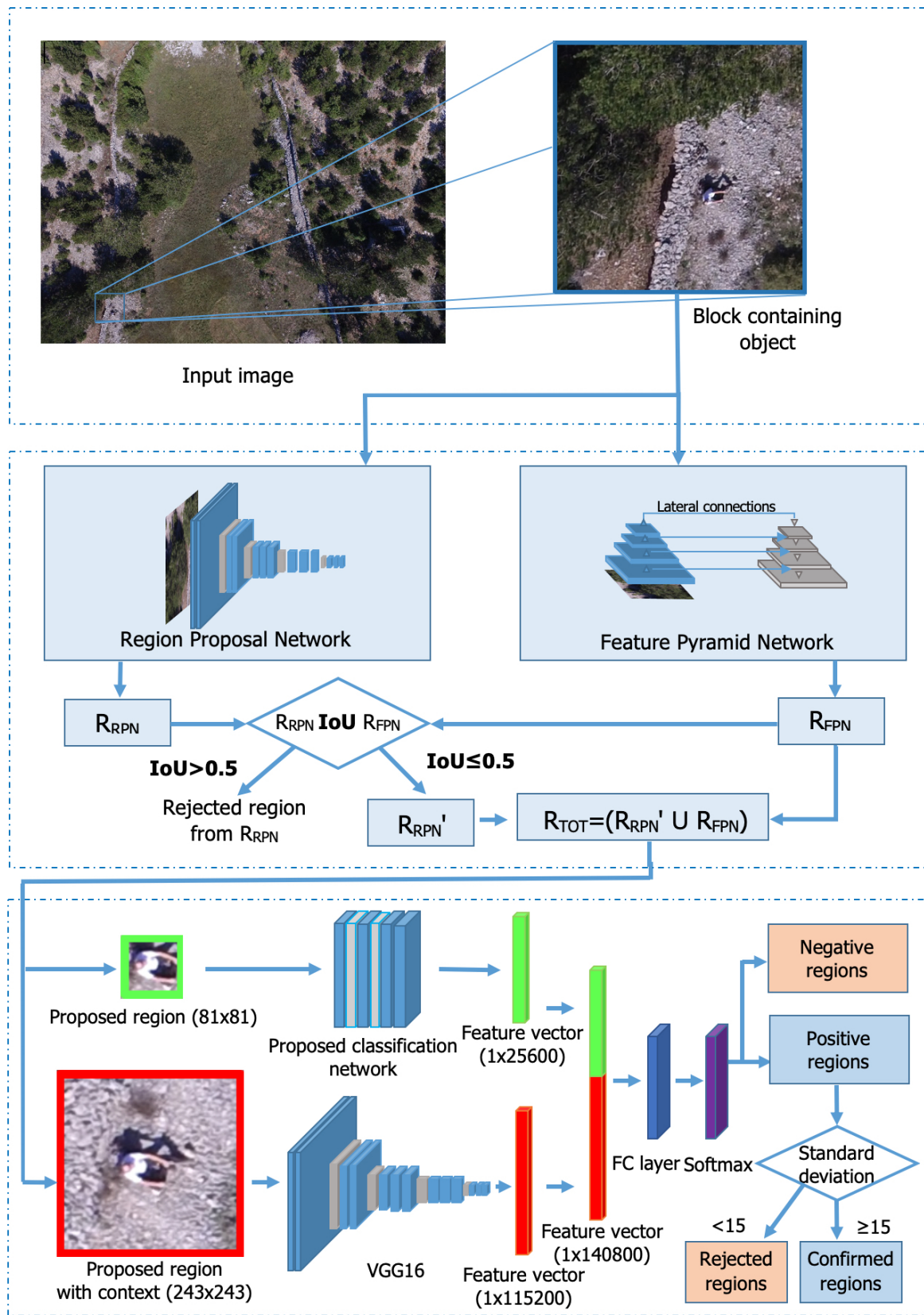


Figure 2. Multimodel approach for person detection in aerial images.

The classification stage with CNN presented in Figure 1 (Section 3.2) was used for output regions contained in the R_{TOT} set. Classification includes a feature map of the isolated object while ignoring the

background features. However, context is a rich source of information that makes it easier to identify objects in different ways. As already mentioned, semantic context cannot be used in this specific problem of person detection in aerial images, but it is assumed that the use of contextual information in the form of pixel based context [42] in the classification stage could improve the classification of regions obtained with the multimodel region proposal approach. Thus, the fusion of the contextual information of the proposed region with all regions surrounding it was performed. The previously mentioned classification network without the last, fully connected layer was used to classify the proposed region size of 81×81 pixels. Simultaneously, a contextual region is generated by taking the central coordinates of the proposed region. Around this central pixel, a region of size 243×243 is formed. This region size allows additional information on object surroundings when compared to the 81×81 pixel region. Furthermore, on the contextual region, the VGG16 network architecture is used for the feature vector extraction. This network is pretrained on ILSVRC2014 for object classification and fine-tuned using positive and negative regions from the HERIDAL dataset. Since images in the HERIDAL dataset (as well as in the real SAR operations) contain a bunch of negative regions while positive regions are limited, for the training task, all positive regions from the training images are taken, while negative regions are generated randomly (approximately the same number as positive regions). After obtaining two feature vectors, concatenating is performed in order to get a joint feature vector for the region and the contextual region, which provides the augmented information of the object. The concatenated feature vector is fed into a fully connected layer and softmax for classification. This approach further improved the results, as is shown in Table 1.

Table 1. Detection results obtained with several methods. The first column contains the names of methods. The second column shows a number of True Positive detections (TP), followed by False Positive detections (FP) and False Negative detections (FN). The last two columns show the standard precision and recall measures calculated as follows: $Precision = \frac{TP}{TP + FP}$ and $Recall = \frac{TP}{TP + FN}$. RPN, Region Proposal Network; FPN, Feature Pyramid Network; RFC, RPN + FPN+ Classification; RFCC, RFC + Context; RFCCD, RFCC + Deviation.

	TP	FP	FN	Precision	Recall
Marušić et al. [25]	301	146	40	67.30%	88.30%
Božić-Štulić et al. [26]	303	568	38	34.80%	88.90%
Edge Boxes + classification	214	581	123	26.91%	63.50%
Mean Shift + classification	172	154	165	52.76%	51.03%
SSD	318	7014	19	4.33%	94.36%
RPN + classification	322	453	15	41.54%	95.54%
FPN + classification	292	88	45	76.84%	86.64%
RFC	322	259	15	55.42%	95.54%
RFCC	320	163	17	66.25%	94.95%
RFCCD	319	144	18	68.89%	94.65%

Additionally, by further visual analysis of positively classified regions, we noticed that some of the false positive detections had small variations between neighboring pixels because it was a part of an image with grass, treetop, shadow, and similar regions. It may be concluded that those regions have a low standard deviation of the pixel values. In order to eliminate those regions and consequently slightly improve precision performance, standard deviation calculation for each detected region was applied. Subsequently, an experiment was performed where regions with a standard deviation below a certain threshold were eliminated. Various thresholds were used, and it was determined that the optimal threshold value was 15. Therefore, in the last step of the proposed approach, the elimination of regions with the standard deviation below this threshold from the set of positive detections was performed. With a lower threshold, fewer false positive detections would be eliminated, while a higher threshold would cause the elimination of the additional true positive detections.

4. Results and Discussion

In order to properly evaluate the proposed methods, it is important to emphasize that in real SAR operations, every false negative detection means that the lost person is actually in the photographed area, but the system missed him/her. Indirectly, this implies that the rescue team cannot provide assistance to a lost person in a timely manner. Based on this fact, it is evident that the false negative detections need to be reduced as much as possible. Consequently, the number of true positive detections would be increased, as well as the recall measure. From another point of view, false positive detections also should be reduced as much as possible because in real SAR operations, every false positive detection may lead the rescue team to the wrong location. This decrease would also affect the increase of the precision measure. Therefore, it is crucial to have the best balance between these two measures. Table 1 shows the results with the current state-of-the-art methods from [25,26] for the comparison, as well as those obtained with all proposed methods in this paper on 101 test images from the HERIDAL dataset, which in total contain 337 objects of interest (persons). The number of True Positive detections (TP) shown in the first column of the table refers to those regions that are proposed and classified as a person correctly. The second column contains the number of False Positive detections (FP), which means that those regions do not actually contain a person, but they are incorrectly proposed and classified as a person. False Negative detections (FN) in the third column represent those regions that actually contain a person, but the model failed to detect them. Depending on the numbers listed, corresponding precision and recall measures were also calculated and presented in the last two columns.

Our first proposed multimodel approach named RFC (RPN + FPN + Classification) achieved a precision of 55.42% and a recall of 95.54% with 322 true positive detections and 259 false positive detections. It is obvious that the proposed model improved the results because the number of true positive detections given by the RPN method for the region proposal stage is maintained, while using the FPN method decreased the number of false positive detections. In real SAR operations, that means less false alarms and less wasted time.

It may be concluded that using contextual information in the classification stage also improves the results; therefore, a new multimodel method named RFCC (RFC + Context) is introduced. Although using contextual information slightly decreased the number of true positive detections down to 320, it also significantly decreased the number of false positive detections. The achieved recall and precision measure were respectively 94.95% and 66.25%.

Finally, in the RFCCD (RFCC + Deviation) method, the standard deviation was calculated for each positive detection in order to eliminate those regions whose standard deviation is below the threshold. After this elimination, there were 319 true positive and 144 false positive detections obtained, which correspond to 68.89% precision and 94.65% recall.

Although the results obtained with the RPN method are the best in terms of the recall measure, at the same time, the precision is inadequate because of the large number of false positive detections. The use of the FPN method achieved the best results if only the precision measure was observed, while the recall measure was significantly reduced. The proposed multimodel approaches keep the recall measure high enough while the precision is increased. This leads to the conclusion that the RFCCD method is most suitable for use because it provides the best balance between the precision and recall measures.

Regarding the quantitative comparison with other methods, it is more appropriate to compare the obtained results with those achieved with current state-of-the-art methods in this particular problem on the HERIDAL dataset presented in [25] (recall: 88.3%; precision: 67.3%) and [26] (recall: 88.9%; precision 34.8%). In comparison with those mentioned, our proposed multimodel approach achieved superior results.

Figure 3 shows a visual example of improving results using different multimodel approaches on the HERIDAL dataset images. The left side of the figure contains a complete image with labeled regions using different methods, while the right side shows enlarged detected regions. Green bounding boxes

represent true positive detections, while false positive detections are represented with red bounding boxes. The first part of the image (a) shows the results obtained with the RPN detector. It is noticeable that all persons in the image are correctly detected, but there are also multiple false positive detections. This is caused by the fact that the feature extractor further reduces the size of already small objects so much that they become unrecognizable. In the second part of the image (b), the results obtained with the FPN method are shown. Using this method, the number of false positive detections is reduced, but it is also evident that some of the true positive detections are missing (two persons in the image are not detected). Accordingly, the precision measure increases, while the recall measure decreases. Using our proposed multimodel approach RFC, the balance between the recall and precision measure is more beneficial, which can be seen in the third part of the image (c). It is obvious that this method retained the advantages of both models used, resulting in all persons detected, and also, the number of false alarms decreased. Furthermore, the results are improved using the RFCC model, which includes contextual information in the classification stage. This improvement can be seen in the fourth part of the image (d), where the number of false positive detections is effectively minimized while the number of true positive detections is retained.

Furthermore, a comment on the applicability and usefulness of the proposed solution compared to the inspection of images by a human observer should be made. Marušić et al. [25] performed a visual inspection experiment on 28 images from the HERIDAL dataset. A group of eight students visually inspected images in order to detect a person, and they achieved a recall of 92% and a precision of 94%. A valuable advantage of the proposed model over visual inspection is also in execution time. Within the experiment, the average time for visual inspection of one image is 43.68 s, while our proposed model performed in less than 15 s per image. The algorithm was run on a workstation equipped with an Intel Xeon E5-2640v4 of 3.40 GHz, 4 × 16 GB DDR4 memory, and multi-GPU 4 × NVIDIA GeForce GTX 1080Ti Turbo with 11 GB memory. Taking into consideration that visual inspection of high resolution images is a really demanding task, it can be concluded that the average time and detection results would be worse if the experiment were performed on a larger number of images because of the performers' tiredness. Supporting SAR operations with drones always results in a huge number of images that need to be processed; therefore, we can state that the visual inspection method is not reliable.

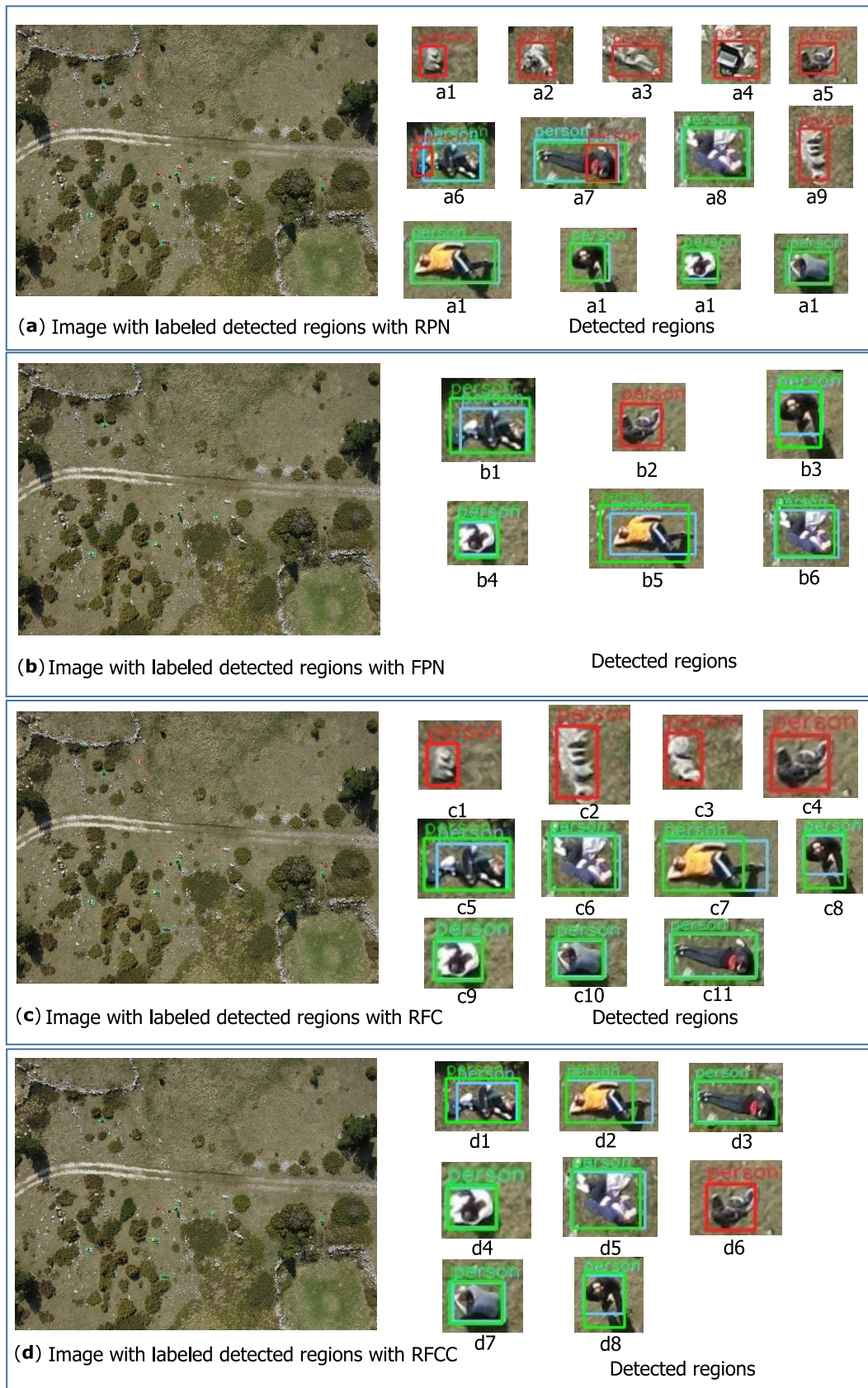


Figure 3. Example of results obtained with different approaches.

5. Conclusions

Search and rescue operations could greatly benefit from using drones for search missions due to their ability to cover a large and wide geographical area faster and provide high quality images. Nevertheless, in order to utilize their full potential, it is important to properly use information provided by drone sensors. For this purpose, an effective system for person detection based on processing of aerial images could significantly improve the outcome of SAR operations.

In this paper, we proposed a novel method for person detection in aerial images based on multimodel deep learning approach in terms of the region proposal stage, as well as in the classification stage. The proposed method achieved 94.66% recall and 68.9% precision, which makes it a state-of-the-art method that can be used as an auxiliary method in real SAR operations in the wilderness.

It is important to mention that the proposed method can be generalized for other remote sensing applications (with some adjustments such as changing the training dataset). However, at the moment, our main goal is to build a reliable system for supporting SAR operations. Therefore, in future work, we intend to improve the proposed model with the capability to detect other small objects in images acquired by drones, e.g., jacket, bag, car, etc. This could also be useful as they may represent helpful traces for searchers and rescuers. Increasing the processing speed in order to achieve a real-time detection is obviously one of our further goals, as well as the use of successive images acquired by a UAV in order to maximize the detection reliability of overlapping suspicious regions.

Author Contributions: Conceptualization, M.K.V. and V.P.; methodology V.P.; software, M.K.V.; validation, M.K.V.; formal analysis, V.P.; investigation, M.K.V. and V.P.; resources, M.K.V. and V.P.; data curation, M.K.V. and V.P.; writing, original draft preparation, M.K.V.; writing, review and editing, V.P.; visualization, M.K.V. and V.P.; supervision, V.P.; project administration, V.P.; funding acquisition, V.P. All authors read and agreed to the published version of the manuscript.

Funding: This research was supported by the project “Prototype of an Intelligent System for Search and Rescue”, Grant Number KK.01.2.1.01.0075, funded by the European Regional Development Fund.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Waharte, S.; Trigoni, N. Supporting Search and Rescue Operations with UAVs. In Proceedings of the International Conference on Emerging Security Technologies, Canterbury, UK, 6–7 September 2010; pp. 142–147.
2. Goodrich, M.A.; Morse, B.S.; Gerhardt, D.; Cooper, J.L.; Quigley, M.; Adams, J.A.; Humphrey, C. Supporting wilderness search and rescue using a camera-equipped mini UAV. *J. Field Robot.* **2008**, *25*, 89–110. [[CrossRef](#)]
3. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*; Curran Associates, Inc.: Red Hook, NY, USA, 2012; pp. 1097–1105.
4. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.W.; Chen, J.; Liu, X.; Pietikainen, M. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vis.* **2018**, *128*, 261–318.
5. Bejiga, M.; Zeggada, A.; Melgani, F. Convolutional neural networks for near real-time object detection from UAV imagery in avalanche search and rescue operations. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 693–696.
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9905, pp. 21–37.
7. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
8. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Las Condes, Chile, 11–18 December 2015; pp. 1440–1448.

9. Shaoqing, R.; Kaiming, H.; Girshick, R.; Jian, S. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)]
10. Valenti, C.F.; Nguyen, N.D.; Do, T.; Ngo, T.D.; Le, D.D. An Evaluation of Deep Learning Methods for Small Object Detection. *J. Electr. Comput. Eng.* **2020**, *2020*. [[CrossRef](#)]
11. Zhang, S.; Wu, R.; Xu, K.; Wang, J.; Sun, W. R-CNN-Based Ship Detection from High Resolution Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 631. [[CrossRef](#)]
12. Zhang, L.; Zhang, Y.; Zhang, Z.; Shen, J.; Wang, H. Real-Time Water Surface Object Detection Based on Improved Faster R-CNN. *Sensors* **2019**, *19*, 3523. [[CrossRef](#)] [[PubMed](#)]
13. Lechgar, H.; Bekkar, H.; Rhinane, H. Detection of cities vehicle fleet using YOLO V2 and aerial images. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *XLII-4/W12*, 121–126. [[CrossRef](#)]
14. Zhang, H.; Wu, J.; Liu, Y.; Yu, J. VaryBlock: A Novel Approach for Object Detection in Remote Sensed Images. *Sensors* **2019**, *19*, 5284. [[CrossRef](#)]
15. Liu, M.; Wang, X.; Zhou, A.; Fu, X.; Ma, Y.; Piao, C. UAV-YOLO: Small Object Detection on Unmanned Aerial Vehicle Perspective. *Sensors* **2020**, *20*, 2238. [[CrossRef](#)]
16. Yun, K.; Nguyen, L.; Nguyen, T.; Kim, D.; Eldin, S.; Huyen, A.; Lu, T. Chow, Small target detection for search and rescue operations using distributed deep learning and synthetic data generation. In Proceedings of the Pattern Recognition and Tracking XXX, Baltimore, MD, USA, 14–18 April 2019; pp. 38–43.
17. Liang, X.; Zhang, J.; Zhuo, L.; Li, Y.; Tian, Q. Small Object Detection in Unmanned Aerial Vehicle Images Using Feature Fusion and Scaling-Based Single Shot Detector With Spatial Context Analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1758–1770. [[CrossRef](#)]
18. Zhang, Z.; Guo, W.; Zhu, S.; Yu, W. Toward Arbitrary-Oriented Ship Detection With Rotated Region Proposal and Discrimination Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1745–1749. [[CrossRef](#)]
19. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [[CrossRef](#)] [[PubMed](#)]
20. Xiao, Z.; Qian, L.; Shao, W.; Tan, X.; Wang, K. Axis Learning for Orientated Objects Detection in Aerial Images. *Remote Sens.* **2020**, *12*, 908. [[CrossRef](#)]
21. Koester, R.J. *A Search and Rescue Guide on where to Look for Land, Air, and Water*; dbS Productions: Charlottesville, VA, USA, 2008.
22. Music, J.; Orovic, I.; Marasovic, T.; Papić, V.; Stankovic, S. Gradient Compressive Sensing for Image Data Reduction in UAV Based Search and Rescue in the Wild. *Math. Probl. Eng.* **2016**, *2016*. [[CrossRef](#)]
23. Burke, C.; McWhirter, P.R.; Veitch-Michaelis, J.; McAree, O.; Pointon, H.A.; Wich, S.; Longmore, S. Requirements and Limitations of Thermal Drones for Effective Search and Rescue in Marine and Coastal Areas. *Drones* **2019**, *3*, 78. [[CrossRef](#)]
24. Leira, F.; Johansen, T.; Fossen, T. Automatic detection, classification and tracking of objects in the ocean surface from UAVs using a thermal camera. In Proceedings of the 2015 IEEE Aerospace Conference, Big Sky, MT, USA, 7–14 March 2015. [[CrossRef](#)]
25. Marušić, Z.; Božić-Štulić, D.; Gotovac, S.; Marušić, T. Region Proposal Approach for Human Detection on Aerial Imagery. In Proceedings of the 2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 26–29 June 2018; pp. 1–6.
26. Božić-Štulić, D.; Marušić, Z.; Gotovac, S. Deep Learning Approach in Aerial Imagery for Supporting Land Search and Rescue Missions. *Int. J. Comput. Vis.* **2019**, *127*, 1256–1278.
27. Divvala, S.; Hoiem, D.; Hays, J.; Efros, A.; Hebert, M. An empirical study of context in object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1271–1278.
28. Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.G.; Lee, S.W.; Fidler, S.; Urtasun, R.; Yuille, A. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 891–898.
29. Torralba, A.; Sinha, P. Contextual Priming for Object Detection. *Int. J. Comput. Vis.* **2003**, *53*, 169–191. [[CrossRef](#)]
30. Sun, J.; Jacobs, D. Seeing What Is Not There: Learning Context to Determine Where Objects Are Missing. In Proceedings of the 2017 Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5716–5724.

31. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2874–2883.
32. Han, S.; Yoo, J.; Kwon, S. Real-Time Vehicle-Detection Method in Bird-View Unmanned-Aerial-Vehicle Imagery. *Sensors* **2019**, *19*, 3958. [[CrossRef](#)]
33. Ma, D.; Wu, X.; Yang, H. Efficient Small Object Detection with an Improved Region Proposal Networks. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *533*, 012062. [[CrossRef](#)]
34. Xia, G.S.; Bai, X.; Zhang, L.; Belongie, S.; Luo, J.; Datcu, M.; Pelilo, M. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
35. Marasović, T.; Papić, V. Person classification from aerial imagery using local convolutional neural network features. *Int. J. Remote Sens.* **2019**, *40*, 1–19. [[CrossRef](#)]
36. Zitnick, C.; Dollar, P. Edge Boxes: Locating Object Proposals from Edges. In Proceedings of the European Conference on Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Volume 8693, pp. 391–405.
37. Turić, H.; Dujmić, H.; Papić, V. Two stage Segmentation of Aerial Images for Search and Rescue. *Inf. Technol. Control.* **2010**, *39*, 138–145.
38. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [[CrossRef](#)]
39. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015.
40. Lin, T.Y.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
41. Liang, Z.; Shao, J.; Zhang, D.; Gao, L. Small Object Detection Using Deep Feature Pyramid Networks. In Proceedings of the Advances in Multimedia Information Processing—PCM 2018, Hefei, China, 21–22 September 2018; pp. 554–564.
42. Fang, P.; Shi, Y. Small Object Detection Using Context Information Fusion in Faster R-CNN. In Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications, Chengdu, China, 7–10 December 2018; pp. 1537–1540.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).