IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 14, 2022, accepted February 28, 2022, date of publication March 4, 2022, date of current version March 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3156903

Detecting Humans in Search and Rescue Operations Based on Ensemble Learning

NAYEE MUDDIN KHAN DOUSAI, (Member, IEEE), AND SVEN LONČARIĆ, (Senior Member, IEEE)
Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia

Corresponding author: Nayee Muddin Khan Dousai (nayee.dousai@fer.hr)

This work was supported by the European Union's Horizon 2020 Research and Innovation Program under the Marie Sklodowska-Curie Grant under Agreement 764951.

ABSTRACT Detection of humans accurately in aerial images is critical for various applications such as surveillance, detecting and tracking athletes on sports fields, and search and rescue operations (SAR). The goal of SAR is to assist, detect, and rescue people who have had accidents in the mountains or other hazardous environments. By using drones in SAR applications, it is desirable to minimize the cost and time spent on SAR operations. In this paper, we present a convolutional neural network-based model for the detection of humans in aerial images of mountain landscapes acquired by unmanned aerial vehicles (UAVs) used in search and rescue operations. Detection of humans in aerial images remains a complex task due to various challenges such as pose and scale variations of humans, low visibility, camouflaged environment, adverse weather conditions, motion blur, and high-resolution aerial images. Due to imaging from high altitudes, in most high-resolution aerial images captured by UAVs, only 0.1 to 0.2 percentage of the image represents humans. To solve the problem of low coverage of the object of interest in high-resolution aerial images, we propose to implement a deep learning-based object detection model. In this paper, we propose a novel method for the detection of humans in aerial images based on the EfficientDET architecture and ensemble learning. The method has been validated on the HERIDAL image dataset. By implementing the proposed methodologies, we achieved an mAP of 95.11%. To the best of our knowledge, this is the highest accuracy result for human detection on the HERIDAL dataset.

INDEX TERMS Deep-learning, detection of humans, EfficientDET, ensemble learning, HERIDAL dataset, image analysis, search and rescue (SAR) operations.

I. INTRODUCTION

Object detection is one of the most researched areas in computer vision. It is the process of determining where exactly the object is in the scene or image and what object has been detected. Object detection refers to finding different types of objects in the scene such as peoples, cars, animals or other existing objects present in the scene [1]–[3]. While normal ground-to-ground imagery has yielded promising results in object detection, detecting objects in aerial imagery is still considered a difficult task [4], [5]. One such important task is to rescue people in search and rescue (SAR) operations from aerial images without loss of life. SAR operations are conducted in wide-open spaces, such as mountains, lowlands, cities, disaster scenarios [6] and marine rescue. In general, search and rescue operations need to be conducted as quickly as possible to identify missing persons. It can be highly expensive and requires distinct types of activities such as sending people in large groups, sniffer dogs and various types of ground and air vehicles such as cars and helicopters.

Object detection in aerial images depends on several factors such as low visibility due to varying altitudes, the object-of-interest, variations in pose and scale, camouflaged environment with rocks and trees, and high-resolution aerial images [4], [7], [8] as shown in Fig. 1. It is expensive and time-consuming to capture aerial images based on these parameters. For example, the UK National Police Air Service (NPAS) logged over 17, 000 of mission hours in 2016/17, with each hour of flight operations costing an estimation of £3000 [9]. To avoid the high costs and

The associate editor coordinating the review of this manuscript and approving it for publication was Szidónia Lefkovits.

FIGURE 1. Few drawbacks of aerial images showing snow and shadow (top-left), scale and pose variations (top-middle), the camouflaged environment with rocks (top-right) and trees (bottom-left), motion blur (bottom-middle), and low illuminance (bottom-right) in the HERIDAL dataset [39], [40].

time commitments associated with traditional SAR methods, we will employ consumer drones in SAR operations, which are readily accessible in the market and significantly less expensive than conventional SAR methods. We can easily identify humans by using drones and various approaches such as machine learning algorithms [10] or thermal imaging cameras [11]. Machine learning is a technology used for a wide range of applications on the road and in the air, such as autonomous driving, mapping a specific region with drones, or finding humans in SAR operations. In machine learning-based human detection, we require a large amount of aerial data to train and detect the humans.

The selection of aerial datasets for training is very crucial in machine learning-based human detection. Detecting humans in aerial images can be carried out in two different ways as offline and onboard detection methods. In offline methods, we need to collect or find the available open-source well-annotated aerial image datasets, train them using a machine learning algorithm and test them with other aerial images. For on-board detection methods, human detection must be carried out live onboard and the data is transmitted to the ground station. The live data stream can be a significant problem due to various drawbacks such as lack of cellular networks, limited bandwidth, the distance between the drone and the ground station, high resolution and numerous images [11]. Given these limitations for the onboard detection system, we preferred to use an offline human detection method for SAR operations described in the proposed section. However, testing the detection models in offline mode is always a safe and prerequisite step before deploying the models online. The other method for human detection in aerial images is to use thermal cameras attached to UAVs and obtain their live feedback. As thermal infrared cameras (TIR) have fewer pixels compared to RGB images, it is easy to get the transmission. The use of thermal cameras is not feasible always, because detecting people with thermal cameras is reliable with weather conditions. In winter or cold regions, the normal temperature of the human body is higher than the environment, so humans appear bright and clear using thermal imaging. While in summer and tropical areas the body temperature is much lower than the environment, so it will be difficult to detect humans in such environments. Using thermal infrared cameras (TIR) [12] for object detection can also have the drawback of carrying heavy cameras. The other reason not to consider thermal imaging in the paper is, the HERIDAL dataset [39]considered for training the model in the below section doesn't have any thermal imaging aerial pictures.

Considering all the above mentioned drawbacks of thermal camera detection and onboard human detection methods, we can conclude that offline human detection is the most reliable method for detecting humans in SAR operations. Using offline human detection, one of the crucial drawbacks to consider is how we can train deep learning architectures for huge image resolutions by minimizing the training costs with better accuracy. This paper considers the drawback of high-resolution aerial images with a small object of interest for search and rescue (SAR) operations. It is challenging to detect and locate humans when the object of interest occupies a minuscule percentage ($0.1 - 0.2\%$) of a huge aerial image.

The paper is organised into five sections and an appendix. The introduction section explains the object detection drawbacks in aerial images and the problems in SAR operations. The next related work section is divided into three subsections. The subsections are explained as handcrafted object detectors, deep learning object detectors, and deep learning object detectors for SAR operations. The third section, methodology explains into four sub-sections as the selection of aerial dataset, dataset preparation, proposed method, and the metric calculation. The section explains the implementation of the ensemble learning model based on Bi-FPN and FC-FPN. Section four, experiments and results explain all the experiments in three sub-sections as experiments with BIFPN, FC-FPN and, final results and discussion. The last section, the conclusion explains the main contribution of the paper and the future work. At the end of the paper, an appendix section is also described with some extra clear figures of EfficientNET backbone with Bi-FPN and FC-FPN.

II. RELATED WORK

A. HAND CRAFTED OBJECT DETECTORS

Object detection has been a significant research area in computer vision over the past two decades. As part of this extensive research, several papers have been published on the detection of objects with exceptional results [13]–[15]. However, most of these papers were focused on detecting objects captured from ground images. When it comes to aerial images, it is still a complicated task because there are many factors and drawbacks to consider, such as small object detection and high resolution of aerial images. Detection of small objects in huge aerial images that represent less than 1% of the size is much more challenging than detecting objects on the ground. Before the evolution of convolutional neural networks, one of the most influential papers proposed for real-time recognition was from Viola-Jones (VJ) in 2001 [16], [17], which has quite impressive results for face recognition. Some researchers have also presented the use of thermal imaging techniques by using thermal infrared (TIR) cameras to detect humans from the aerial images [18]. One such paper is from Burke *et al.* [12], who explains the limitations and requirements of thermal object detection for effective search and rescue in marine and coastal environments. Another paper by Doherty *et al.* [19], explains the detection of human by using a simple hardware-based onboard model using thermal and colour imagery.

B. DEEP LEARNING OBJECT DETECTORS

The efficiency of object detection has evolved significantly with the emergence of convolutional neural networks (CNNs) [20], [21]. AlexNet [22] is one such model which has been shown to outperform most handcrafted models such as the VJ detector [16], [17], the Scale Invariant Feature Transform (SIFT) [23], and the Histogram of Oriented Gradients (HOG) [24], [25]. Since then, CNNs have made a huge leap in object detection and many other computer vision applications such as feature extraction, autonomous driving, and others. Based on the region of interest (ROI) [13], most deep learning architectures are categorized into two types: one-stage and two-stage architectures [13], [14]. Single-stage detectors are directly approached models without any intermediate object proposals called end-to-end object detection models whereas, two-stage object detection models have a two-way approach with a regional proposal stage followed by object detection and bounding box regression. For the regional proposal stage, there are few methods considered such as Mean Shift [26], Region Proposal Network (RPN) [27], and Feature Pyramid Network (FPN) [28]. The main difference between these stages is the accuracy and the use of the application, two-stage detectors are considered to be more accurate as compared to the one-stage detectors while single level detectors are better in real-time applications.

There are many papers proposed on both one-stage and two-stage object detectors [13], [14], but most of them are based on Faster RCNN [27], FPN [28], EfficientDET [29], [30], YOLO [31]–[33], and SSD [34] through various modifications and improvements. Most CNN's require a fixed small input size for training and testing, which limits network depth, width, or image resolution. This is one of the challenges in aerial datasets, as the images captured by drones are usually high-resolution images. Considering the drawback of scaling and aspect ratio few researchers from Google has proposed a one-stage object detector EfficientDET [29], [30], which has proven to be more efficient and accurate than two-stage object detectors. They proposed a new scaling factor that can uniformly scale all dimensions of depth, width and resolution. By proposing a new compound coefficient and using Bi-directional Feature Pyramid Network (BiFPN) [29], EfficientNet-B7 achieves a state-of-the-art 84.4% accuracy on ImageNet [35] and an average precision of 52.2% on COCO test-dev [36].

C. DEEP LEARNING OBJECT DETECTORS FOR SEARCH AND RESCUE (SAR) OPERATIONS

Many more publications have been recommended for object detection [37], [38]. However, according to our application of search and rescue operations on an aerial dataset, there are relatively few of them. Among those handfuls, a study from the University of Split has proposed a model [39] based on the HERIDAL aerial dataset [40], [41]. Their work is based on

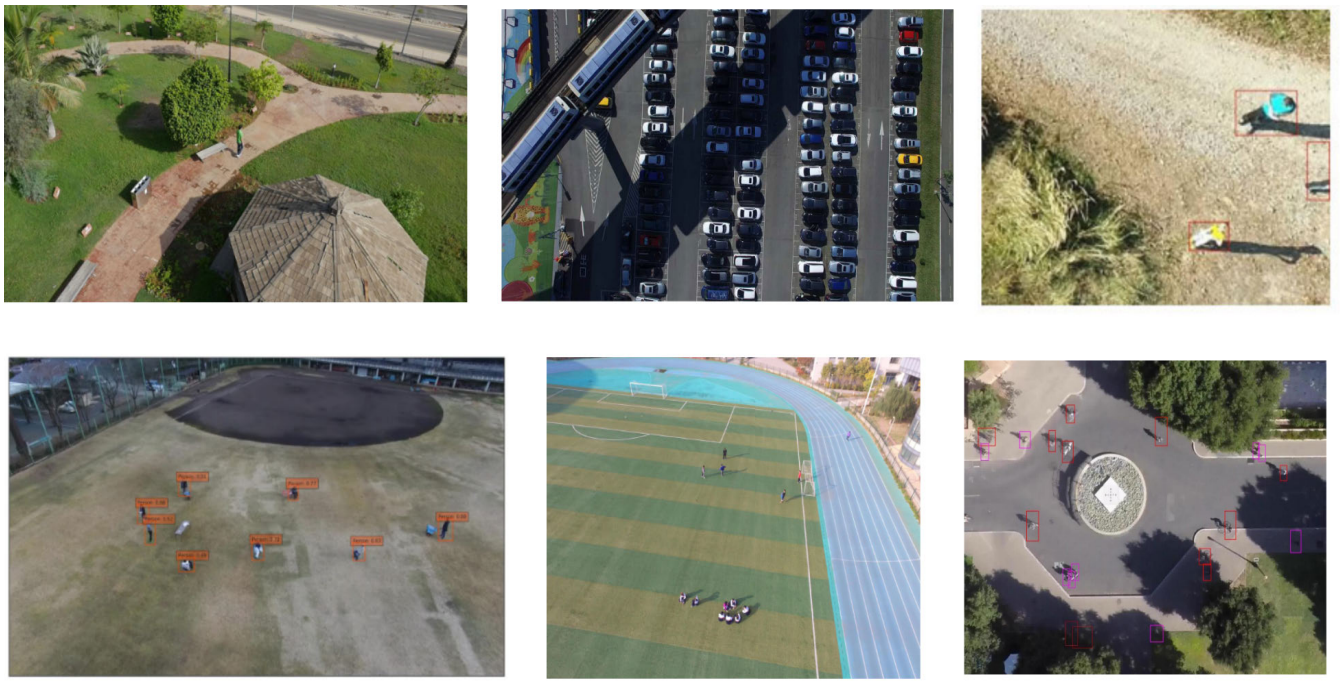

FIGURE 2. Few example images from the existing aerial datasets, Top-left: UAV123 [48], Top-middle:CARPK [52],Top-right: SARD [53], Bottom-left: Okutama-action [49], Bottom-middle: VisDrone [46], Bottom-right: Campus [50], [51].

TABLE 1. Comparing state-of-the-art models in SAR operations with our proposed model in the paper.

SAR state-of-the-art models	Proposed architectures
Kundid Vasić et al. [42]	Multimodel deep learning approaches based on RPN + FPN
Božić-Štulić et al. [39]	Two step attention algorithm based on Faster RCNN
Nayee Muddin Khan et al. [44]	EfficientDET architecture with pre-processing step
Our proposed model	Ensemble learning based on EfficientDET architecture with Bi-FPN + FC-FPN

the implementation of a two-step frame utilizing the Faster-RCNN object detector [27]. The salient features are extracted from the HERIDAL dataset using the Region Proposal module in the first step, and a CNN module has been used to categorize people and non-people in the HERIDAL dataset for the second stage. In the end, non-maxima suppression was performed to reduce false-positive detections by using clustering proposals and achieved an accuracy of 88.9%. Another paper based on the same dataset was proposed by Vasić*et al.* [42]. The research in the paper explains the use of multi-modal deep learning architectures for both regional proposals and classification stages. The following methods are used in the region proposal phase: Edge Boxes [43], Mean Shift [26], Region Proposal Network (RPN) [27] and Feature Pyramid Network [28]. During the classification phase, each proposed region is binary classified using a small convolutional neural network designed for classification problems by utilizing the patches from the HERIDAL dataset for training and testing. By this approach, they achieved an accuracy of 68.89% and a recognition of 94.65%. Since our research was also conducted on the same dataset, the results from these two papers are considered benchmark results, which we compare in the Experiments and Results section. We would also like to mention our previous work published at the beginning of 2021 [44] based on EfficientDET [29] and the HERIDAL dataset [39]. The paper describes how to regenerate the HERIDAL dataset to reduce computation time, as well as how to train deep learning architectures for aerial images based on the object of interest while ignoring the rest of the aerial image that does not contain an object of interest. All the uniqueness of state-of-the art models with the proposed model from the paper is shown in table 1. By considering the EfficientDET architecture at an image resolution of 512 × 512 and optimised hyperparameters, we achieved an mAP of 93.29% with RMSprop optimizer. The research presented in the proposed section below is part of the continuity of the paper.

III. METHODOLOGY

A. SELECTION OF AERIAL DATASET

There are several open-sourced datasets available for ground-to-ground images [35], [36], [45], but to find a well-annotated aerial dataset for the suitable application is difficult. However, a few aerial image datasets have recently been captured in diverse scenarios, e.g., for crowded places, playgrounds, mountains, urban and non-urban areas as shown in Fig.2. We would like to mention few of those aerial

images as: Vision Meets Drones (VisDrone) [46], [47], The UAV123 [48], Okutama-action by Barekatain [49], Campus [50], [51], and car parking dataset [52]. Most of these aforementioned aerial datasets lack in various parameters like not designed for SAR operations or not well annotated for human object class. Finding an authentic and reliable aerial dataset for search and rescue operations is considered crucial as the testing results are dependent on the selection of the training dataset. After the PASCAL VOC dataset [45] has been publicly open-sourced, there have been many challenging and well labelled open-sourced object detection datasets available from different researchers and organisations.

However, for our research, we are specifically aiming for aerial images which are suitable for search and rescue operations. One of such important and well labelled aerial datasets we are using for our research is the HERIDAL dataset [39], [40]. Few researchers from the University of Split, flew a few hundred hours on drone flights across the Mediterranean and Sub-Mediterranean landscapes to collect the data specifically suited for SAR operations. HERIDAL dataset contains over 1500 well-labelled training images and 500 testing images where 29, 050 are positive image patches with humans and 39, 700 are negative image patches without any objects labelled. Image patches are the set of cropped images with positive and negative patches recognising humans from the original image resolution. All the captured images for the dataset has 4000×3000 image resolution as shown in Fig. 3. The other main advantage for choosing the HERIDAL dataset for our research over VisDrone [46], [47] and other existing aerial datasets are all the images in the HERIDAL dataset are labelled in particularly to detect humans in search and rescue operations. Considering all these favourable factors, we will build our deep learning models using the HERIDAL dataset in the following section.

B. DATASET PREPARATION

Due to the drawback of high-resolution aerial images, most published research on aerial detection using deep-learning methods lacks accuracy and computation time. Even for our research and selection of the aerial dataset section, we have found that the HERIDAL database has an image pixel size of 4000×3000. Deep-learning models will require a large amount of RAM for the Graphical Processing Unit (GPU) and enormous amount of time to train 1500 well-labelled training database. To overcome this drawback, we have proposed a preprocessing step [44], where we have proposed two different scenarios for the generation of new HERIDAL database into various image sizes such as 512, 640, and 1024 as shown in Fig. 3. In the first scenario, we will consider and include all existing humans in the defined picture window varying from 512 to 1024 image resolution. In the second scenario, when there are single or more humans for the defined image window we will generate them as two or more images. One of the main reason to include this pre-processing step is to decrease the HERIDAL dataset image resolution by neglecting the image area where there is no presence of humans.

In each of the previous scenarios, we will exclude objects with pixel values less than 10×10 from the original image size. The main reason to implement this idea is due to the presence of humans above 10-pixel value from the patches of original HERIDAL dataset. By proposing this step we can save a lot of computational time in regeneration of the HERDIAL dataset. Based on the results in the paper [44], we can see that the first scenario is proven to be more accurate. Therefore in this paper, we will only consider the first scenario to generate the new HERIDAL dataset. As the input size of the object detector increases the computational time on GPUs also increases, as explained and illustrated in the experimental section.

C. PROPOSED METHOD

The next step after preprocessing the HERIDAL dataset is to train the newly generated dataset on deep-learning architectures. We propose the implementation of EfficientDET architecture and ensemble learning based on Bidirectional Feature Pyramid Network (Bi-FPN) and Fully Connected Feature Pyramid Network (FC-FPN).

One of the main idea of selecting and implementing the EfficientDET architecture for our research is due to its better accuracy compared to FasterRCNN [27], YOLOv3 [32] and the other existing object detectors. From the EfficientDET paper [29], [30], we can observe that the method can overcome the drawback of the multi-scale feature fusion problem. EfficientDET can be explained into two steps: imagenet-based pre-trained EfficientNETs and repeated feature extraction networks. It is difficult to scale the network uniformly based on depth, width, and resolution for most other well-known object detectors.The EfficientDET paper describes how to utilize a compound coefficient (ϕ) to effectively scale network parameters. The method explains the implementation of cross-scale connections with the EfficientNET backbone followed by various feature extractors such as Feature Pyramid Network (FPN) [28], BiFPN, Path Aggregation Network (PANet) [54], Neural Architecture Search (NAS-FPN) [55], and FC-FPN. In this paper, we propose to implement ensemble learning based on BiFPN and FC-FPN. One of the main reasons to choose Bi-FPN and FC-FPN over the other existing feature fusion models is due to better performance and accuracy of the modules. The proposed approaches are organised into three scenarios, which are explained below.

The BiFPN layer is implemented in the first scenario to get the features from level 3 to level 7 of the EfficientNET backbone as shown in the Fig. 6. Based on the compound coefficient, the HERIDAL dataset is considered as an input image, where the smallest resolution of 512 is considered as zero, while the other image resolutions 640, 768, 896 and 1024 are counted numerically from 1 to 4 on ϕ value. The width of the BiFPN feature network depends on the value of the compound coefficient (ϕ) as explained in the EfficientDET paper. If the network input is 512 image resolution then the ϕ value of zero is added with numerical three, which

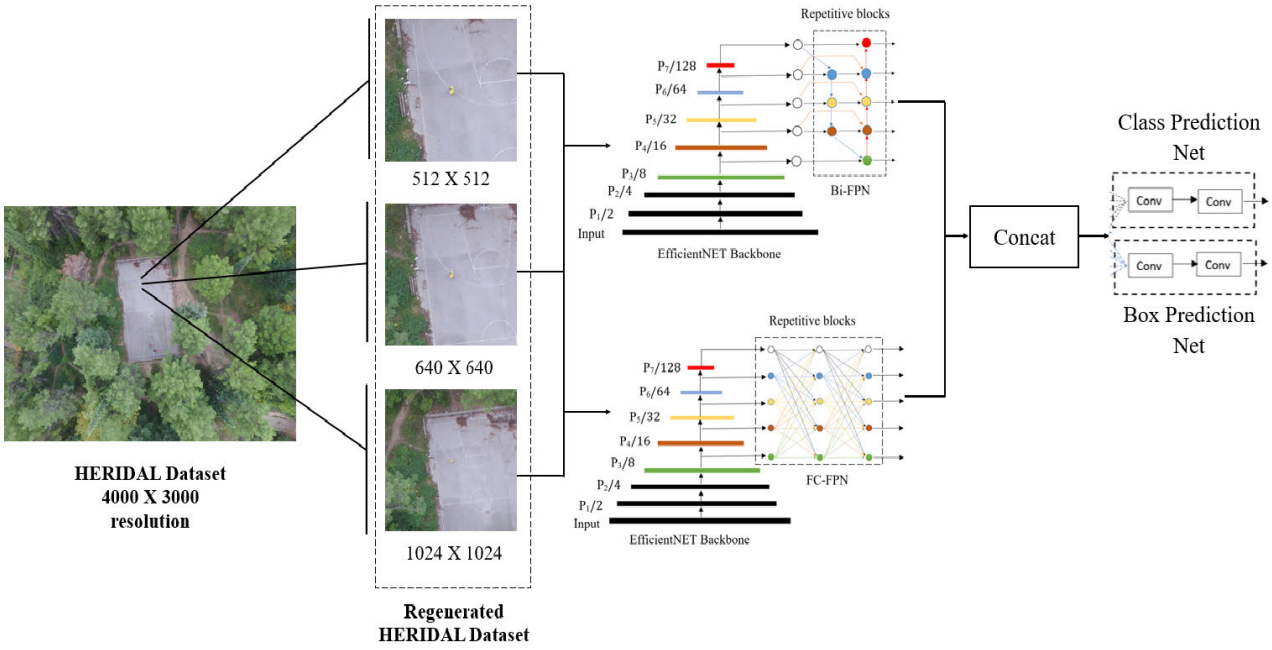

FIGURE 3. A complete proposed object detection model based on EfficientDET and ensemble learning [29], [30].

makes the total of three BiFPN feature networks. If the input size is 640 image resolution, then the width of the BiFPN feature network will be four, adding ϕ value one with the constant three, making a versatile architecture. As the width of the network increases, it takes more computational time on GPUs. We do not implement the continuous step provided by the EfficinetDET network study, which discusses how to use the object class identification and bounding box prediction. Instead, we will save the finest feature maps from training the BiFPN network, which will be used in ensemble learning. In the second scenario, we will train the EfficientNET backbone with FC-FPN by mostly following the same procedure as in the first scenario by considering the compound coefficient for input image generation and also for the width of the FC-FPN network as shown in Fig. 7. After training the network, we will save the best features as we did in the first step of BiFPN. Once we have the best features from both networks, we explain the final step based on ensemble learning. As shown in the equation (1) of ensemble learning, we will concatenate the best features from multiple networks to enhance the performance. To improve the accuracy of our research, we will concatenate the best feature maps from both BiFPN and FCFPN feature networks, as indicated in the Fig. 3. After concatenating the feature maps from both cases, we will use the object class and bounding box prediction networks described in the EfficientDET. The architecture considers the fast normalized fusion method over softmax based fusion as the softmax will slow down the GPU hardware, whereas the fast normalized fusion method has increased the GPU memory by 30%. This part is significant while implementing ensemble learning, as training the two different networks will consume a lot of memory and time. The equation (2) represents the final learning results. In the next section of the paper, all of the experimental results are discussed and shown.

$$Concat = Featuremaps(BiFPN + FCFPN) \quad (1)$$

$$Finalresult = Concat + (Class, Boxprediction) \quad (2)$$

D. METRIC CALCULATION

In most object detector models, mAP (mean Average Precision) is the popular metric to evaluate the results. The mAP computes the score by comparing the ground-truth bounding box to the detected bounding box. The higher the score, the better the model detection. The mAP is the AP mean, it is computed for each class and averaged to obtain the mAP. In SAR operations, we must detect only humans as a class. As mentioned in the equations (3) (4), average precision computes the precision value for recall values ranging from 0 to 1. The precision score is a measure of how accurate your predictions are i.e., the percentage of predictions are exactly correct, whereas recall evaluates how well you can locate all of the models positives.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

where TP = True positive, TN = True negative, FP = False positive, and FN = False negative

In general, the Average Precision (AP) may be defined by calculating the area under the precision-recall curve using the

equation (5), which can be visualized using the preceding calculations.

$$AP = \int_0^1 p(r)dr \tag{5}$$

Intersection over Union (IoU) is calculated by using the equation (6) by dividing the area of overlap or intersection between the two boxes by the area of their union. The higher the IoU value the higher will be the accuracy. In our scenario, we have considered to detect human object class if the $IoU \geq 0.5$.

$$IoU = \frac{area\ of\ overlap}{area\ of\ union} \tag{6}$$

IV. EXPERIMENTS AND RESULTS

To evaluate the proposed and stated methods in this paper, we will consider implementing and evaluating the results on regenerated HERIDAL dataset [44] in diverse resolutions of 512, 640 and 1024. Following the HERIDAL dataset regeneration, the images has nearly doubled from 1545 to 3000 for training and from 256 to 500 for validating images. By opting for the regeneration paper [44] for the HERIDAL dataset, we have increased the possibility to train various image resolutions. The main reason to choose the image resolution of 512 but not lower than this is due to the coefficient value proposed in EfficientDET paper [29]. The EfficientDET models B0 to B7 depends on this coefficient value, by changing the input size lower than 512 there will be a mismatch in the layers and the EfficientDET network will be out of bound. In this section, we will explain all the potential experiments and results based on the proposed ensemble learning model. We intend to conduct experiments in three different sections, as mentioned below.

- In the first step, we will train EfficientDET architecture with Bidirectional Feature Pyramid Network (BiFPN)
- For second step, we will train EfficientDET architecture with Fully Connected Feature Pyramid Network (FC-FPN)
- In the last step, we concatenate the best features from the above two sections and train the model to obtain our class prediction and bounding box

A. EXPERIMENTS WITH BiFPN

This section shows the continuation of experiments from paper [44], published at the beginning of 2021. The paper describes the training techniques for EfficientDET with BiFPN for 512 image resolution on distinct hyperparameters by considering multiple optimizers to discover the optimum and promising results to detect the humans in the HERIDAL regenerated dataset. According to the paper, the results for step 1 have achieved 91.27% mAP(Mean Average Precision) by adopting RMSprop optimizer, where we freeze the backbone and train the BiFPN from level 3 to level 7 to extract the best features for human detection. The fine-tuning of hyperparameters is quite critical. After experimenting and observing with batch normalization, we have set our hyperparameters as batch size 32, epoch 50, a step of 1000 and a learning rate of 0.001.

In the second step, we will unfreeze the backbone and train the EfficientDET network by using relevant and compatible parameters to train the dataset on Graphics Processing Unit (GPU), for this step the accuracy has improved to 93.29% by adopting RMSprop optimizer. We have set the hyperparameters of batch size 4, epoch 50, a step of 1000 and a low learning rate of 0.0001 for step 2. The main reason to select lower batch size in the second step is due to the utilization of memory on GPUs and also the computational training time. If the step size is higher then the model will be computational expensive also. So opting for lesser batch size in step 2 was a better choice to balance the model and the computational training time. We have also trained the model on various hyper-parameter tuning like varying epoch size from 50-150, batch size from 4-16 and also varying other parameters, but we have presented the most optimized parameters in the paper. We will continue the experiments on multiple HERIDAL dataset image resolutions of 640 and 1024, by various tuning parameters and optimizers to identify the most appropriate results by plotting the differences in computational time for training our model and mAP, as shown in the table 6. By observing the table 2 of computational time for training various image resolutions, we can observe 640 and 1024 image resolution takes more time to train the proposed model compared to 512 image resolution in both step 1 and step 2. By using 640 image resolution, we have achieved the mAP of 91.05% for step 1 and 91.52% mAP for step 2 by considering the same hyperparameters and optimizer as 512 image resolution experiments above. For the last step, we have tested the model on 1024 image resolution with the same hyperparameters as in other image resolutions to get the mAP of 88.07% and 89.56% respectively for step 1 and step 2 by adopting RMSprop optimizer. We can observe from these statistics that the mAP decreases as image resolution increases because the model extracts fewer features,

TABLE 2. Table showing the computational time comparison for training various HERIDAL image resolutions on BiFPN.

Experiments	Computational time for training
HERIDAL Dataset with 512 image resolution using EfficientDET BiFPN for step 1	5.24 hours
HERIDAL Dataset with 512 image resolution using EfficientDET BiFPN for step 2	2.05 hours
HERIDAL Dataset with 640 image resolution using EfficientDET BiFPN for step 1	9 hours
HERIDAL Dataset with 640 image resolution using EfficientDET BiFPN for step 2	5.08 hours
HERIDAL Dataset with 1024 image resolution EfficientDET BiFPN for step 1	11.14 hours
HERIDAL Dataset with 1024 image resolution EfficientDET BiFPN for step 2	8.54 hours

and the computational time also increases as image resolution increases. Based on these results, we can conclude that the HERIDAL 512 image resolution has significant advantages in terms of computational training time and accuracy.

B. EXPERIMENTS WITH FC-FPN

This section describes the experimental training approach on various HERIDAL image resolutions based on EfficientDET with Fully Connected- Feature Pyramid Network (FC-FPN). We will initially train the proposed model in two steps based on 512 image resolution. In the first step, we will freeze the EfficientNET backbone and train the FC-FPN from level 3 to level 7 to get the best features from the network. While training the model, tuning hyperparameters is computationally expensive and crucial. After testing and experimenting, we decided on a batch size of 32, an epoch of 50, a step of 1000, and a learning rate of 0.001 using batch normalization and the RMSprop optimizer. By following these parameters, we have achieved an mAP of 91.46%.

For the second step of FC-FPN, we will unfreeze the EfficientNET backbone and train the whole network using the optimal hyperparameters and improve the accuracy to 93.31% as shown in table 6. To achieve this accuracy, we have set the hyperparameters of batch size 4, epoch 50, a step of 1000 and a low learning rate of 0.0001 using RMSprop optimizer. We will also conduct additional experiments on different image resolutions of 640 and 1024 to calculate the computational time for training the model as shown in table 3. By examining the table, we can say that the 1024 image resolution will require more time to train the FC-FPN model than both the 512 and 640 image resolutions. We can also observe the mAP of various image resolutions in the mentioned table 6. From the table, we can notice that using 640 image resolution has achieved 90.47% mAP for step 1 and 91.86% for the step 2 and 1024 image resolution has achieved 88% mAP for step 1 and 89.45% for step 2 by adopting RMSprop optimizer. Based on the results of these experiments, we may conclude that the HERIDAL 512 image resolution achieved better accuracy than other image resolutions with low computational training time.

TABLE 3. Table showing the computational time comparison for training various HERIDAL image resolutions on FCFPN.

Experiments	Computational time for training
HERIDAL Dataset with 512 image resolution using EfficientDET FC-FPN for step 1	5.14 hours
HERIDAL Dataset with 512 image resolution using EfficientDET FC-FPN for step 2	2.11 hours
HERIDAL Dataset with 640 image resolution using EfficientDET FC-FPN for step 1	8.49 hours
HERIDAL Dataset with 640 image resolution using EfficientDET FC-FPN for step 2	5.25 hours
HERIDAL Dataset with 1024 image resolution EfficientDET FC-FPN for step 1	11 hours
HERIDAL Dataset with 1024 image resolution EfficientDET FC-FPN for step 2	9.1 hours

TABLE 4. Table showing the computational time comparison for training various HERIDAL image resolutions using ensemble learning.

Experiments	Computational time for training
HERIDAL Dataset with 512 image resolution using ensemble learning	9.42 hours
HERIDAL Dataset with 640 image resolution using ensemble learning	18.35 hours
HERIDAL Dataset with 1024 image resolution using ensemble learning	36.71 hours

C. FINAL RESULTS AND DISCUSSION

The possible experiments and observations based on ensemble learning using BiFPN and FC-FPN are discussed in this final section. We will select the best features from the previous two sections of Bi-FPN and FC-FPN and train the model on various HERIDAL image resolutions as mentioned in the above experiments. At first, we will take the best features of the HERIDAL dataset from 512 image resolution of both Bi-FPN step 2 and FC-FPN step 2 and train the network on class prediction and bounding box. By doing this, we can determine the precise position and number of persons detected in the 512 regenerated image resolution. We conducted experiments on several optimizers using a variety of hyperparameters and discovered the optimal tuning values as the batch size of 32, epoch of 50, a step of 1000 and a learning rate of 0.001 with batch normalization on RMSprop optimizer to get the accuracy of 95.11%mAP. For the 640 image resolution of the HERIDAL dataset, we will

TABLE 5. Table comparing the results of different proposed models based on HERIDAL dataset with our results from the paper.

Object detection model	mAP Calculated
Kundid Vasić et al. [42]	68.89%
Božić-Štulić et al. [39]	88.9%
Nayee Muddin Khan et al. [44]	93.29%
mAP based on EfficientDET and Ensemble learning on 1024 image resolution	90.06%
mAP based on EfficientDET and Ensemble learning on 640 image resolution	92.63%
mAP based on EfficientDET and Ensemble learning on 512 image resolution	95.11%

TABLE 6. Table showing the mAP results for different HERIDAL dataset image resolution based on BiFPN, FC-FPN and ensemble learning.

Experiments	HERIDAL dataset with 512 image resolution	HERIDAL dataset with 640 image resolution	HERIDAL dataset with 1024 image resolution
mAP based on EfficientDET with BiFPN step 1	91.27%	91.05%	88.07%
mAP based on EfficientDET with BiFPN step 2	93.29%	91.52%	89.56%
mAP based on EfficientDET with FC-FPN step 1	91.46%	90.47%	88%
mAP based on EfficientDET with FC-FPN step 2	93.31%	91.86%	89.45%
mAP based on EfficientDET and Ensemble learning	**95.11%**	**92.63%**	**90.06%**

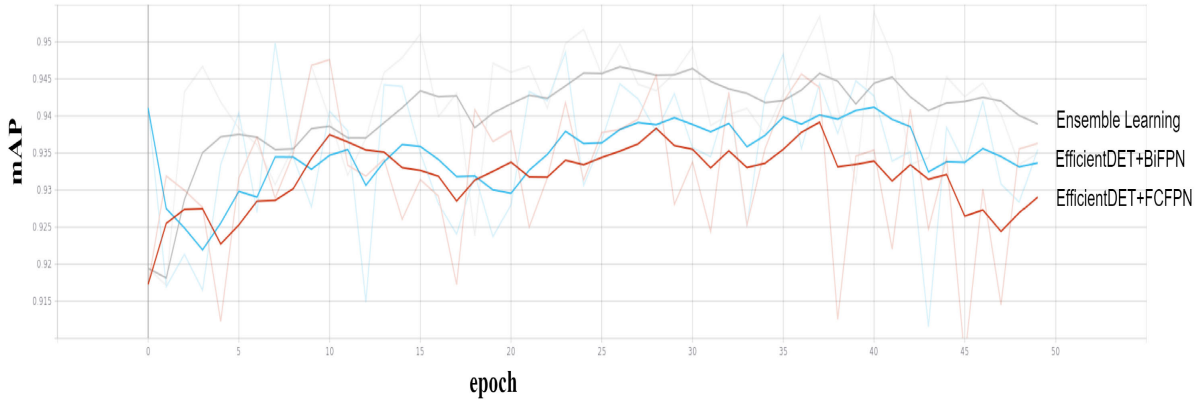

FIGURE 4. Showing the plotted results of BiFPN, FCFPN, and Ensemble Learning based on EfficientDET architecture [29].

TABLE 7. System specifications used for training the proposed model.

Hardware/ Software	Specifications/ Version
CPU	Intel Core i5 8 Generation
GPU	NVIDIA GeForceGTX1080
RAM	12GB
Anaconda	3-5.1.0
Python	3.4.2.17
Tensorflow	1.15
Keras	2.2.5

conduct the experiments by following the same procedure of getting the best features from Bi-FPN and FC-FPN step 2 and train the network to detect the humans on the 640 image resolution. By implementing the same optimal parameters as above, we have achieved the accuracy of 92.63% as shown in table 6. Based on the results of the previous two experiments, we can see that the accuracy of 512 image resolution is significantly better than that of 640 image resolution. For the final ensemble learning experiment, we conducted our experiments on HERIDAL 1024 generated image resolution using the same pattern and procedures as in the previous steps, extracting the best features and using the same hyperparameters, we achieved an accuracy of 90.06% mAP, which is significantly lower than the previous two experiments. From this, we can conclude that the accuracy is much better in 512 image resolution compared to the other image resolutions and the computational time for training the networks is also much lesser using 512 resolution as shown in the table 4. The Fig. 4 explains the graph for calculating mAP values based on three different scenarios as EfficientDET with BiFPN, EfficientDET with FCFPN and Ensemble Learning. From the Fig. 4, the mAP is calculated on y-axis with respect to epochs on the x-axis. It can be observed by implementing ensemble learning, we have achieved better accuracy than BiFPN and FC-FPN feature networks. In the table 5, we have compared all the previously mentioned papers in the literature for the detection of humans on the HERIDAL dataset. We can conclude that the results plotted in this paper have outperformed with an mAP of 95.11%. From the Fig. 5, we can view the detected humans on the HERIDAL dataset in various image resolutions of 512, 640, and 1024. By observing the figure, we can say detection of humans in 512 has better human detection, as the other image resolutions of 640 and 1024 have some undetected humans. By observing Fig. 5, you can notice there are some falsely detected objects at the corners of the higher resolution images, this is one of the issues we will consider to improve for the future work. By considering ensemble learning on BiFPN and FC-FPN feature models, we have used two GPUs running simultaneously for training the models. The entire experiments were carried out on two NVIDIA GEFORCE GTX 1080 GPUs with 12 GB of RAM. The other

FIGURE 5. Example of detecting humans in 512 (Left), 640 (Middle), and 1024 (Right) HERIDAL dataset image resolutions based on ensemble learning.

feature networks, like PANet and NAS-FPN are not included in ensemble learning as it needs to have more powerful GPUs, which makes the model computationally highly expensive. By this proposed detection model from the paper, we can save many lives by solving the problem of detecting humans in SAR operations using ensemble learning.

V. CONCLUSION

We can save many individuals who are involved in mountain accidents by detecting humans in SAR missions. We can minimize the cost and time involved in traditional SAR operations by using drones. We have examined the state-of-the-art person detectors implemented on the HERIDAL dataset and proposed an ensemble learning-based method for detecting humans for our research.

As the existing HERIDAL dataset images are 4000×3000, it is difficult to train object detection models for such large image resolution. We have explained the implementation of a preprocessing step to solve this problem. The proposed method demonstrates how to regenerate the HERIDAL dataset at various image resolutions ranging from 512 to 1024.

Following the preprocessing step, we proposed an ensemble learning method in which we adopted EfficientDET architecture with BiFPN as one branch and EfficientDET architecture with FC-FPN as the other. Both the steps are

concatenated with the best features to follow up with the class and bounding box networks to detect humans in SAR operations. All of the experimental results are discussed and presented in the paper with various figures and plotted tables. The proposed model in the paper has implemented on the HERIDAL dataset, which is specifically created for SAR operations. By implementing an ensemble learning method in the paper based on EfficientDET architecture using feature networks, we have improved our accuracy from the existing state-of-the-art human detectors. For our future research work on human detection in aerial images, we are working on modifying the backbone architecture of EfficietDET and finding the solutions for knowledge distillation from the other object detection models.

APPENDIX

In this section, we would like to show the clear EfficientNET architectures explained in the proposed section based on BiFPN and FCFPN feature networks. The Fig. 6 shows the detail EfficientNET backbone with the repetitive blocks of Bi-FPN based on the compound coefficient, while the Fig. 7 is based on FC-FPN feature network.

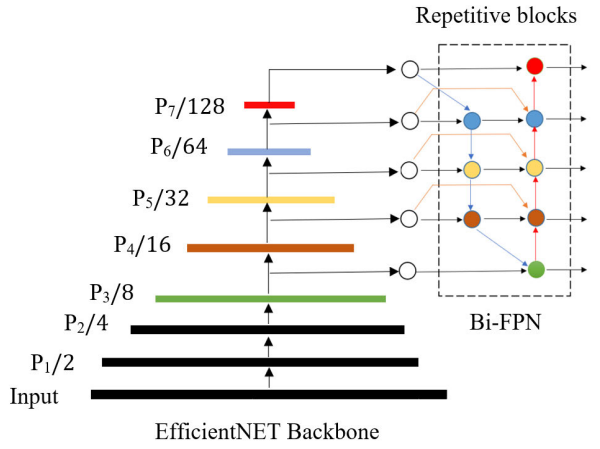

FIGURE 6. EfficientNET backbone with Bi-FPN.

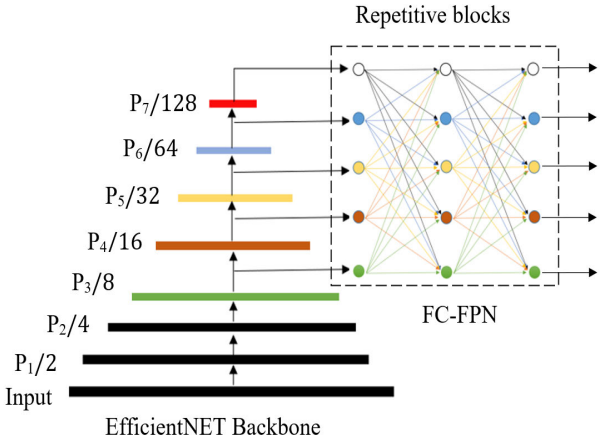

FIGURE 7. EfficientNET backbone with FC-FPN.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPUs used for this research.

REFERENCES

[1] X. Wang, "Deep learning in object recognition, detection, and segmentation," *Found. Trends Signal Process.*, vol. 8, no. 4, pp. 217–382, 2016.

[2] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 694–711, May 2006.

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[4] I. Martinez-Alpiste, G. Golcarenarenji, Q. Wang, and J. M. Alcaraz-Calero, "Search and rescue operation using UAVs: A case study," *Expert Syst. Appl.*, vol. 178, Sep. 2021, Art. no. 114937.

[5] M. Kampouraki, G. A. Wood, and T. R. Brewer, "Opportunities and limitations of object based image analysis for detecting urban impervious and vegetated surfaces using true-colour aerial photography," in *Object-Based Image Analysis*. Berlin, Germany: Springer, 2008, pp. 555–569.

[6] S. Liao. (2019). [Online]. Available: https://www.theverge.com/2019/4/16/18410723/notre-dame-fire-dji-drones-tracking-stopped-thermal-cameras DJI drones helped track and stop the Notre Dame fire. The Verge

[7] M. Paul, S. M. E. Haque, and S. Chakraborty, "Human detection in surveillance videos and its applications—A review," *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, pp. 1–16, Dec. 2013.

[8] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[9] H. Majesty's, "Inspectorate of constabulary and fire and rescue services. Planes, drones and helicopters: An independent study of police air support," HMICFRS, London, U.K., 2017, pp. 1–97.

[10] P. Zhang, Y. Zhong, and X. Li, "SlimYOLOv3: Narrower, faster and better for real-time UAV applications," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9.

[11] P. Rudol and P. Doherty, "Human body detection and geolocalization for UAV search and rescue missions using color and thermal imagery," in *Proc. IEEE Aerosp. Conf.*, Mar. 2008, pp. 1–8.

[12] C. Burke, P. R. McWhirter, J. Veitch-Michaelis, O. McAree, H. A. G. Pointon, S. Wich, and S. Longmore, "Requirements and limitations of thermal drones for effective search and rescue in marine and coastal areas," *Drones*, vol. 3, no. 4, p. 78, Oct. 2019.

[13] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019.

[14] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, Jul. 2020.

[15] Z. Wu, K. Suresh, P. Narayanan, H. Xu, H. Kwon, and Z. Wang, "Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1201–1210.

[16] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Kauai, HI, USA, Dec. 2001, pp. 1–13, doi: 10.1109/CVPR.2001.990517.

[18] M. Kristo, M. Ivasic-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using Yolo," *IEEE Access*, vol. 8, pp. 125459–125476, 2020.

[19] P. Doherty and P. Rudol, "A UAV search and rescue scenario with human body detection and geolocalization," in *Proc. Australas. Joint Conf. Artif. Intell.* Berlin, Germany: Springer, 2007, pp. 1–13.

[20] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Aug. 2017, pp. 1–6.
[21] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, *arXiv:1511.08458*.
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.* vol. 60, no. 2, pp. 91–110, 2004.
[24] R. K. McConnell, "Method of and apparatus for pattern recognition," U.S. Patent 4 567 610, Jan. 28, 1986.
[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.
[26] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–799, 1995.
[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[28] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[29] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.
[30] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.
[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.
[32] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[33] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[34] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Zürich, Switzerland: Springer, 2014, pp. 740–755.
[37] D. A. Zherdev, E. Y. Minaev, V. V. Procudin, and V. A. Fursov, "Object recognition using real and modelled SAR images," *Proc. Eng.*, vol. 201, pp. 503–510, Jan. 2017.
[38] D. Cao, Z. Chen, and L. Gao, "An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks," *Hum.-centric Comput. Inf. Sci.*, vol. 10, no. 1, pp. 1–22, Dec. 2020.
[39] D. Božić-Štulic, Ž. Marušić, and S. Gotovac, "Deep learning approach in aerial imagery for supporting land search and rescue missions," *Int. J. Comput. Vis.*, vol. 127, no. 9, pp. 1256–1278, 2019.
[40] S. Gotovac, V. Papic, and Z. Marusic, "Analysis of saliency object detection algorithms for search and rescue operations," in *Proc. 24th Int. Conf. Softw., Telecommun. Comput. Netw. (SoftCOM)*, Sep. 2016, pp. 1–6.
[41] Ž Marušic, D. Božić-Štulić, S. Gotovac, and T. Marušić, "Region proposal approach for human detection on aerial imagery," in *Proc. 3rd Int. Conf. Smart Sustain. Technol. (SpliTech)*, Jun. 2018, pp. 1–6.
[42] M. K. Vasić and V. Papić, "Multimodel deep learning for person detection in aerial images," *Electronics*, vol. 9, no. 9, p. 1459, Sep. 2020.
[43] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 391–405.
[44] N. M. K. Dousai and S. Loncaric, "Detection of humans in drone images for search and rescue operations," in *Proc. 3rd Asia Pacific Inf. Technol. Conf.*, Jan. 2021, pp. 69–75.
[45] M. Everingham, A. Zisserman, C. K. I. Williams, L. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, and S. Duffner, "The 2005 Pascal visual object classes challenge," in *Proc. Mach. Learn. Challenges Workshop*. Berlin, Germany: Springer, 2005, pp. 117–176.
[46] P. Zhu, L. Wen, D. Du, X. Bian, H. Ling, Q. Hu, H. Wu, Q. Nie, H. Cheng, C. Liu, and X. Liu, "VisDrone-VDT2018: The vision meets drone video detection and tracking challenge results," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 1–23.
[47] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," 2018, *arXiv:1804.07437*.
[48] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 445–461.
[49] M. Barekatain, M. Marti, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017.
[50] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 549–565.
[51] A. Robicquet, A. Alahi, A. Sadeghian, B. Anenberg, J. Doherty, E. Wu, and S. Savarese, "Forecasting social navigation in crowded complex scenes," 2016, *arXiv:1601.00998*.
[52] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4145–4153.
[53] S. Sambolek and M. Ivasic-Kos, "Automatic person detection in search and rescue operations using deep CNN detectors," *IEEE Access*, vol. 9, pp. 37905–37922, 2021.
[54] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
[55] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.

NAYEE MUDDIN KHAN DOUSAI (Member, IEEE) was born in India, in 1990. He received the master's degree in robotics engineering from the University of Petroleum and Energy Studies, Dehradun, India, in 2014, and the master's degree in computer vision from the University of Bourgogne, Dijon, France, in 2018. He is currently pursuing the Ph.D. degree in computer science from the University of Zagreb, Croatia. His research interests include computer vision, artificial intelligence, and object detection.

SVEN LONČARIĆ (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Cincinnati, Cincinnati, OH, USA, in 1994, as a Fulbright Scholar. He is currently a Full Professor of electrical engineering and computer science with the Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia. With his students and collaborators, he has coauthored more than 200 publications in scientific journals and conferences. He is the Founder of the Center for Computer Vision, University of Zagreb, where he is also the Head of the Image Processing Group. He has served as the Co-Director for the National Center of Research Excellence in Data Science and Cooperative Systems. He is a member of the Croatian Academy of Technical Sciences. He received several awards for his scientific and professional work. He was the Chair of the IEEE Croatia Section.