Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Human Detection from Drone using You Only Look Once (YOLOv5) for Search and Rescue Operation

Fadhlan Hafizhelmi Kamaru Zaman[1,*], Nooritawati Md Tahir[2], Yusnani Mohd Yusoff[3], Norashikin M. Thamrin[3], Ahmad Hafizam Hasmi[4]

[1] Vehicle Intelligence and Telematics Lab (VITAL), College of Engineering, Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia
[2] Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia
[3] College of Engineering, Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia
[4] Institut Perubatan Forensik Negara, Hospital Kuala Lumpur, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 9 December 2022
Received in revised form 14 April 2023
Accepted 22 April 2023
Available online 13 May 2023

Keywords:
Human detection; object detection; deep learning; drone search and rescue

ABSTRACT

Drones are unmanned aerial vehicles that can be remotely operated to perform a variety of tasks. They have been used in search and rescue operations since the early 2000s and have proven to be invaluable tools for quickly locating missing persons in difficult terrain and environment. In certain cases, automated human detection on drone camera feed can help the responder to locate the victims more effectively. In this work, we propose the use of a deep learning method called You Only Look Once version 5, or YOLOv5. The YOLOv5 model is trained using data collected during a simulation of search and rescue operations, where mannequins were used to represent human victims. Video was acquired using DJI Matrice 300 drone with Zenmuse H20T camera which flew around an area with various terrains such as farms, ravines, and river of more than 15,000 m2, at a height of 40 meters. The drone used grid, circular and zigzag flying patterns, with three different levels of camera zooms, and the data was captured on different days and times. The total duration of the video collected at 1080p@30fps is 148 minutes 26 seconds. Five pretrained models of YOLOv5 with different complexities were trained and tested using this dataset. Results showed that pretrained yolov5l6 model delivered the best precision, recall and mAP50 rate at 0.668, 0.303 and 0.346 respectively. Besides, the experiment also showed that we can improve the overall performance by using images acquired at 6x zoom magnification level where precision, recall, and mAP50 rate are increased to 0.846, 0.543, and 0.591 respectively. yolov5l6 model also delivered an acceptable inference time of 43ms per 1920x1080 resolution image, thus it can run at a respectable 23fps.

* *Corresponding author.*
E-mail address: fadhlan@uitm.edu.my

https://doi.org/10.37934/araset.30.3.222235

1. Introduction

Drones are unmanned aerial vehicles which can fly quickly, maneuver easily, and capable to access areas that are difficult or impossible to reach by traditional search and rescue methods. The use of drones has revolutionized search and rescue (SAR) operations where they are not only efficient but also cost-effective, making it possible to launch multiple searches in a short period of time.

Furthermore, the use of drones can help reduce the risk of injury to personnel, as well as provide better visibility and situational awareness for first responders. Thus, drones have become an indispensable tool for search and rescue operations and are transforming the way we respond to emergencies. As a matter of fact, drones can be equipped with a variety of sensors, including thermal imaging cameras, to quickly locating and rescuing people in difficult and dangerous situations. They can also be used to swiftly survey an area to assess the best approach for a rescue operation. Furthermore, they can provide video feeds to search and rescue teams, allowing them to quickly assess the situation and take the necessary action. Besides, drones are useful also for a wide range of tasks, including monitoring crime scenes, spotting marine life, assessing habitat loss, monitoring crops, and mapping vegetation [1,2]. This has spurred research including human detection from drone, drone computer vision [3], smart search system consisting of autonomous flying drone, search algorithms, protocols [4] and can be enhanced with algorithms to detect early forest fire [5].

In cases where the drones are used in search and rescue operation that covers a large area, the use of automated human detection from drone camera feed is necessary to help the responder locate the victims more effectively. Drone camera has been used previously by Al-Naji *et al.,* [6] to remotely detect survivors' periodic chest movements that generate cardiopulmonary motion. Eight human test subjects and one mannequin in various attitudes were used, and the results indicate that motion detection on the body surface of the survivors is probably beneficial for spotting live signals without making direct physical contact.

Mishra *et al.,* [7] suggests a methodology for human detection and activity identification that draws inspiration from the pyramidal feature extraction of Single Shot Detector (SSD). When used with the suggested dataset, the proposed model achieves 0.98mAP, which is a considerable contribution. Additionally, the suggested model outperforms the most recent detection models in the literature by 7% when applied to the standard Okutama dataset. Previously, Baeck *et al.,* [8] employed mannequin-like dummies and nadir drone images that combines deep learning and photogrammetric algorithms to find humans and place them on an overview orthomosaic and 3D terrain map with their location.

Meanwhile, an overall architecture for drone hardware that enables fast exploration of GPS-denied environment was proposed by Lee *et al.,* [9], including practical methods for victim detection by using DJI Matrice 100 drone and utilize LIDAR for global mapping and Intel RealSense for local mapping. Similarly, Rizk *et al.,* [10] looked on adding processing units running emergent AI-based detectors to UAVs. The suggested system can relay the correct coordinates to the ground station and detect people in real time.

Among recent computer vision methods, You Only Look Once (YOLO) has received a lot of attention and it has been successfully used in many computer vision applications. YOLOv5 has been successfully used in several drone-assisted applications such as tassel detection in maize using UAV-based RGB imagery [11], object detection from drone [12], forest fire detection [13], and real-time monitoring of isolated places and automatically detecting stranded humans during floods [14]. Recently, Lagman *et al.,* [15] proposed a human detection and counting algorithms from drone images and thermal cameras based on YOLOv5. Besides, Sruthi *et al.,* [14] incorporates a YOLOv5 object tracking convolutional neural network method for quicker detection of people, an Open-

Source autopilot system model, an APM 2.8 multicopter flight controller that effectively stabilizes the flight, and an Open-Source autopilot system model.

YOLO is a real-time object detection system developed by Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. The first version of YOLO was published in 2015, called YOLOv1 [16]. It was designed to be fast and accurate, able to run in real-time on a standard desktop computer. YOLOv2 made several improvements over the original YOLO system [17]. It used a modified version of the Darknet neural network architecture and was trained on the COCO dataset, which contains 80 object classes. YOLOv3 introduced several important improvements over the previous versions [18]. It used a new neural network architecture called "Darknet-53" and was trained on the COCO dataset. It also introduced several techniques to improve the speed and accuracy of the model, including multi-scale predictions and a new loss function called "focal loss." YOLOv4 was published in 2020 [19] and made further improvements to the model architecture and training techniques. YOLOv5 then introduced several new features and improvements over the previous versions.

In this work, YOLOv5 is proposed to be used to detect human from drone aerial images during a simulation of search and rescue operations. Mannequins are used as substitute to human, and they are scattered around the search and rescue area. The collected aerial images are then used to train YOLOv5 with variations in terms of different magnification levels and drone flying patterns to prove its effectiveness in detecting human in a challenging research area. As a result, the used of automated human detection in drones in search and rescue operations has several advantages over traditional search methods. For example, drones can cover large areas much more quickly than human search teams, and they can do so without putting additional people at risk. Drones can also operate in hazardous or hard-to-reach areas, such as steep mountain terrain or areas affected by natural disasters.

2. Methodology

In this section, the methods used in this work are described, including the data collection, data processing and labelling, modelling and training of YOLOv5 models, as well as the performance metrics used to evaluate the results.

2.1 Data Collection

The data used in this work has been collected on 9th November – 11th November 2021 at Pulau Sebang 78000 Alor Gajah, Melaka (GPS coordinate: 2.457139, 102.257239). The exact location is shown in Figure 1, within the blue shaded region. The surface of the area is around 15,000 m2 where the terrain includes farms, ravines, and rivers. The data collection was performed during a search and rescue simulation activity in collaboration between Universiti Teknologi MARA, Institut Perubatan Forensik Negara (IPFN), Angkatan Pertahanan Awam Malaysia (APM) and Aerodyne. In this simulation, similar to Baeck *et al.,* [8], mannequins are used to represent human victims, scattered around the area.

Fig. 1. The satellite image view showing the area used for data collection

DJI Matrice 300 drone was used to scour the area to look for victims and the location of the victim was then transmitted to the rescuers. The drone is flying roughly 40m above the ground and it was equipped with Zenmuse H20T camera which is used to capture the aerial view of the search area. The specification of the camera is given in Table 1.

Table 1
Zenmuse H20T camera specification

Sensor	1/1.7" CMOS, 20 MP
Lens	DFOV: 66.6°-4° Focal length: 6.83-119.94 mm (equivalent: 31.7-556.2 mm) Aperture: f/2.8-f/11 (normal), f/1.6-f/11 (night scene) Focus: 1 m to ∞ (wide), 8 m to ∞ (telephoto)
Resolution	3840x2160@30fps, 1920x1080@30fps

To perform the data collection process, the drone must fly around the search area by performing three types of flying patterns, which are grid flying pattern, circular flying pattern, and zigzag flying pattern. The flying patterns are shown in Figure 2. The flight distance is between 600 to 800 meters and the time taken to perform a single flight over the search area is between 5-6 minutes. While flying, the drone will capture the video at 3 different levels of magnification – no zoom, 2x zoom and 6x zoom. The videos are recorded by the drone in 1920x1080@30fps format resolution where the total duration of captured video is 148 minutes 26 seconds.

(a) Grid flying pattern (b) Circular flying pattern

(c) Zigzag flying pattern

Fig. 2. The flying patterns (a) grid flying pattern, (b) circular flying pattern and (c) zigzag flying pattern used by the drone during data collection

Once the videos are collected, the next process is to extract image frames from the videos. We define sampling frequency of 1 frame per second that will allow us to extract a single image every second. For example, for a 30-second video, we can get 30 still images. This is sufficient considering that the drone was not flying fast and the transition of focus area between images happened slowly. The videos are divided into 5 datasets and the details of the dataset including total size, total duration, number of images and the date taken are given in Table 2. Examples of no zoom, 2x zoom and 6x zoom images are shown in Figure 3.

Table 2
Detailed description of collected data in this work.

Dataset	Flying Patterns	Magnification levels	Total Size	Total Duration	# Images	Date Taken
Circle-1	Circle	No zoom, 2x, 6x	4.86GB	22min 26sec	1348	09/11/2021
Grid-1	Grid	No zoom, 2x, 6x	8.45GB	39min 00sec	2343	09/11/2021
Grid-2	Grid	No zoom, 2x, 6x	5.47GB	25min 14sec	1519	10/11/2021
Grid-3	Grid	No zoom, 2x, 6x	3.68GB	16min 58sec	1018	10/11/2021
Zigzag-1	Zigzag	No zoom, 2x, 6x	9.74GB	44min 58sec	2698	09/11/2021

(a) No zoom (b) 2x zoom

(c) 6x zoom

Fig. 3. Sample images showing the difference between aerial images taken with different zoom magnification levels at no zoom, 2x zoom and 6x zoom

2.2 Data Processing and Labeling

In total we manage to collect 32.2GB worth of videos, with a total of 8926 images extracted from the videos. The next step would be to label the subjects in the image to be used in our model training. The labeling process was carried out using Label Studio (https://labelstud.io/) which is a python-based open-source data annotation tool. Example of labelled data is shown in Figure 4. In total we have labelled 5,210 images. Out of these images 1,586 images contain our subjects (mannequins) and 3,624 images only contain background. Distribution of dataset used for training and validation of our models is shown in Table 3. According to Table 3, Circle-1 and Grid-1 dataset is used as training dataset, while Grid-2 is used as test dataset. Grid-3 and Zigzag-1 datasets are unlabeled and are used as inference datasets.

Fig. 4. Example of labelled instances showing the bounding box around the subject mannequin

Table 3
Distribution of dataset used in this work

Dataset	# Images	# Positive Images	Training/Test
Circle-1	1348	860	Training
Grid-1	2343	576	Training
Grid-2	1519	150	Test
Grid-3	1018	unlabelled	Inference
Zigzag-1	2698	unlabelled	Inference

2.3 YOLOv5 Training and Modelling

YOLOv5 has five different models called yolov5n, yolov5s, yolov5m, yolov5l and yolov5x. These models are built with the same components; however, they differ in the complexity of the model which is determined by times of module execution and the numbers of convolution kernels. In this work, five pre-trained YOLOv5 variants are used that accepts input size of 1280 pixels and the number of parameters is shown in Table 4. The overview of YOLOv5 architecture is given in Figure 5.

The core of the YOLOv5 architecture is a series of downsampling layers, which are used to reduce the spatial resolution of the input image while increasing the number of feature maps. This is done by using convolutional layers with a stride of 2, which effectively reduces the spatial dimensions of the input by half. After the downsampling layers, the feature maps are passed through a series of residual blocks, which are used to increase the representational capacity of the model. Each residual block is composed of two 3x3 convolutional layers, with a ReLU activation function applied after the first convolutional layer. The feature maps are then upsampled back to the original resolution using a series of upsampling layers. Finally, a series of convolutional layers are used to predict the bounding boxes and class probabilities for each object in the image.

Based on Figure 5, there are three prediction heads used by YOLOv5 to detect large, medium, and small objects respectively. The same head as YOLOv3 and YOLOv4 is used by YOLOv5. It is made up of three convolution layers that forecast where the bounding boxes (x, y, height, and width), scores, and object classes will be. Besides, the model neck is used to extract feature pyramids. This helps the

model to generalize well to objects on different sizes and scales. A feature pyramid network called PANet was utilized in YOLOv4 to enhance information flow and aid in the accurate localization of pixels for the purpose of mask prediction. This network has been changed in YOLOv5 by implementing the CSPNet approach.

Table 4
Number of parameters in YOLOv5 pre-trained models

YOLOv5 models	yolov5n6	yolov5s6	yolov5m6	yolov5l6	yolov5x6
#Parameters (M)	3.5	12.6	35.7	76.8	140.7

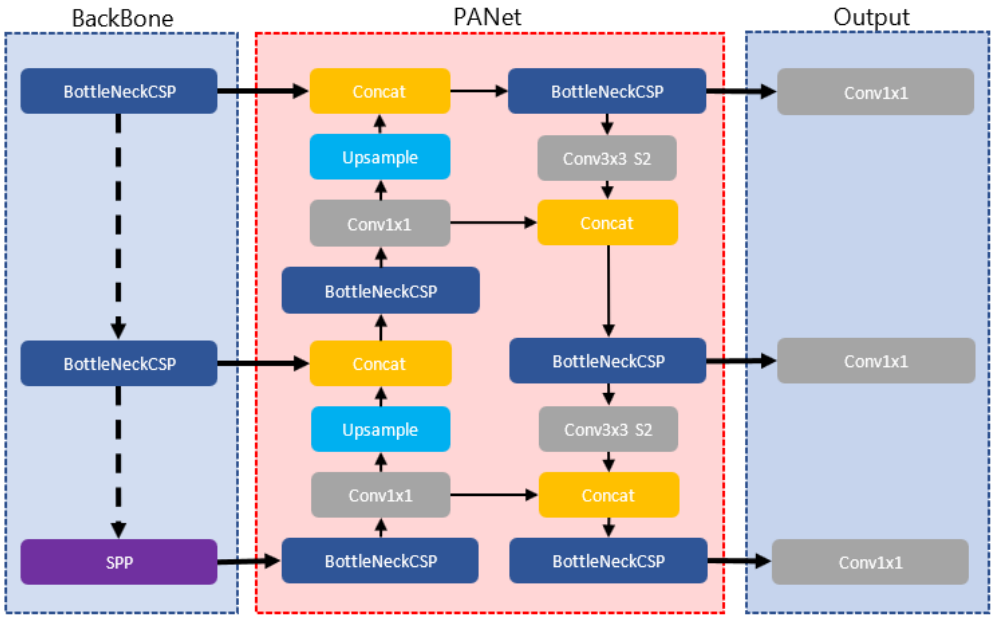


Fig. 5. The overview of YOLOv5 architecture

Additionally, YOLOv5 adopts the structure of PANet. First, the feature map is downsampled to reduce its size. Increasing the dimension after scaling also makes it easier to extract deep features through C3 layers and conv3 layers are used to change channels of the input feature maps. Then up sampling the feature map and enlarge the feature size. In this process, feature maps of the same size during down sampling are spliced in dimension. Finally, the feature map is down sampling again and the output feature map will be detected. This structure (PANet) is helpful for feature fusion of different detection layers.

2.4 Performance Evaluation Metrics

Performance metrics are used to evaluate machine learning models because they provide a way to measure how well the model is performing. They help us to understand the strengths and weaknesses of a model and can be used to compare different models to determine which one is the best. Without performance metrics, it would be difficult to know whether a model is improving or not, or whether one model is better than another. To ensure that the evaluation is correct, the right performance metrics must be used. However, evaluation metrics differ from work to work, often making their comparative assessment confusing and misleading [20]. Due to the scope of the implementation of our work, we chose to use precision, recall, mAP50 and mAP50:95 as our evaluation metrics.

Precision refers to the percentage of correct positive predictions made by the model out of all the positive predictions it made. It is a measure of how well the model can identify relevant objects. Recall, on the other hand, is the percentage of correct positive predictions made by the model out of all the relevant objects in the dataset. It is a measure of the model's ability to find all relevant cases [20]. To compute precision and recall, we must determine from each detected bounding box to be classified as

i. True positive (TP): A correct detection of a ground-truth bounding box.
ii. False positive (FP): An incorrect detection of a non-existing object or a misplaced detection of an existing object.
iii. False negative (FN): An undetected ground-truth bounding box.

Precision and recall can be calculated from Eq. (1) and Eq. (2)

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

The average precision (AP) is a performance metric that evaluates the precision-recall trade-off of a model. It is based on the area under a precision-recall curve that has been modified to remove the zig-zag pattern that can occur due to variations in the model's confidence levels. AP provides a single value summary of the model's performance on the task of predicting bounding boxes such as used in PASCAL Visual Object Classes (VOC) Challenge [21]. mean average precision (mAP) refers to average AP over all classes as shown in Eq. (3). mAP50 on the other hand denotes average mAP over different IoU thresholds, from 0.5 to 1.0 where a detection is considered as a TP only if its IoU is larger than 0.5. Similarly, mAP50:95 takes the average of mAP over different IoU thresholds, from 0.5 to 0.95, with a step of 0.05 [22, 23]. This puts significantly larger emphasis on localization compared to the PASCAL VOC metric which only requires IoU of 0.5 [24] .

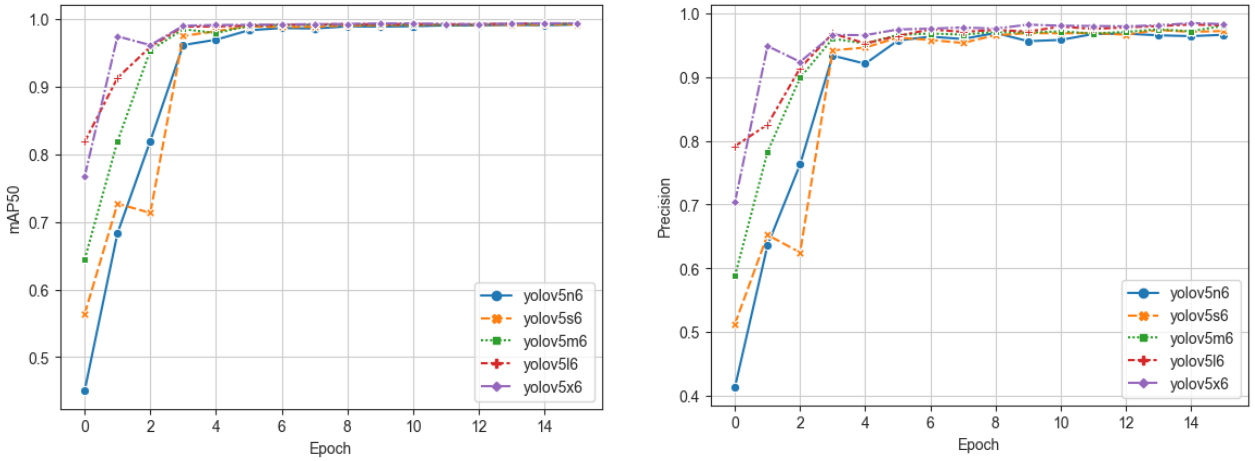
$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{3}$$

3. Results and Discussions

In this section, we discuss the results of human detection using YOLOv5 variants, namely yolov5n6, yolov5s6, yolov5m6, yolov5l6 and yolov5x6 as described earlier. The experiments were conducted specifically to evaluate YOLOv5 models and find the best performing model in terms of precision, recall and inference time. Then, the best model is used to evaluate the detection performance using aerial images with different zoom magnification levels. Firstly, YOLOv5 models are trained using the labelled data from the Circle-1 and Grid-1 datasets. These datasets include videos with different zoom magnification levels. The yolov5n6, yolov5s6, yolov5m6, yolov5l6 and yolov5x6 models are trained using the training parameters defined in Table 5. The training performance for the models is shown in Figure 6.

Table 5
Training parameters used in YOLOv5 training

Training parameters	Value
Image size	1920
Epoch	16
Batch size	2
Learning Rate	0.01
Momentum	0.937
Weight decay	0.0005
Optimizer	Stochastic Gradient Descent (SGD)



(a) (b)

Fig. 6. Training performance in terms of (a) precision and (b) recall over increasing epochs during YOLOv5 training based on different model variants

According to Figure 6, the precision and recall rates for all models started to increase significantly during epoch 2. The initial precision and recall at epoch=1 however depend on the complexity of the model, where yolov5l6 started with the highest precision and recall both at 0.78 and yolov5n6 started with the least precision and recall at 0.41 and 0.58 respectively. In general, all models reached the maximum precision and recall at epoch=10. Figure 7 shows the mAP50 score for all trained models based on epoch=1, epoch=10 and epoch = 16. Based on this figure, yolov5l6 and yolov5 started with higher initial mAP50 compared to other models. Like precision and recall, mAP50 scores for all models reached maximum score by epoch=10. Subsequently, all models are tested using unseen test dataset which is Grid-2 dataset. Grid-2 dataset is a challenging test dataset since the Grid-2 dataset is collected on different day, and the locations of the mannequins in this dataset are completely different from those used in training. The results are tabulated in Table 6.

According to Table 6, yolov5l6 gives the best precision, recall, mAP50 and mAP50:95 at 0.668, 0.303. 0.380 and 0.247 respectively. Despite having more parameters and higher complexity, yolov5x6 delivers lower performance compared to yolov5l6 model. As expected, yolov5n6 with the least number of parameters gives the worst performance and the fastest inference time at 22 s total inference time, or 14ms inference time per image. yolov5x6 model on the other hand has slowest inference time with 129 s total inference time, or 84ms inference time per image. yolov5l6 model gives total inference time of 66 s, where the inference time per image is 43ms. For reference, to achieve a 30fps video, a frame should be processed every 33ms. Thus, yolov5l6 model has just slightly higher processing time per frame, allowing the video to run at roughly 23fps which is sufficiently quick for real-time applications. Since yolov5l6 model has the highest performance and reasonable inference time, this model is used in the next experiment.

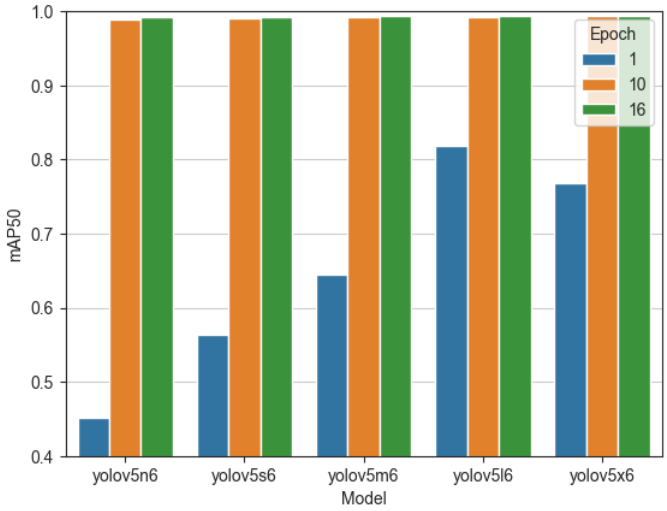


Fig. 7. The mAP50 scores for all trained YOLOv5 models

Table 6
The detection performance of variants of YOLOv5 models

Model	Precision	Recall	mAP50	mAP50:95	Inference time	
					Total (s)	*Image (ms)*
yolov5x6	0.631	0.298	0.346	0.232	129	84
yolov5l6	0.668	0.303	0.380	0.247	66	43
yolov5m6	0.506	0.246	0.280	0.172	46	30
yolov5s6	0.597	0.273	0.343	0.197	28	18
yolov5n6	0.428	0.297	0.250	0.140	22	14

Using yolov5l6 model, we evaluate the detection performance on images acquired from different type of videos captured under different zoom magnification levels, namely no zoom, 2x zoom and 6x zoom. This is important to determine which magnification level is more suitable to be used by the drone to get the best detection performance from the model. According to Figure 8, we showed that the yolov5l6 model performance can be improved by using images with higher magnification levels. The highest overall performance is achieved by 6x zoom images whereas no zoom gives the worst performance. The precision for 6x zoom is 0.846 compared to no zoom which is 0.572 while the recall for 6x zoom is 0.543 and the recall for no zoom is 0.258. Similarly, the highest mAP50 and mAP50:95 is obtained when using 6x zoom that is 0.591 and 0.428 respectively. Additionally, Figure 9 shows the samples of prediction made by yolov5l6 model to detect the location of subject mannequin. Successful detections are shown in red bounding box, whereas misdetection is highlighted in blue bounding box.

According to Figure 9, we can observe that there are several instances of false negatives where mannequins are not detected. There are several possible reasons for this including obstructions and occlusions, deformed mannequins shape, viewing angle and background contrast. However, since the model also works with video, the detection would continuously run and the drone flies around the area, there is also a possibility that the undetected mannequins get detected in the subsequent frames. Besides that, it is apparent from the sample images, there are no false positives predicted. For readers who are interested to see the prediction applied on a real-time video, sample of inference video using yolov5l6 model applied one video from Zigzag-1 dataset with 2x zoom magnification can be accessed directly from this online video.

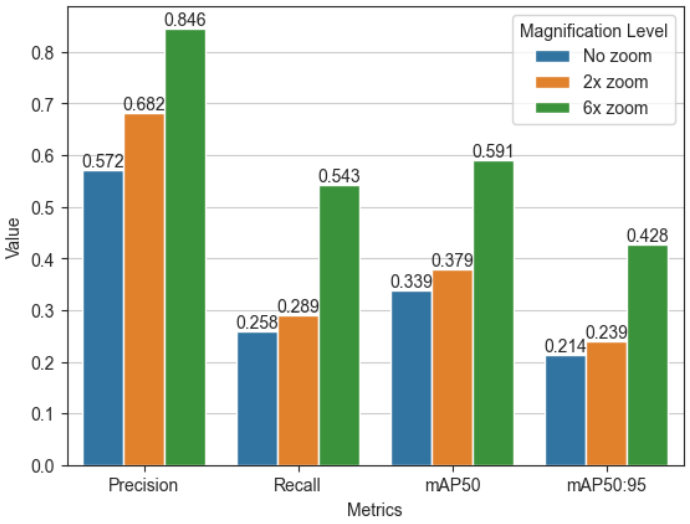


Fig. 8. Detection performance of yolov5l6 on Grid-2 test dataset using different magnification levels

Fig. 9. Sample of images showing the successful detection by yolov5l highlighted by red boxes. Blue boxes on the other hand show several instances of misdetections (false negatives)

4. Conclusions

The work presented here proposed a deep learning method called You Only Look Once version 5, or YOLOv5 as a human detector which can be used to assist in search and rescue operations, by providing drones with automated human detection capability to help locate victims quicker. Based on experiments using data collected during a simulation of search and rescue operations, a variant of YOLOv5 model called yolov5l6, produced the best performance. This model delivered overall test precision, recall and mAP50 rate at 0.668, 0.303 and 0.346 respectively. We have also shown that using images acquired from video with 6x zoom magnification can significantly improve the model

performance. The model performance was improved such that precision, recall, and mAP50 rate were increased to 0.846, 0.543, and 0.591 respectively. Another important factor to consider for a successful application of the model is the inference time. We showed that the best model has 43ms inference time per image which is sufficient to allow the video to run at 23fps. From the experiment, we can also conclude that there is still a lot of room for improvement if we were to get higher detection performance, with less false detections and miss detections. For a critical mission such as search and rescue, higher performance should be expected to ensure that the victims can be located and rescued immediately. Thus, for future work we are looking into other approaches such as vision transformer and ensembles of detectors to improve the overall performance.

Acknowledgement
We would like to extend our gratitude to Universiti Teknologi Mara (UiTM) and to those who have directly or indirectly contributed to our research. This work was supported by the Trans-Disciplinary Research Grant Scheme (600-IRMI/TRGS 5/3 (001/2019)-1), Ministry of Higher Education Malaysia. We would like to thank Institut Perubatan Forensik Negara (IPFN), Aerodyne Group and Angkatan Pertahanan Awam Malaysia (APM) which have involved in the disaster rescue simulation and data collection.

References

[1] Daud, Sharifah Mastura Syed Mohd, Mohd Yusmiaidil Putera Mohd Yusof, Chong Chin Heo, Lay See Khoo, Mansharan Kaur Chainchel Singh, Mohd Shah Mahmood, and Hapizah Nawawi. "Applications of drone in disaster management: A scoping review." *Science & Justice* 62, no. 1 (2022): 30-42. https://doi.org/10.1016/j.scijus.2021.11.002

[2] Yusof, Ahmad Anas, Mohd Khairi Mohamed Nor, Nur Latif Azyze Mohd Shaari Azyze, Anuar Mohamed Kassim, Shamsul Anuar Shamsudin, Hamdan Sulaiman, and Mohd Aswad Hanafi. "Land Clearing, Preparation and Drone Monitoring using Red-Green-Blue (RGB) and Thermal Imagery for Smart Durian Orchard Management Project." *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences* 91, no. 1 (2022): 115-128. https://doi.org/10.37934/arfmts.91.1.115128

[3] Kanellakis, Christoforos, and George Nikolakopoulos. "Survey on computer vision for UAVs: Current developments and trends." *Journal of Intelligent & Robotic Systems* 87 (2017): 141-168. https://doi.org/10.1007/s10846-017-0483-z

[4] Oh, Donggeun, and Junghee Han. "Smart Search System of Autonomous Flight UAVs for Disaster Rescue." *Sensors* 21, no. 20 (2021): 6810. https://doi.org/10.3390/s21206810

[5] Noviarini, Diena, Mutia Delina, Ananda Mochammad Rizky, Umi Widyastuti, Osly Usman, and Akhmad Yamani. "Early Warning System for Fire Catcher in Rain Forest of Sumatera Using Thermal Spots." *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences* 103, no. 1 (2023): 30-39. https://doi.org/10.37934/arfmts.103.1.3039

[6] Al-Naji, Ali, Asanka G. Perera, Saleem Latteef Mohammed, and Javaan Chahl. "Life signs detector using a drone in disaster zones." *Remote Sensing* 11, no. 20 (2019): 2441. https://doi.org/10.3390/rs11202441

[7] Mishra, Balmukund, Deepak Garg, Pratik Narang, and Vipul Mishra. "Drone-surveillance for search and rescue in natural disaster." *Computer Communications* 156 (2020): 1-10. https://doi.org/10.1016/j.comcom.2020.03.012

[8] Baeck, P. J., N. Lewyckyj, B. Beusen, W. Horsten, and K. Pauly. "Drone based near real-time human detection with geographic localization." *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 42 (2019): 49-53. https://doi.org/10.5194/isprs-archives-XLII-3-W8-49-2019

[9] Lee, Seoungjun, Dongsoo Har, and Dongsuk Kum. "Drone-assisted disaster management: Finding victims via infrared camera and lidar sensor fusion." In *2016 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, pp. 84-89. IEEE, 2016. https://doi.org/10.1109/APWC-on-CSE.2016.025

[10] Rizk, Mostafa, Fatima Slim, and Jamal Charara. "Toward AI-assisted UAV for human detection in search and rescue missions." In *2021 International Conference on Decision Aid Sciences and Application (DASA)*, pp. 781-786. IEEE, 2021. https://doi.org/10.1109/DASA53625.2021.9682412

[11] Liu, Wei, Karoll Quijano, and Melba M. Crawford. "YOLOv5-Tassel: detecting tassels in RGB UAV imagery with improved YOLOv5 based on transfer learning." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022): 8085-8094. https://doi.org/10.1109/JSTARS.2022.3206399

[12] Ding, Kaiwen, Xianjiang Li, Weijie Guo, and Liaoni Wu. "Improved object detection algorithm for drone-captured dataset based on yolov5." In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pp. 895-899. IEEE, 2022. https://doi.org/10.1109/ICCECE54139.2022.9712813

[13] Zhanying, Zhang, and Chen Xinyuan. "Research on Forest Fire Detection Algorithm Based on Yolov5." In *2021 International Conference on Intelligent Computing, Automation and Systems (ICICAS)*, pp. 354-357. IEEE, 2021. https://doi.org/10.1109/ICICAS53977.2021.00080

[14] Sruthi, M. S., Manal Jaleel Poovathingal, V. N. Nandana, S. Lakshmi, Mohamed Samshad, and V. S. Sudeesh. "YOLOv5 based Open-Source UAV for Human Detection during Search And Rescue (SAR)." In *2021 International Conference on Advances in Computing and Communications (ICACC)*, pp. 1-6. IEEE, 2021. https://doi.org/10.1109/ICACC-202152719.2021.9708269

[15] Lagman, Jewel Kate D., Alden B. Evangelista, and Charmaine C. Paglinawan. "Unmanned Aerial Vehicle with Human Detection and People Counter Using YOLO v5 and Thermal Camera for Search Operations." In *2022 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pp. 113-118. IEEE, 2022. https://doi.org/10.1109/I2CACIS54679.2022.9815490

[16] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788. 2016. https://doi.org/10.1109/CVPR.2016.91

[17] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263-7271. 2017. https://doi.org/10.1109/CVPR.2017.690

[18] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." *arXiv preprint arXiv:1804.02767* (2018).

[19] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934* (2020).

[20] Padilla, Rafael, Wesley L. Passos, Thadeu LB Dias, Sergio L. Netto, and Eduardo AB Da Silva. "A comparative analysis of object detection metrics with a companion open-source toolkit." *Electronics* 10, no. 3 (2021): 279. https://doi.org/10.3390/electronics10030279

[21] Everingham, Mark, L. Van Gool, and Williams KI. "C., Winn, J., & Zisserman, A.(2010)." *The PASCAL Visual Object Classes (VOC) Challenge*: 303-338. https://doi.org/10.1007/s11263-009-0275-4

[22] Huang, Jonathan, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer et al. "Speed/accuracy trade-offs for modern convolutional object detectors. arXiv 2017." *arXiv preprint arXiv:1611.10012* (2012). https://doi.org/10.1109/CVPR.2017.351

[23] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems* 28 (2015).

[24] Bell, Sean, C. Lawrence Zitnick, Kavita Bala, and Ross Girshick. "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2874-2883. 2016. https://doi.org/10.1109/CVPR.2016.314