Illumination-Aware Image Fusion for Around-The-Clock Human Detection in Adverse Environments from Unmanned Aerial Vehicle

Gelayol Golcarenarenji[a,*], Ignacio Martinez-Alpiste[a], Qi Wang[a] and Jose Maria Alcaraz-Calero[a]

[a]*School of Computing, Engineering and Physical Sciences, University of the West of Scotland, United Kingdom*
**Corresponding author*
g.golcarenarenji@uws.ac.uk (Gelayol Golcarenarenji)
ignacio.alpiste@uws.ac.uk (Ignacio Martinez-Alpiste)
qi.wang@uws.ac.uk (Qi Wang)
jose.alcaraz-calero@uws.ac.uk (Jose Maria Alcaraz-Calero)

ARTICLE INFO

Keywords:
Human detection
UAV
Deep machine learning
Image fusion
Image registration

ABSTRACT

This study proposes a novel illumination-aware image fusion technique and a Convolutional Neural Network (CNN) called BlendNet to significantly enhance the robustness and real-time performance of small human objects detection from Unmanned Aerial Vehicles (UAVs) in harsh and adverse operation environments. The proposed solution is particular useful for mission-critical public safety applications such as search and rescue operations in rural areas. The operation environments of such missions are featured with poor illumination condition and complex background such as dense vegetation and undergrowth in diverse weather conditions, and the missions have to address the challenges of detecting humans from UAVs at high altitudes, with a moving platform and from various viewing angles. To overcome these challenges, the proposed solution register and fuse the images using Enhanced Correlation Coefficient (ECC) and arithmetic image addition with customized weights techniques. The result of this fusion is fuelled with our new BlendNet AI model achieving 95.01 % of accuracy with 42.2 Frames Per Second (FPS) on Titan X GPU with input size of 608 pixels. The effectiveness of the proposed fusion method has been evaluated and compared with other methods using the KAIST public dataset. The experimental results show competitive performance of BlendNet in terms of both visual quality as well as quantitative assessment of high detection accuracy at high speed.

1. Introduction

Advanced human detection is highly demanded in many mission-critical application fields such as search and rescue missions (Martinez-Alpiste et al., 2020b, 2021), surveillance and intruder detection systems, autonomous driving and so on. When coupled with Unmanned Aerial Vehicles (UAVs), developing a robust human detector becomes even more challenging specially to achieve high-accuracy and high-speed detection from UAVs that fly at far distances for efficient operations (Yu et al., 2020). Furthermore, such applications require high robustness in diverse and adverse operation environments, such as unstable ambient illumination conditions, high altitudes and diversity in backgrounds, viewing angles, human poses and clothing (Rudol and Doherty, 2008).

From the different sensors that UAVs implement, the most commonly used are thermal and optical cameras. Most of the existing human detectors use optical images with good lighting conditions, and thus they do not perform well with images captured under low visibility conditions, in bad weather or at night (Guan et al., 2019). Moreover, solely relying on the optical imagery leads to performance issues in low contrast scenarios such as bright clothing detection. In addition, human detection is not feasible for lands covered intensely with undergrowth, forest or dense foliage with the existing solutions. Therefore, using only optical imagery is insufficient for all-weather and around-the-clock human detection with the aforementioned adverse conditions (Liu et al., 2016a).

Yet thermal imagery can improve the accuracy of the human detection in around-the-clock applications. The atmospheric conditions will not effect the thermal imaging and the hidden heat source targets can be distinguished using these images. Moreover, the human can be differentiated from hard negative samples such as trees, poles and telephone booth using thermal imagery (Pei et al., 2020). However, although more robust in these conditions, thermal cameras have their own weaknesses in situations of lower resolution of imagery and when the temperature signature of the object is similar to its surroundings (Dawdi et al., 2020). As a case in point, a hot sunny day will yield a lot of hot areas on the entire thermal image, which in turn will hinder an effective detection.

The above analysis can conclude that fusion of both optical and thermal spectra would make the detection systems more robust under varying lighting conditions, both day and night, in forests and dense vegetation. The aim of thermal and optical imagery is to combine different sets of information from two images which contain abundant detailed information of optical images and effective target areas of thermal images. Meanwhile, integration of the information from both sources is a challenging task. Moreover, switching between the different imagery modes, optical only, thermal only as well as fusion, should be enabled to increase the flexibility of a practical integrated solution. In a nutshell, the environment where a UAV platform is located is usually complex and changing, which leads to a big challenge in optimising solutions. For UAV-based applications such as search and rescue missions, the major challenge is a trade-off between detection speed and accuracy. It requires the proposed technique to be highly accurate which is critical when saving human lives but it also needs to be of small size and light weight. Thermal and optical image fusion is a beneficial processing task for UAV surveillance, which can improve the visibility by combining the advantages of the thermal and optical imaging. This will increase the detection rate where difficult conditions such as high altitude, moving

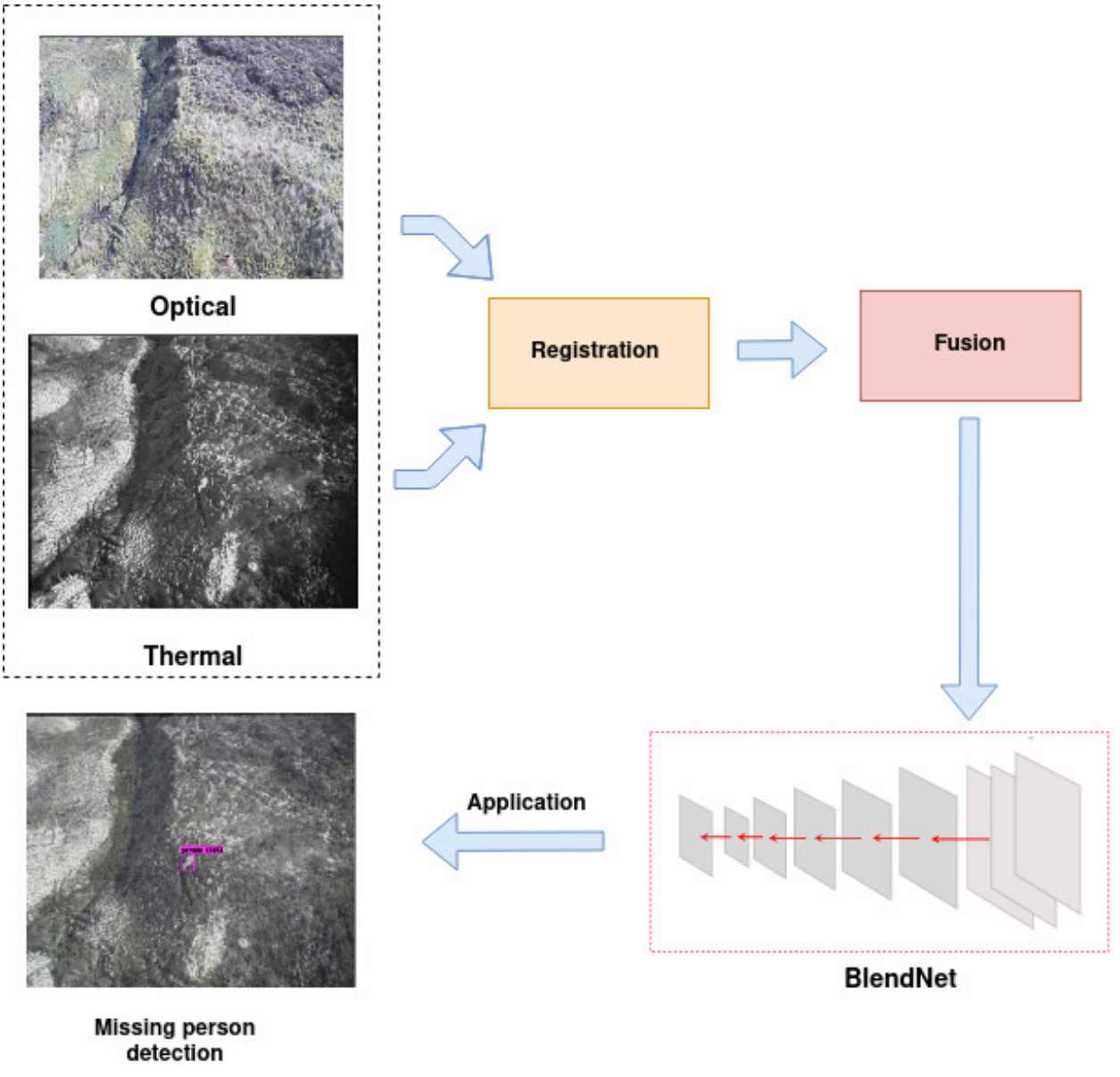


Figure 1: Overview of methodology.

platform and viewing angles make human detection challenging. Hence, this research proposes an effective fusion technique with flexible mode switching, with a new end-to-end real-time machine-learning-based human detection system empowered by a novel convolutional neural network (CNN) algorithm called BlendNet. The propose solution is able to detect people around-the-clock with high accuracy and high speed while addressing the challenges involved in human detection from UAVs. Figure 1 presents a high-level overview of the proposed Illumination-Aware Image Fusion technique for this accurate and real-time human detection system. Based on the figure, the thermal and optical images are blended after registration and go through BlendNet to detect the small human object. The application of the developed integrated system has been validated in real scenarios by Police Scotland in search and rescue operation trials.

As a result, our contributions are summarised as follows:

- Design and development of a new three-step customised fusion method for minimal fused imagery distortion by utilising and leveraging the relationship between optical and thermal images.
- Design and development of a new CNN to create a new real-time machine learning based small human object detection algorithm (BlendNet) considering all the challenges of human detection from UAVs.
- Training and validation of the proposed fusion method and the BlendNet algorithm based on real-world trials undertaken by Police Scotland.
- Allowing flexible operation capabilities such as optical only, thermal only or optical-thermal fusion modes in response to different illumination conditions.
- Both qualitative and quantitative testing evaluation empirically performed to show high-speed real-time human detection with high accuracy, in comparison with the-state-of-the-art solutions.

2. Related Work

This section focuses on the description of the techniques used in this paper. In addition, it reviews state-of-the-art work related to fusion techniques and human detection.

2.1. Object Detection Techniques

Several methods have been employed for the purpose of object detection in the last decade. For instance, traditional object detection methods were used for handcraft feature extraction (Dollar et al., 2009; Surasak et al., 2018; Felzenszwalb et al., 2010; Teutsch and Krüger, 2012). CNN-based object detection methods have also been used. Two major categories used for CNN-based detectors are two-stage and one-stage detectors. The common two-stage object detectors are Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), and R-FCN (Dai et al., 2016). In these techniques, the classification occurs in the second stage after the Regions Of Interest (ROIs) are extracted in the first stage. These methods are accurate but computationally expensive. They are not suitable for constrained devices and real-time object detection. To overcome these shortcomings, "You Only Look Once" (YOLO) series (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018), "You Look Only Twice" (Van Etten, 2018) and "Single Shot Detector (SSD)" (Liu et al., 2016b) are used as one-stage detectors. In these methods, the ROI and the classification are performed all at once (Jiang and Wang, 2016). A modified YOLOV4 was developed to fulfill the requirements of the study.

2.2. Path Aggregation Network (PAN)

Most object detection techniques are mainly focused on medium-to-large sized objects. Small object detection is significantly more challenging as the feature extraction of small objects with sometimes being just few pixels becomes much more difficult and the features get vanished during the down-sampling process in deeper layers of CNN. Furthermore, large datasets for small objects are scarcely available in the public domain (Krishna and Jawahar, 2017). To improve the accuracy of small object detection, additional techniques are needed. Inspired by the Feature Pyramid Network (FPN) (Lin et al., 2017), the accuracy of small objects is improved by using the PAN architecture (Golcarenarenji et al., 2021, 2022; Liu et al., 2018b). This is a method that improves the accuracy of small object detection by reducing the data path between the initial layers and deeper layers. This is realised by including an extra bottom-up path augmentation to combine features from top layers with more information, and the deeper layers with more meaningful information as both information is needed. An extra up-sampling was added to the PAN architecture used in the proposed architecture to keep more shallow features that are essential for successful small object detection.

2.3. Receptive Field Block (RFB)

The RBF module is a multi-branch pooling used with different kernels and applies dilated convolution layers to improve the deep features learned to improve speed and accuracy of object detectors (Liu et al., 2018a). In the proposed architecture the hybrid dilation rates were modified to 1,2, and 3 which resulted in a better improvement. The 5×5 convolution layer was also replaced by 3×3 convolutional layers to reduce parameters.

Table 1
Comparison of fusion-based human detection solutions

Ref	Algorithm	Exec Env.	Platform	Fusion	Input size	Accuracy (%)	Speed (FPS)	Model Size (MB)	Altitude (m)
Dawdi et al. (2020)	Canny Edge/ORB	RaspPi	OpenCV	Arithmetic addition	152× 197	All victims	0.55	NG	10
Song et al. (2021)	MFMFN (Improved-YOLOv3)	PC	NG	Feature maps fusion	416×416	85	56	NG	N/A
Vandersteegen et al. (2018)	YOLOv2	PC	Darknet	Channel Composition	640×512	31.2*	80	NG	N/A
Pei et al. (2020)	F-RetinaNet	PC	NG	Early/Res/ FPN	640×768	27.60*	0.13	NG	N/A
Fu et al. (2021)	ASPFF	PC	Pytorch	Pixel-level	NG	15.43*	25(CPU)/ 35(GPU)	NG	N/A
Cao et al. (2019)	DCNN/VGGnet	PC	NG	TS-RPN	NG	95.5**	4.5	25.2	N/A
Cao et al. (2021)	YOLOv4	PC	NG	Half-way	640×640	4.91/23.14*	32.3	NG	N/A
Li et al. (2019)	Fast RCNN	PC	NG	Half-way	56×56	15.73*	4.8	NG	N/A
Xue et al. (2021)	MAF-YOLO	PC	NG	Modal weighted	416×416	87.8	40	NG	N/A
Wu et al. (2022)	Modified Faster RCNN	PC	NG	GAN	NG	95.36	NG	NG	N/A
Li et al. (2021)	SSD	Nvidia TX1/ Zedboard	Opencv	Thermal target/texture feature map	640×480	92.6	36.6/205.3	NG	NG
TP	BlendNet	PC	Darknet	ECC	608×608	95.01	42.2	42.6	up to 75

TP = This Paper; NG = Not Given; N/A=Not applicable; Green = information provided; Red = information missed
*=miss rate; **=Recall

2.4. Image registration

Before Image fusion, the camera should be calibrated to provide intrinsic camera parameters and distortion coefficients to correct image distortion (Dandrifosse et al., 2021). OpenCV library was used to calibrate the camera. Image registration is another process to align and match two or more images from different cameras. A good optical and thermal image fusion method should be able to keep the thermal radiation information in thermal images and the texture detail information in optical images (Piao et al., 2019). The registration of optical and thermal images is a vital preliminary step for image fusion, object detection and tracking, and remote sensing to eliminate the offset between images (Yu et al., 2019; Ding et al., 2021; Dandrifosse et al., 2021). Although many studies exist for optical and thermal image registration, studies for UAV-based platforms are still rare (Meng et al., 2021). The registration of visible and infrared images has certain complexities due to different resolutions, field of view and spatial position (Ding et al., 2021). An optimised Enhanced Correlation Coefficient (ECC) method (Evangelidis and Psarakis, 2008), was designed due to being highly effective and fast for registration of heterogeneous images. ECC is a gradient-based image registration algorithm invariant to global illumination changes (Raudonis et al., 2021; Choi et al., 2017; López et al., 2021; Hwooi and Sabri, 2017). The ECC algorithm was optimised in terms of iterations and precision factor to fulfil the requirements of the use case.

2.5. Fusion-based Human Detection

This subsection highlights representative published results related to fusion-based human detection, in comparison with the proposed solution in this paper. The comparative analysis is summarised in Table 1.

Cao et al. (2021) proposed a fusion architecture based on YOLOv4 for multi-spectral human detection and a novel attention fusion method. Multispectral channel feature fusion (MCFF) was introduced to fuse the thermal and optical streams based on illumination conditions with half-way fusion being the best fusion option. However, although accurate, the detection is not from UAVs.

In another study, Li et al. (2019) proposed an illumination-aware faster Regions with Convolutional Neural Network (R-CNN) for robust multi-spectral pedestrian detection taking illumination conditions into consideration. Similarly, the detection is not from UAVs and computationally expensive for our use-case. Dawdi et al. (2020) developed an autonomous system for victim detection using Canny edge detection template matching on UAVs. The detection was not tested from an altitude of more than 10m and is slow for our use case. Song et al. (2021) implemented a robust Multi-Spectral Feature Fusion Network (MSFFN) for pedestrian detection, which fully integrated the features extracted from visible light and infrared channels using improved YOLOv3. Although accurate, the detection was not tested

from UAVs and not fast enough for our use case. Vandersteegen et al. (2018) developed a real-time multi-spectral pedestrian detector using single-pass CNNs. Although the accuracy is close to the performance of other state-of-the-art multi-spectral CNNs with a log-average miss-rate of 31.2% measured on the KAIST dataset, the detection was not tested from UAVs. Pei et al. (2020) designed a three Deep Convolutional Neural Network (DCNN) fusion architectures were designed and the sum fusion strategy showed best performance for their detector. Their study is adaptable to the around-the-clock applications. However, it is computationally expensive for our use case and it is not tested for UAVs. Fu et al. (2021) introduced a light end-to-end dual-modality multi-scale human detection framework, which can achieve real-time detection speed using an adaptive spatial pixel-level feature fusion (ASPFF) Network. In another study, a unified framework was proposed by Cao et al. (2019) which combined the auto-annotation method with a two-stream region proposal network (TS-RPN) detector to learn the semantic features of thermal and optical images to achieve unsupervised learning of multi-spectral features for human detection. Both studies have not been tested from UAVs. Xue et al. (2021) proposed a novel approach denoted as Multi-modal attention fusion based YOLO (MAF-YOLO) for multi-modal pedestrian detection based on YOLOv3. Although this study is very accurate, it is not fast enough for our use-case and not from UAVs. In another study, Li et al. (2021) extracted the thermal target in an infrared image and combines it with the background of the scene. The performance was tested using the SSD-based target detection algorithm. The results were implemented on Nvidia TX1 and the NVIDIA Jetson TX1 and the ZedBoard (FPGA). Although very accurate, it was not mentioned from which altitudes, this accuracy was obtained due to the dataset being private. In another study, Wu et al. (2022) used GAN and a modified fast RCNN for human detection. Although the algorithm achieved very good detection results for information from both modalities (95.36%), the model used is computationally expensive (modified Faster RCNN) and thus not suitable for our use case where a trade-off between speed and accuracy is essential. To sum up, current literature does not look deeply at the challenges of detecting humans from UAVs due to difficult conditions such as high altitude, moving platform and viewing angles. They typically rely on computationally expensive models whereas this proposed solution in our study minimises the computational power to make it usable for a portable, inexpensive, resource-constrained devices with capability of 24/7 human detection in woodlands. To overcome the shortcomings of the existing systems, this study is proposed to combine thermal and optical sensors and implement a novel CNN-based model, to significantly improve the robustness and usefulness of human objects detection in harsh and adverse operation environments, potentially at 24 hours for life-saving scenarios. The target environments are featured with poor illumination condition and complex background such as dense vegetation and undergrowth in diverse weather conditions. In addition, large geometrical distortions are observed in UAV images, making the registration of optical and thermal images difficult. Our proposed three-Step Customised Low-Distortion Image Fusion of Optical and Thermal Imagery have reduced this by optimising simple yet highly effective, and fast method for calibration registration and fusion of heterogeneous images. To sum up, the first main contribution is design and development of a three-Step Customised Low-Distortion Image Fusion of Optical and Thermal Imagery. In order not to introduce overhead to our UAV-based object detection system to be suitable for resource-constrained devices, we had to keep it as simple as possible. Hence, after many trials, the Enhanced Correlation Coefficient (ECC) method was optimised for registration of heterogeneous images. This method has been optimised using grid-search in terms of iterations, precision factor, and motion model to fulfil the requirements of the use case. The images then fused using arithmetic image addition with customised weight. The weights have varied for both optical and thermal components of the fused image. Moreover, an additional training was carried out where the selection of the weights was not statically selected and chosen instead according to metadata named as Dynamic Metadata-based Weight to gain the best performance. This weight variation allowed best optical-thermal fusion modes in response to different illumination conditions. The second contribution is a design and development of a real-time highly accurate small human object detector algorithm from UAV using infrared and visible image fusion in a real UAV scenario considering all the challenges involved in UAV-based scenario. To fulfil this, a novel CNN-based (Tiny YOLOv4 +modified RBF+ modified PAN) was created. The rest of the paper is organised as follows. Section 2 presents the related work. Section 3 describes the design of the proposed solution to detect humans, followed by the experimental setup in Section 4. Section 5 discusses the results of the proposed solution. Section 6 concludes the paper.

3. Proposed Solution

In this section, the use case overview and requirements which are high accuracy, high speed, and portability, the proposed Image-fusion method and the design of the CNN-based model are all explained in details.

3.1. Use case overview and Requirements

To achieve high-accuracy, high-speed, and portability which is imperative for the success of primary Search and rescue missions used by police of Scotland and other applicable use cases such as intruder detection from UAVs in the EU Horizon 2020 5G-PPP 5G-INDUCE project and Angel Drone in EU Horizon 2020 ARCADIAN-IoT project, speed of FPS $\geq$24, accuracy of more than 90% (to find missing people and save lives), and resource efficiency for portability of the solutions on constrained devices such as smartphone, and tablets (Model size$\leq$ 54 MB and BFLOPS $\leq$ 28) are required (Golcarenarenji et al., 2021; Martinez-Alpiste et al., 2020c; Martinez-Alpiste et al., 2019; Martinez-Alpiste et al., 2020a). Hence, the following steps are proposed solutions to achieve the requirements of these projects.

3.2. Proposed Three-Step Customised Low-Distortion Image Fusion of Optical and Thermal Imagery

High-quality imagery fusion requires that two images are blended with minimal distortion, or displacements. To accurately fuse two images, camera calibration and Image registration must be performed. To this end, a three-step image fusion scheme is proposed as follows.

In the first step, camera calibration is carried out. This is done to avoid the introduced radial and tangential distortion to images by the optical and thermal cameras. The images were calibrated using 30 photos taken of a 9×7 chessboard (30mm squares) for each camera. The chessboard was built using a mirror and black velvet for both thermal and optical cameras. Figure 2 shows the chessboard photos taken simultaneously with both thermal and optical cameras.

In the second step, image registration is performed. RGB imagery typically have higher resolution than thermal imagery; therefore, it is resized to match the size of the thermal one. For performing a successful registration, three transformation should be considered: translation, rotation and scale, for different perspective views. The translation and rotation are supposed to fix the differences between RGB and thermal imagery and the scale searches for an appropriate size where both images get an accurate overlapping.

The translation and rotation offset are minimised thanks to an image registration process named as ECC which is used as a gradient-based algorithm and recommended for registration of heterogeneous images (López et al., 2021). Two main motion models were considered to be used in this method: Affine and Homography, from lower to higher computational complexity. The model stays linear ($O(n)$) even with the most complex algorithm. The Affine motion model includes translation, rotation and scale transformations and proved to give the best results with an warp matrix of 2×3 instead of 3×3 against the Homography method. This is obtained by varying iterations and precision factor of the ECC algorithm using the grid search. The implementation and optimisation of the ECC is done in Python. Figure 2 shows the schematic of the proposed image registration method where the Affine motion model is applied. In terms of the motion model for fusion, the number of iterations, and the threshold were tuned using the grid search. The number of iterations changed between 5 and 5,000 by increments of 5. The threshold of the increment in the correlation coefficient between two iterations was changed between 1×10^{-10} and 1×10^{-50} by the increments of 1×10^{-10}. The best results obtained with the increment threshold of 1×10^{-50} and 5,000 iterations. In the third and last step, the fusion of imaged is performed. An arithmetic image addition with customised weights was selected as the fusion technique to keep the most important features of one image and enhance the less important feature of another image to have for maximum control. Figure 3 illustrates the results of using image registration and customised weight fusion technique. As can be seen on the first and the last row of the figure, the persons are not visible due to being behind the trees. However, they can be seen on thermal images and with more details on fused images. The second row from the top shows an image taken at night with person not being visible. However, the person is visible in the thermal image and with more details in the fused image. In the third row from the top, the person is visible in the optical and fused images but not in the thermal image due to the temperature signature being similar to its surroundings. To sum up, in order not to introduce overhead to our UAV-based object detection system, we optimised the ECC method as a simple yet highly effective, and fast method for registration of heterogeneous images. This method was optimised using grid-search in terms of iterations, precision factor, and motion model to fulfil the requirements of the use case. The images were then fused using arithmetic image addition with customised weight.

3.3. Proposed New CNN-Based BlendNet Algorithm for Fast and Accurate Small Object Detection

The proposed BlendNet comprises three primary components: a Backbone to create a base lightweight object detection algorithm for small object detection, a Receptive Field Block to strengthen the lightweight features for improved cost-efficiency of accuracy and speed, and a PAN to enhance the process of instance segmentation for higher accuracy for small object detection.

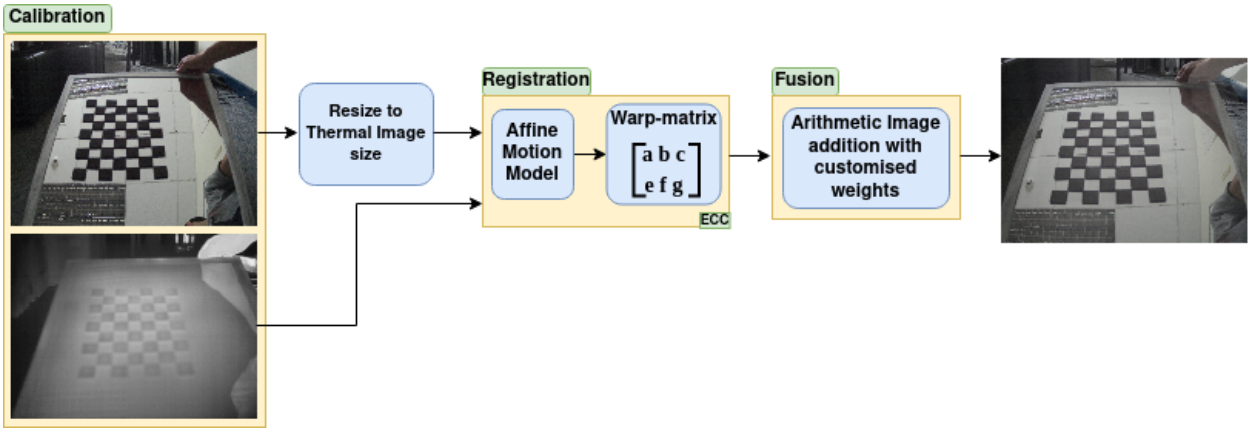


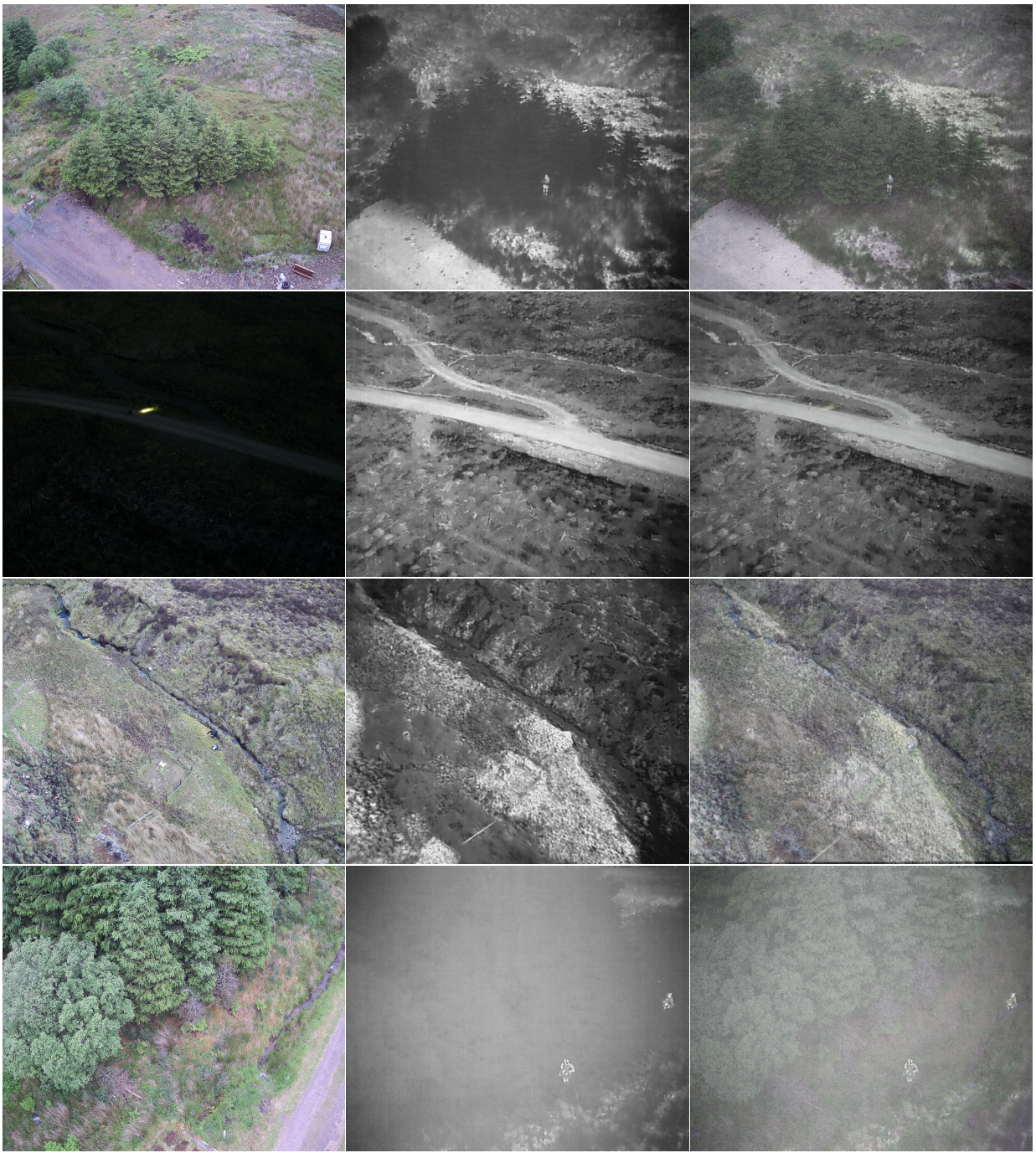
Figure 2: Image fusion process composed of three stages: calibration, registration and fusion.

As shown in Figure 4, firstly, the Backbone starts with a 3×3 convolution with the number of filter being 32 with stride 2. The Tiny-YOLOv4 (AlexeyAB, 2020a) was used as the backbone of the proposed method. The Tiny-YOLOv uses the CSPBlock module in cross stage partial network (Wang et al., 2020) instead of the ResBlock module in residual network (labelled as a). The feature map in the CSPBlock module is divided into two parts and cross stage residual edge is used to combine the two parts. This increase the correlation difference of gradient information and can enhance the learning ability of convolution network when compared with ResBlock module. Tiny-YOLOv4 uses three CSPblock modules with the number of filters being 64, 128 and 256, respectively. The first module starts with the filter number being 64 and divides into two convolutions with filters of 32 and then combined with the first convolution with filter 32. A 3×3 convolution with filter 64 was run on the result. The result is again combined with the shortcut from the first convolution with filter 64. The second module starts with the filter number being 128 and divides into two convolutions with filters of 64 and then combined with a shortcut from the first convolution with filter 64. A 3×3 convolution with filter 128 was run on the result. The result is again combined with the shortcut from the first convolution with filter 128. Similarly, the last module starts with the filter number being 256 and divides into two convolutions with filters of 128 and then combined with a shortcut from the convolution with filter 128. A 3×3 convolution with filter 256 was run on the result. The result is again combined with the shortcut from the first convolution with filter 256.

Secondly, Receptive field block (RFB) was utilised (labelled as b) at the end of the backbone. In the first branch from the top, a 1×1 convolutional layer was used to decrease the number of channels in the feature map accompanied by one 3×3 convolution with dilation rate of 1. Similarly, in the second branch, a 1×1 convolutional layer was used followed by two 3×3 convolutions. The 5×5 convolution layer in (Liu et al., 2018a) was replaced by 3×3 convolutional layer to reduce the parameters. The dilution rate was selected two in this branch. Lastly, the third branch comprises one 1×1 convolutional layer followed by two 3×3 convolutions with dilation rate of 3. This selection of dilation rates resulted in a better improvement in our study.

Finally, a modified PAN module was also added (labelled as c) in the algorithm with three up-samplings (compared to YOLOv4) and one top-down pathway. The number of filters selected was 128, 256, 128 and 64, respectively. The extra bottom-up path augmentation was down-sampled (4 down-sampling compared to two down-sampling in YOLOv4) with factors of 16, 8, 4 and 2, respectively. The number of filters selected was 32. The features from the bottom-up path were then concatenated and 1×1 convolution was run on the result. The YOLOv3 headers were used to outputs the coordinates, probability and level of confidence. The feature map of three detection heads were 19×19, 38×38 and 76×76.

The combination and integration of Tiny-YOLOv4, RFB-module and PAN created the BlendNet. To summarise, to develop a real-time highly accurate small human object detector algorithm from UAV using infrared and visible image fusion in a real UAV scenario considering all the challenges involved, a modified RFB module was used to improve the accuracy without incurring too much computational burden by increasing the receptive field of the backbone in our approach. To increase the accuracy of small humans at high altitudes which is crucial for this use case, a modified PAN module was implemented and added into the architecture. The PAN architecture was modified by adding an extra up-sample (3 up-samplings) compared to that of YOLOv4 (2 up-samplings) to keep more shallow features which is



(a) Optical (b) Thermal (c) Fusion

Figure 3: Fusion results.

essential for small object detection. The extra bottom-up path augmentation was gradually down-sampled (4 down-samplings) compared to the 2 down-samplings in YOLOv4. Finally, to improve object detection speed, Tiny-YOLOv4 were proposed in the backbone.

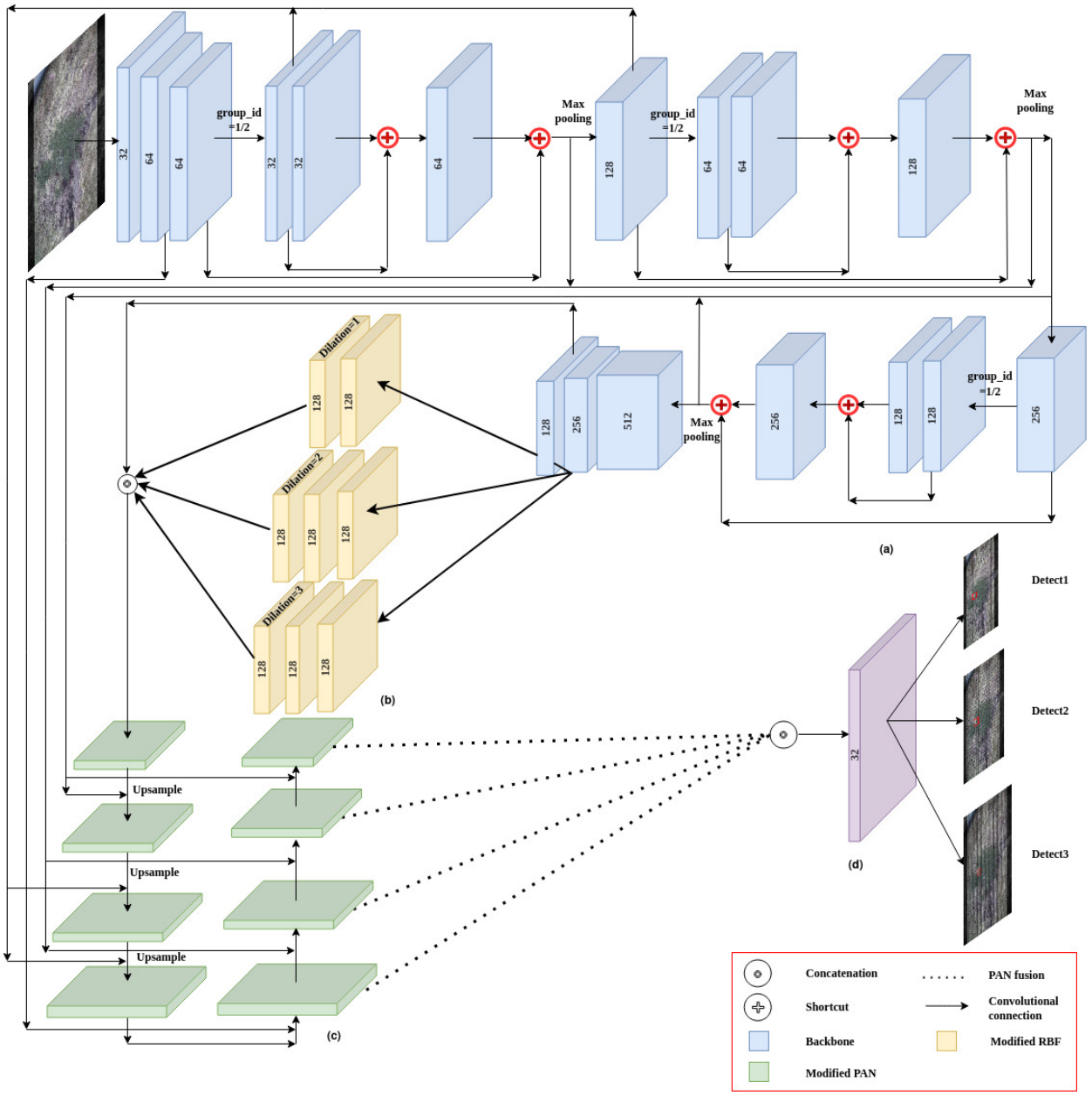


Figure 4: BlendNet architecture composed of modified RFB-module, modified PANet and Tiny-YOLOv4.

4. Experimental Setup

This section is divided into three subsections. First, the real-world trials are explained for dataset creation. Second, the hyper-parameters used in the study are discussed. Finally, the execution environment is introduced.

4.1. Dataset Creation Based on Real-world Trials

4K live videos were recorded and still images were extracted from the collected videos for further processing. The UAV-based experimental trials were performed by Police Scotland during the day and night in cloudy and sunny weather conditions at various UAV altitudes from 5 to 75 meters. The Police Scotland crew (men and women) comprised two teams, one operating the system whilst the other hiding in the wild to act as missing people. The

Table 2
Execution Hyperparameters

Hyperparameters	Values
Image size in pixel	608×608
Number of iteration	10000
Batch size	64
Initial learning rate	0.008
Solver	SGDR
Momentum coefficient	0.9
Weight decay	0.001

participating people in the second team were wearing clothing in red, green, white, yellow, pink, camouflage and chartreuse. The data were collected at different operational locations in Scotland wilderness with various background. Different poses, figures, postures, scales, angles, orientations, sizes and altitudes were taken into account when collecting the dataset.

A thermal camera (DJI Zenmuse XT2) was mounted on board of the UAV recording videos with the input size of 640×512 at 30 Frames Per Second (FPS). The optical camera recorded videos with the input size of 3840×2160 at 30 FPS. Totally, 10,000 image pairs (thermal and optical) with 5,232 positive and 4,768 negative images were extracted and manually annotated to create the training dataset.

YOLO-Mark2 (AlexeyAB, 2020b) was used as the labelling tool for the dataset. To generate similar number of positive and negative images, synthetic images were created using copy-paste strategy (Dwibedi et al., 2017) and data augmentation techniques such as flipping, rotation, blurring, Gaussian noise (Jung et al., 2020; Perez and Wang, 2017).

4.2. Hyper-parameters

The hyper-parameters used in this study are specified in this subsection. To recalculate the anchor boxes for the dataset, the K-means technique was used with the input size of (608×608 pixels) and 9 anchors boxes (AlexeyAB, 2020a). To have a true comparison of the proposed algorithm with existing state-of-the-art algorithms, the same values of the hyper-parameters were used for all the algorithms. The number of training iterations was set to be 10,000. The Stochastic Gradient Descent with Warm Restarts (SGDR) (Loshchilov and Hutter, 2016) was selected as the solver. The initial learning rate and the momentum coefficient of learning policy, the weight decay and the subdivision (the number of mini-batches in a batch) were set to be 0.008, 0.9, 0.001 and 8, respectively. Table 2 shows the summary of all the hyper-parameters.

Mean average precision (mAP) on validation data and FPS as the speed of the algorithms were used for the validations of the model. The best mAP validation was chosen for further comparison.

To increase the performance of the models, pre-trained weights were used for Transfer learning (Kohavi, 1995) from COCO dataset (Lin et al., 2014).

4.3. Experimental platform

The experiments were executed on a computer with an Intel(R) Xeon(R) E5-2630 v4 at 2.20GHz with 20 cores and 32 GB RAM, running Ubuntu 20.04 with a kernel version of 5.11.0.

An NVIDIA Titan X GPU with 12 GB RAM was used for training and validating the various convolutional neural networks in this study.

All the algorithms were implemented and executed on Darknet (AlexeyAB, 2020a), which is an open source framework for neural networks. It is written in C and CUDA, and enables the execution on CPUs or GPUs.

OpenCV (Bradski, 2000) is also employed for the purpose of image processing.

5. Empirical Results and Discussion

This section provides the results of applying the proposed method. First, we apply the proposed detector on optical, thermal and fused images by varying customised weights to obtain the best customised weights to get the optimal performance. Next, we show the accuracy, speed, model size on our dataset. An Ablation study is performed to validate

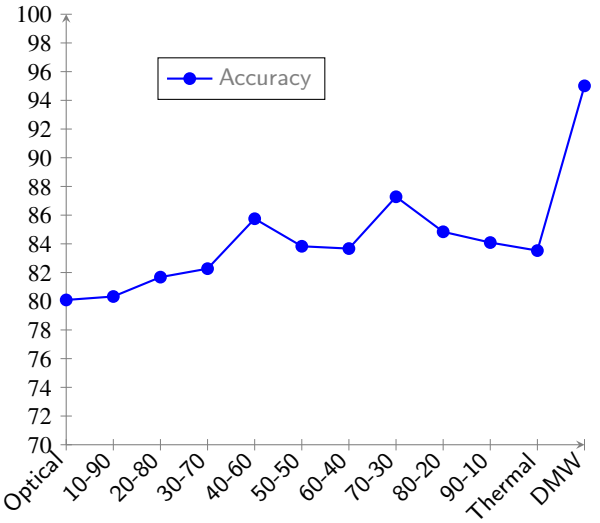


Figure 5: Customised weights for both optical and thermal components of the fused image.

the combination of the components selected for this study. The accuracy of the proposed technique is tested using a public dataset (KAIST). The speed is also tested in a constrained environment. Finally, the results are compared visually.

5.1. Result of proposed detector using customised weights

To validate and evaluate the efficacy of the proposed algorithm, a set of videos were collected by Police Scotland at various sites. The proposed machine learning model was compared with different state-of-the-art models. We applied the proposed detector on optical, thermal and fused images with customised weights. The weights were varied from 10% to 90% for both optical and thermal components of the fused image. Moreover, an additional training was carried out where the selection of the weights was not statically selected and chosen instead according to metadata provided by the police regarding the weather condition (sunny, hazy, cloudy) and lighting condition (day, night). This additional mode has been named as "Dynamic Metadata-based Weights", DMW in the figure. The idea behind this was the weather and lighting condition to be inserted manually by the user or automatically set by the application (out of scope of this contribution). Consequently, the weights in this mode were adjusted manually according to weather and lighting conditions provided in metadata following a set of rules. For instance, the weights associated with images taken at day light in sunny (hot) weather were selected as 40% (thermal)-60% (optical). In low illumination conditions and cloudy weather the associated weights were selected as 70% (thermal)-30% (optical). The low illumination condition does not include the foggy conditions due to operational limitations of the police Scotland. The image fusion was skipped for images taken at night in darkness and just thermal images were used in this regard.

Figure 5 shows the comparison of customised weights. As it can be seen the weights starts from 10% (thermal)-90% (optical) to 90% (thermal)-10% (optical). As results show, the best customised weights for our study were obtained when adjusted according to illumination conditions (DMW). Without this adjustment, the best trade-off is 70% thermal and 30% optical which is much lower in accuracy (95% versus 87.28). The accuracy (mAP%) of other customized weights were 80.33, 81.68, 82.27, 85.75, 83.83,83.67, 87.28, 84.84, 84.09 for 10% (thermal)-90% (optical), 20% (thermal)-80% (optical), 30% (thermal)-70% (Optical), 40% (thermal)-60% (optical), 50% (thermal)-50% (optical), 60% (thermal)-40% (optical), 70% (thermal)-30% (optical), 80% (thermal)-20% (optical), and 90% (thermal)-10% (optical), respectively.

5.2. Quantitative Results

Table 3 shows the comparison results of various state-of-the-art models including YOLOv4 which is the improved version of YOLOV3 (Bochkovskiy et al., 2020) against the BlendNet approach (our approach).

Table 3
Accuracy, speed, Model size and BPLOPS of different models with input size 608 trained with the police Scotland dataset.

Index	Model	Accuracy (mAP%)	Speed (FPS)	Model size (MB)	BFLOPS	Parameter (M)
1	Standard Tiny-YOLOv4	86.06	59.61	23.5	14.498	6
2	Standard YOLOv3	97.03	15.3	246.3	139.496	62
3	Standard Tiny-YOLOv4(3l)	92.58	57.39	24.5	17.127	6
4	Standard YOLOv4	98.57	17.5	256	127.232	64
5	Our approach	95.01	42.2	42.6	21.823	11

5.3. Accuracy and Speed Comparison

According to the results in Table 3, Figure 6, and Figure 7, the high accuracy of 98.57% and 97.03% was obtained by Standard YOLOv4 and Standard YOLOv3, respectively, but with 17.5 and 15.3 FPS on Titan X GPU with input size of 608, which makes them unsuitable for our mission-critical real-time use case. Consequently, a simplified, light version of YOLOv4, the standard Tiny-YOLOv4 (Bochkovskiy et al., 2020), was also trained and compared with the proposed BlendNet approach. It achieved an accuracy of 86.06% with 59.6 FPS on Titan X GPU with input size 608. However, it reduces the accuracy of humans at higher altitudes due to using only two output layers instead of three output layers in standard YOLOv4. Hence, an extra output layer has been added to the standard Tiny-YOLOv4 (Bochkovskiy et al., 2020). Tiny-YOLOv4 with 3 layers (3l) achieved an accuracy of 92.58% with 57.4 FPS. Our approach achieves an accuracy of 95.01% and 42.2 FPS on Titan X GPU. Considering the trade-off between speed and accuracy, BlendNet has high accuracy on small human detection at high altitudes while being fast, which is imperative for our use case to find missing people in the wild.

5.4. Model size and Complexity

Table 3 shows three metrics the model size, the Billion Floating-Point Operations (BFLOPS) and the model complexity which is the number of learnable parameters reported in Million (M) (Bianco et al., 2018). According to the results, the least complex model is standard Tiny-YOLOv4 with 23.5 MB and 14.5 BFLOPS and 5874116 as the number of learning parameters, respectively. The standard Tiny-YOLOv4 (3l) has a model size and BFLOPS of 24.5 MB and 17.1 BFLOPS and 6,114,390 number of parameters respectively. The standard YOLOv4 has a model size of 256, 127 BFLOPS and 64363101 parameters, which is not light enough for our use case. Similarly standard YOLOv3 has the model size of 246 MB, 139.5 BFLOPS and 61523734 learnable parameters, which is not suitable for our use case either. The model size of BlendNet is 43 MB with 22 BFLOPS. The number of learning parameters is also 10630102. Our approach is highly accurate whilst still fulfilling the requirements of complexity that the use case needs to achieve an optimised trade-off between detection accuracy and speed.

5.5. Ablation Experiment

To further prove the effectiveness of the proposed BlendNet, ablation experiments were performed. In ablation experiment 1, the accuracy result of Blendnet is presented removing the PAN architecture (YOLOv4(Tiny) +RBF). In the second experiment, RBF module was removed from the architecture (YOLOv4(Tiny) +PAN). The results in Table 4 show that BlendNet (combination of all modules) outperforms all the other alternative solutions.

5.6. Results with Public Dataset and Resource-Constrained Device

To further evaluate our method for human detection, the challenging public datasets KAIST has been selected and compared with the prevalent state-of-the-art algorithms. The KAIST dataset contains 96,312 aligned color-thermal images with total of 103,128 dense annotations and 1,182 unique pedestrians (Hwang et al., 2015). The size of each image is 640×512 pixels. The algorithm was trained on the train set of KAIST (set00-set05). The results were compared with BlendNet.

According to Table 5, standard YOLOv3 and YOLOV4 trained with the KAIST dataset obtained the average accuracy of 66.5 and 79.4 at a speed of 36 and 42 FPS, respectively on a Geforce GTX 2080ti GPU device (Xue et al., 2021). As apparent from the results, YOLOv3 and YOLOv4 introduce high power consumption and computational overhead and are slow for our use case. The M2Det model (Zhao et al., 2019) achieved 72.8% of accuracy with input size of 512 with speed of 13 FPS, which is less than that of the proposed BlendNet with the input size of 416 and not unsuitable for our use-case. The RefineDet (Duan et al., 2019) achieved 79.6%, which is slightly higher than that of our

Table 4
The Ablation Experiment

Model	Average
Standard Tiny-YOLOv4	92.58
YOLOv4(Tiny) + Modified RBF (PAN removed)	92.74
YOLOv4(Tiny) +Modified PAN (RBF removed)	93.47
BlendNet(YOLOv4(Tiny) +PAN +RBF)	95.01

Table 5
Accuracy and speed of different models with Kaist dataset

Model	Input size	Accuracy (mAP%)	Speed (FPS)	GPU
PftNet (Wei et al., 2020)	416 × 416	65.4	32	Geforce GTX 2080ti
MAF-YOLO (Xue et al., 2021)	416 × 416	87.8	40	Geforce GTX 2080ti
RefineDet (Zhang et al., 2018)	512 × 512	79.6	10	Geforce GTX 2080ti
M2det (Zhao et al., 2019)	512 × 512	72.8	13	Geforce GTX 2080ti
Standard YOLOV3	416 × 416	66.5	36	Geforce GTX 2080ti
YOLOv4	416 × 416	79.4	42	Geforce GTX 2080ti
Optimised Tiny YOLOV4 (Roszyk et al., 2022)	416 × 416	55.7	410	Nvidia RTX 3080
MFCG (Hua et al., 2022)	640 ×512	77.16	NG	Geforce GTX 2080ti
BlendNet	416 × 416	78.48	105	TITAN X

Table 6
The speed on constrained environment (Jetson)

Model	Speed
Standard Tiny-YOLOv4	31
Standard YOLOv3	5.9
Standard Tiny-YOLOv4(3l)	28
Standard YOLOv4	6.3
BlendNet	24

approach but with the input size of 512 with speed of 10 FPS and thus not suitable for our use case. PftNet (Wei et al., 2020) achieved 65.4% of accuracy with 32 FPS, which is slower and less accurate than that of BlendNet. MAF-YOLO (Xue et al., 2021) achieved higher accuracy of 87.8% but with the FPS of 40, which is slower than BlendNet. In (Roszyk et al., 2022), YOLOv4 and the tiny middle fusion approach to YOLOv4 were used in multi-spectral pedestrian detection. The Tiny version achieved 55.7%, which is less than ours (78.48), although a very high speed 410 was achieved due to optimisation using Tensorflow RT. In (Hua et al., 2022), the accuracy of 77.16 was achieved with input size 640×512, which is higher than 416; the speed was not mentioned in that study. BlendNet achieved the accuracy of 78.48% with 105 FPS on TITAN X GPU device. The results demonstrate that the proposed model can achieve the best accuracy and speed trade-off that was imperative for the success of the real-world use case.

The results in Table 6 show the speed of different models on a resource-constrained small-sized device (in this case a NVIDIA Jetson Xavier). As can be seen from the table, Standard YOLOv3 and standard YOLOv4 are not fast enough with input size 608 for our use case on this device. Tiny-YOLOv4 is the fastest among all but not accurate enough for our use case. Considering the trade-off between speed-accuracy, BlendNet can fulfill the requirements of both accuracy and speed for the use case.

5.7. Qualitative Results

To visually compare the detection results over the different operation modes (Optical only, Thermal only and Fusion), detection was conducted using a set of typical scenes. To this end, new testing videos (not used for training) were taken by Police Scotland in different real-world scenarios, including an open fields and cluttered natural environments with dense vegetation. As seen in Figures 8(a), 8(d), positive detection was missed in the Optical mode

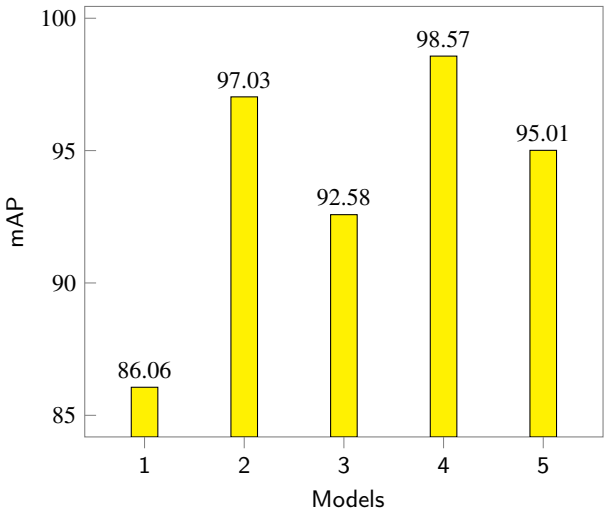


Figure 6: Comparison of mAP of different models with input size 608. See the models' indices in Table 3.

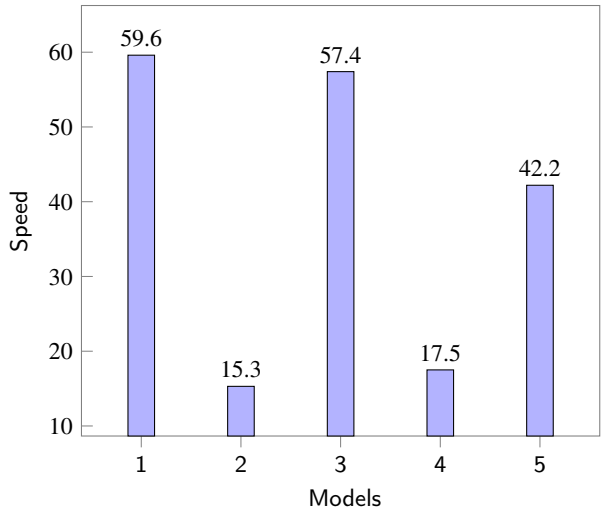


Figure 7: Comparison of speed of different models with input size 608. See models' indices in Table 3.

due to the human object wearing low-visibility clothing (camouflage in Figure 8(a) and chartreuse in (Figure 8(d)). In contrast, the human objects were successfully detected in both thermal and fused images (Figures 8(b), 8(c), 8(e), and 8(f)), although the the proposed Fusion mode achieved significantly higher confidence in detection, e.g., 61% (Fusion) vs. 35% (Thermal) and 47% (Fusion) vs. 30% (Thermal) in Figure 8(f) vs. Figure 8(e). In the dense vegetation scene (Figure 8(g)), positive detection was missed in the optical mode due to the human object being under trees. In the Thermal mode (Figure 8(h)), the human object on the left was not detected. However, both human objects were detected in the Fusion mode (Figure 8(i)). In the last scenario, a human object was detected in both the Optical and Fusion modes (Figures 8(j), 8(l)). However, the detection was missed in the Thermal mode (Figure 8(k)) due to the sunny and hot weather and thus human object's temperature signature being similar to the surrounding environment. The results show the superior effectiveness in the detection using the proposed illumination-aware human detection system with no negative detection observed.

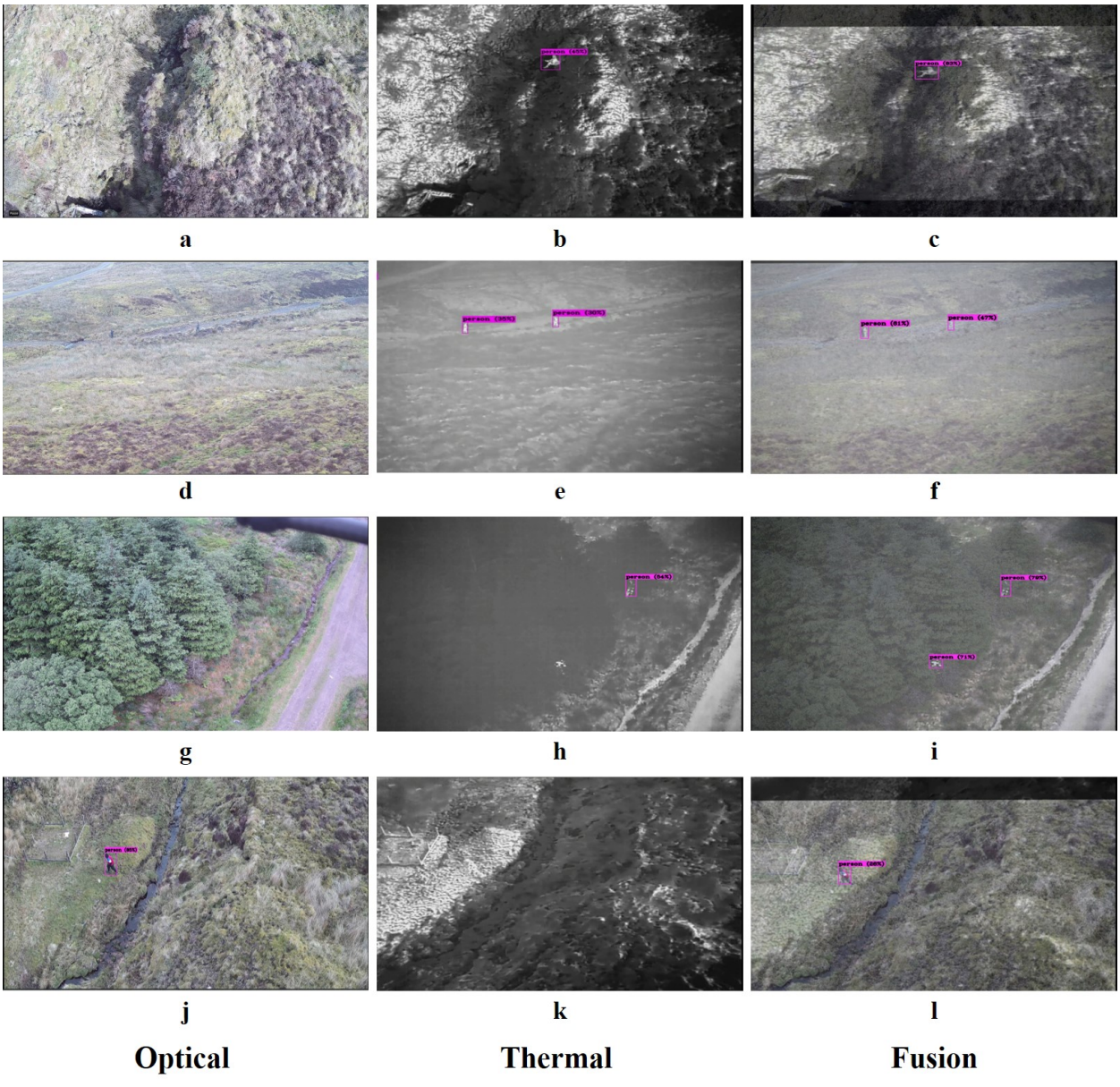


Figure 8: Fusion results.

6. Concluding Remarks and Future Work

In this study, we have proposed a new illumination-aware all-around machine-learning-based human detection system for providing cost-effective high-accuracy and real-time small human object detection in UAV-based use cases in adverse environments. The proposed detection system features three-step customised low-distortion image fusion scheme of optical and thermal imagery as well as a new CNN-based BlendNet algorithm for fast and accurate small object detection by modifying the RBF and PAN architectures to increase the accuracy of small humans from distance. The proposed solution is superior to the state-of-the-art machine learning based solutions in terms of accuracy and speed trade-off for small humnan objects. BlendNet has achieved the mAP accuracy of 95.01% and 42.2 FPS. The proposed system can be used in many areas of UAV-based applications including search and rescue operations, surveillance such as intruder detection, emergency and vigilance (e.g., the Drone Angel), to name a few. In future

work, the proposed solution will be embedded on more portable devices such as smartphones or tablets, and will be tested in an adversarial scenario. Automatic selection of customised weights for image fusion based on metadata is another future work.

Supplementary Material

Video of a missing people operation: http://beyond5ghub.uws.ac.uk/index.php/search-and-rescue/

CRediT authorship contribution statement

Gelayol Golcarenarenji: Writing original draft, Software, Methodology, Validation. **Ignacio Martinez-Alpiste:** Writing original draft, Data curation, Methodology. **Qi Wang:** Supervision, Writing, review, editing, Project administration, Funding acquisition. **Jose Maria Alcaraz-Calero:** Supervision, Writing, review, editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was in part funded by the Innovation Centre for Sensor and Imaging Systems (CENSIS) and in collaboration with Thales under the Grant number CAF-0680, the EU Horizon 2020 5G-PPP 5G-INDUCE project ("Open cooperative 5G experimentation platforms for the industrial sector NetApps") under the Grant number H2020-ICT-2020-2/section*101016941, and the EU Horizon 2020 ARCADIAN-IoT project ("Autonomous Trust, Security and Privacy Management Framework for IoT") under the Grant number 101020259. The authors would like to thank all the partners in these projects for their support.

References

AlexeyAB (2020a). Darknet. https://github.com/AlexeyAB.

AlexeyAB (2020b). Darknet. https://github.com/AlexeyAB/Yolo_mark.

Bianco, S., Cadene, R., Celona, L., Napoletano, P. (2018). Benchmark analysis of representative deep neural network architectures. IEEE Access, 6, 64270–64277.

Bochkovskiy, A., Wang, C.Y., Liao, H. (2020). Yolov4: Optimal speed and accuracy of object detection. ArXiv, abs/2004.10934.

Bradski, G. (2000). The opencv library. Dr. Dobb's Journal: Software Tools for the Professional Programmer, 25, 120–123.

Cao, Y., Guan, D., Huang, W., Yang, J., Cao, Y., Qiao, Y. (2019). Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. information fusion, 46, 206–217.

Cao, Z., Yang, H., Zhao, J., Guo, S., Li, L. (2021). Attention fusion for one-stage multispectral pedestrian detection. Sensors, 21, 4184.

Choi, S., Kwon, O.J., Lee, J. (2017). A method for fast multi-exposure image fusion. IEEE Access, 5, 7371–7380.

Dai, J., Li, Y., He, K., Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks, in: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (Eds.), NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, Curran Associates, Inc.. p. 379–387. https://proceedings.neurips.cc/paper/2016/file/577ef1154f3240ad5b9b413aa7346a1e-Paper.pdf.

Dandrifosse, S., Carlier, A., Dumont, B., Mercatoris, B. (2021). Registration and fusion of close-range multimodal wheat images in field conditions. Remote Sensing, 13. https://www.mdpi.com/2072-4292/13/7/1380, doi:10.3390/rs13071380.

Dawdi, T.M., Abdalla, N., Elkalyoubi, Y.M., Soudan, B. (2020). Locating victims in hot environments using combined thermal and optical imaging. Computers & Electrical Engineering, 85, 106697.

Ding, L., Wang, Y., Laganière, R., Huang, D., Luo, X., Zhang, H. (2021). A robust and fast multispectral pedestrian detection deep network. Knowledge-Based Systems, 227, 106990. https://www.sciencedirect.com/science/article/pii/S0950705121002537, doi:https://doi.org/10.1016/j.knosys.2021.106990.

Dollar, P., Tu, Z., Perona, P., Belongie, S. (2009). Integral channel features, in: Proceedings of the British machine vision conference, BMVA Press. (pp. 91.1–91.11). Doi:10.5244/C.23.91.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q. (2019). Centernet: Keypoint triplets for object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, (pp. 6569–6578).

Dwibedi, D., Misra, I., Hebert, M. (2017). Cut, paste and learn: Surprisingly easy synthesis for instance detection, in: Proceedings of the IEEE international conference on computer vision, (pp. 1301–1310).

Evangelidis, G.D., Psarakis, E.Z. (2008). Parametric image alignment using enhanced correlation coefficient maximization. IEEE transactions on pattern analysis and machine intelligence, 30, 1858–1865.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A. (2010). Cascade object detection with deformable part models, in: 2010 IEEE computer society conference on computer vision and pattern recognition, (pp. 2241–2248).

Fu, L., Gu, W.b., Ai, Y.b., Li, W., Wang, D. (2021). Adaptive spatial pixel-level feature fusion network for multispectral pedestrian detection. Infrared Physics & Technology, 116, 103770.

Girshick, R. (2015). Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, (pp. 1440–1448).

Golcarenarenji, G., Martinez-Alpiste, I., Wang, Q., Alcaraz-Calero, J.M. (2021). Efficient real-time human detection using unmanned aerial vehicles optical imagery. International Journal of Remote Sensing, 42, 2440–2462.

Golcarenarenji, G., Martinez-Alpiste, I., Wang, Q., Alcaraz-Calero, J.M. (2022). Machine-learning-based top-view safety monitoring of ground workforce on complex industrial sites. Neural Computing and Applications, 34, 4207–4220.

Guan, D., Cao, Y., Yang, J., Cao, Y., Yang, M.Y. (2019). Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion, 50, 148–157.

Hua, C., Sun, M., Zhu, Y., Jiang, Y., Yu, J., Chen, Y. (2022). Pedestrian detection network with multi-modal cross-guided learning. Digital Signal Processing, 122, 103370.

Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline, in: Proceedings of the IEEE conference on computer vision and pattern recognition, (pp. 1037–1045).

Hwooi, S.K.W., Sabri, A.Q.M. (2017). Enhanced correlation coefficient as a refinement of image registration, in: 2017 IEEE international conference on signal and image processing applications (ICSIPA), IEEE. (pp. 216–221).

Jiang, H., Wang, S. (2016). Object detection and counting with low quality videos. Technical Report. Stanford University.

Jung, A.B., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C., Yadav, S., Banerjee, J., Vecsei, G., Kraft, A., Rui, Z., Borovec, J., Vallentin, C., Zhydenko, S., Pfeiffer, K., Cook, B., Fernández, I., De Rainville, F.M., Weng, C.H., Ayala-Acevedo, A., Meudec, R., Laporte, M., et al. (2020). Data augmentation. https://github.com/aleju/imgaug.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the fourteenth international joint conference on artificial intelligence, 1995, American Association for Artificial Intelligence. (pp. 1137–1143).

Krishna, H., Jawahar, C. (2017). Improving small object detection, in: 2017 4th IAPR Asian conference on pattern recognition (ACPR), IEEE. (pp. 340–345).

Li, C., Song, D., Tong, R., Tang, M. (2019). Illumination-aware faster r-cnn for robust multispectral pedestrian detection. Pattern Recognition, 85, 161–171.

Li, J., Peng, Y., Jiang, T. (2021). Embedded real-time infrared and visible image fusion for uav surveillance. Journal of Real-Time Image Processing, 18, 2331–2345.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), (pp. 2117–2125).

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. (2014). Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. (pp. 740–755).

Liu, J., Zhang, S., Wang, S., Metaxas, D.N. (2016a). Multispectral deep neural networks for pedestrian detection, in: Richard C. Wilson, E.R.H., Smith, W.A.P. (Eds.), Proceedings of the British machine vision conference (BMVC), BMVA Press. (pp. 73.1–73.13).

Liu, S., Huang, D., et al. (2018a). Receptive field block net for accurate and fast object detection, in: Proceedings of the European conference on computer vision (ECCV), (pp. 385–400).

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. (2018b). Path aggregation network for instance segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, (pp. 8759–8768).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016b). Ssd: Single shot multibox detector, in: Proceedings of the European conference on computer vision (ECCV), Springer. (pp. 21–37).

López, A., Jurado, J.M., Ogayar, C.J., Feito, F.R. (2021). A framework for registering uav-based imagery for crop-tracking in precision agriculture. International Journal of Applied Earth Observation and Geoinformation, 97, 102274.

Loshchilov, I., Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, .

Martinez-Alpiste, I., Casaseca-de-la-Higuera, P., Alcaraz-Calero, J., Grecos, C., Wang, Q. (2019). Benchmarking machine-learning-based object detection on a uav and mobile platform, in: 2019 IEEE wireless communications and networking conference (WCNC), (pp. 1–6). doi:10.1109/WCNC.2019.8885504.

Martinez-Alpiste, I., Golcarenarenji, G., Wang, Q., Alcaraz-Calero, J.M. (2020a). Altitude-adaptive and cost-effective object recognition in an integrated smartphone and uav system, in: 2020 European conference on networks and communications (EuCNC), (pp. 316–320). doi:10.1109/EuCNC48522.2020.9200951.

Martinez-Alpiste, I., Golcarenarenji, G., Wang, Q., Alcaraz-Calero, J.M. (2020b). Real-time low-pixel infrared human detection from unmanned aerial vehicles, in: Proceedings of the 10th ACM symposium on design and analysis of intelligent vehicular networks and applications, (pp. 9–15).

Martinez-Alpiste, I., Golcarenarenji, G., Wang, Q., Alcaraz-Calero, J.M. (2021). Search and rescue operation using uavs: a case study. Expert Systems with Applications, 178, 114937.

Martinez-Alpiste, I., Casaseca-de-la Higuera, P., Alcaraz-Calero, J.M., Grecos, C., Wang, Q. (2020c). Smartphone-based object recognition with embedded machine learning intelligence for unmanned aerial vehicles. Journal of Field Robotics, 37, 404–420.

Meng, L., Zhou, J., Liu, S., Ding, L., Zhang, J., Wang, S., Lei, T. (2021). Investigation and evaluation of algorithms for unmanned aerial vehicle multispectral image registration. International Journal of Applied Earth Observation and Geoinformation, 102, 102403.

Pei, D., Jing, M., Liu, H., Sun, F., Jiang, L. (2020). A fast retinanet fusion framework for multi-spectral pedestrian detection. Infrared Physics & Technology, 105, 103178. https://www.sciencedirect.com/science/article/pii/S1350449519305845, doi:https://doi.org/

10.1016/j.infrared.2019.103178.

Perez, L., Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621, .

Piao, J., Chen, Y., Shin, H. (2019). A new deep learning based multi-spectral image fusion method. Entropy, 21, 570.

Raudonis, V., Paulauskaite-Taraseviciene, A., Sutiene, K. (2021). Fast multi-focus fusion based on deep learning for early-stage embryo image enhancement. Sensors, 21, 863.

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

Redmon, J., Farhadi, A. (2017). Yolo9000: better, faster, stronger, in: Proceedings of the IEEE conference on computer vision and pattern recognition, (pp. 7263–7271).

Redmon, J., Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, .

Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks, in: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in neural information processing systems, Curran Associates, Inc.. (pp. 91–99). https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

Roszyk, K., Nowicki, M.R., Skrzypczyński, P. (2022). Adopting the yolov4 architecture for low-latency multispectral pedestrian detection in autonomous driving. Sensors, 22, 1082.

Rudol, P., Doherty, P. (2008). Human body detection and geolocalization for uav search and rescue missions using color and thermal imagery, in: 2008 IEEE aerospace conference, IEEE. (pp. 1–8).

Song, X., Gao, S., Chen, C. (2021). A multispectral feature fusion network for robust pedestrian detection. Alexandria Engineering Journal, 60, 73–85.

Surasak, T., Takahiro, I., Cheng, C.h., Wang, C.e., Sheng, P.y. (2018). Histogram of oriented gradients for human detection in video, in: 2018 5th International conference on business and industrial research (ICBIR), IEEE. (pp. 172–176).

Teutsch, M., Krüger, W. (2012). Detection, segmentation, and tracking of moving objects in uav videos, in: 2012 IEEE ninth international conference on advanced video and signal-based surveillance, IEEE. (pp. 313–318).

Van Etten, A. (2018). You only look twice: Rapid multi-scale object detection in satellite imagery. arXiv preprint arXiv:1805.09512, .

Vandersteegen, M., Beeck, K.V., Goedemé, T. (2018). Real-time multispectral pedestrian detection with a single-pass deep neural network, in: International conference image analysis and recognition, Springer. (pp. 419–426).

Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H. (2020). Cspnet: A new backbone that can enhance learning capability of cnn, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, (pp. 390–391).

Wei, X., Zhang, H., Liu, S., Lu, Y. (2020). Pedestrian detection in underground mines via parallel feature transfer network. Pattern Recognition, 103, 107195.

Wu, Z., Wu, X., Zhu, Y., Zhai, J., Yang, H., Yang, Z., Wang, C., Sun, J. (2022). Research on multimodal image fusion target detection algorithm based on generative adversarial network. Wireless Communications and Mobile Computing, 2022, 1740909.

Xue, Y., Ju, Z., Li, Y., Zhang, W. (2021). Maf-yolo: Multi-modal attention fusion based yolo for pedestrian detection. Infrared Physics & Technology, 118, 103906. https://www.sciencedirect.com/science/article/pii/S1350449521002784, doi:https://doi.org/10.1016/j.infrared.2021.103906.

Yu, K., Ma, J., Hu, F., Ma, T., Quan, S., Fang, B. (2019). A grayscale weight with window algorithm for infrared and visible image registration. Infrared Physics & Technology, 99, 178–186.

Yu, X., Gong, Y., Jiang, N., Ye, Q., Han, Z. (2020). Scale match for tiny person detection, in: The IEEE winter conference on applications of computer vision, (pp. 1257–1265).

Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z. (2018). Single-shot refinement neural network for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, (pp. 4203–4212).

Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., Ling, H. (2019). M2det: A single-shot object detector based on multi-level feature pyramid network, in: Proceedings of the AAAI conference on artificial intelligence, (pp. 9259–9266).