International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137

Volume – 08, Issue – 04, April 2023, PP – 18-27

Using Yolo Technology and SWIR Images to Identify and Analyze Human and Vehicles in Smog

Chih-Wei Kuan[1], Mei-Ling Yeh[2], Tien-Yin Chou[2], Chih-Ling Wu[2]

[1]*(PHD student of Infrastructure Planning and Engineering, Feng Chia University, Taiwan)*

[2]*(GIS center, Feng Chia University, Taiwan)*

Abstract: When haze or sandstorms occur in the environment, there will be smog and foggy conditions, which cannot effectively identify human and vehicles. In this paper, SWIR images have the characteristics of penetrating smoke, and Yolo deep learning technology is used to conduct automatic identification of human and vehicles. The research results show that SWIR images can successfully display their high penetration when visible light images cannot effectively identify human and vehicles, thereby improving the recognition of human and vehicles. The research results of this paper are quite successful, and can be applied to traffic safety in the future, by embedding the module into the vehicle-mounted lens module, it can be recognized and given a warning in real time to improve driving safety.

Keywords: Short-wave infrared (SWIR), Smoke penetration, Yolo, Deep learning

I. INTRODUCTION

In the environment, due to natural disasters, such as forest fires, sandstorms, air pollution, and even smoke from wars, it is impossible to effectively identify human and vehicles in the environment, which will put passers-by in a high-risk situation. In order to keep abreast of the dynamics ahead even when the visibility is not high, it is very important to improve the driving perspective smoke equipment. Visible light devices installed in general cars cannot see through smog, but in the optical field, short-wave infrared light (SWIR) has the property of seeing through smog. This research will independently develop and design the SWIR sensor to improve the device's see-through ability, and transmit the sensor data to the computer for target recognition. This study believes that if the SWIR intelligent identification system can be combined with related functions of vehicle-mounted equipment in the future, it can be more widely used in various use scenarios. It is hoped that after the development of the SWIR intelligent identification system is mature, it can use SWIR images to quickly and smoothly identify targets in smoky environments to achieve driving safety.

II. RELATED WORKS

The purpose of this study is to use the characteristics of SWIR that can penetrate smoke to evaluate the feasibility of automatic AI identification of human and vehicles in a smoky environment. It is necessary to understand the characteristics of SWIR and the analysis principle of AI identification modules. Therefore, this study collects SWIR smoke penetration and AI Identify the relevant characteristics, principles and application cases of the literature, as the implementation direction and reference basis in the research process.

1. Characteristics of SWIR smoke penetration

SWIR Short-wave infrared light refers to non-visible light with a wavelength of 1,400 to 3,000 nm. Zeng Xinmiao et al. (2011) mentioned in "Analysis and Research on Shortwave Infrared Imaging Systems" that the current main product representatives of optical sensors are CCD and CMOS, and CCD optical sensors can only detect projected visible light or infrared light. The detection of the image cannot achieve the penetration effect on the target object.The SWIR sensor senses the radiant heat of an object. Because the temperature difference between an object active at night and the background is greater than that during the day, and it has a penetrating effect on smoke, fog, sand, dust or camouflage, it can work no matter in day or night , and can be used as all-day and night security monitoring, as shown in Fig. 1.

Fig. 1 Differences in penetration and clarity of the two images (a) Visible light image (b) SWIR image. (Image referenced from Sensors Unlimited.)

Bertozzi et al. studied the application of low-cost sensors in advanced driver assistance systems (Advanced Driver Assistance Systems, ADAS) in 2013. The paper mentions the contribution of short-wave infrared light to pedestrian detection and sensing under different visibility conditions (such as under smoke). Smoke and fog in the atmosphere have different characteristics of absorption, refraction and reflection of short-wave infrared light. The image of the research and test results is shown in Fig. 2. The author conducts image detection with sensors in different bands of clear (Clear), haze (Haze) and water mist (Fog), using three filters: C is visible light to short-wave infrared Light (400 to 1,700um), F1 is short-wave infrared light (1,000 to 1,700um), F2 is short-wave infrared light (1,300 to 1,700um), and the research results are shown in haze (dust, smoke and other wet or dry particles) The penetration performance of medium and short-wave infrared rays (F1, F2) is the best, while the detection and avoidance of short-wave infrared light in water mist still requires the assistance of other sensors.

Fig. 2 Detection situation of different bands for different visibility. (Image referenced from https://www.hindawi.com/journals/isrn/2014/858979/fig12/)

2. AI recognition and identification

YOLO (you only look once, YOLO) is a neural network-like calculation suite for object detection. It does not use any famous deep learning framework for implementation. It is lightweight, has few dependencies,

and has high algorithm efficiency. It is widely used in the industry. The application fields are very valuable, such as pedestrian detection, industrial image detection and so on.

Early target detection methods usually extract some robust features of the image (such as HAAR, SIFT, HOG, etc.), use deformable parts model (DPM), and use sliding window (sliding window) to predict objects with Better bounding box. But this method is very time-consuming, and the accuracy is not high.

Later, the object proposal method appeared (selective search is a typical representative of this type of method). Compared with the exhaustive method of sliding window, it reduces a lot of calculations and greatly improves the performance.Using the results of selective search, after the appearance of regions with convolutional neural network (R-CNN) combined with convolutional neural network, the performance of object detection has been improved qualitatively. The SPPnet, Fast R-CNN, Faster R-CNN and other methods developed based on R-CNN have proved the effectiveness of the "Proposal + Classification" method in object detection. Compared with the R-CNN series of methods, Ren et al. (2016) provided another way of thinking, transforming the problem of object detection into a regression problem. Given an input image, directly regress the outer frame of the object and its classification category at multiple positions of the image. YOLO is a convolutional neural network that can predict the position and category of multiple outer frames at one time. It can realize point-to-point target detection and recognition. Its biggest advantage is its speed. In fact, the essence of target detection is regression, so a CNN that implements a recursive function does not require a complicated design process. YOLO did not choose to slide the window or extract the target to train the network, but directly used the whole image as the training model. The advantage of this is that it can better distinguish the target area and the background area. In contrast, Fast R-CNN using the target training method often misjudges the background area as a specific target.

YOLO is somewhat similar to R-CNN, each grid region is a potential bounding box, and convolutional features are used to score these boxes. However, the YOLO system uses spatial constraints on the grid to help us reduce multiple detections of the same object. The YOLO system proposes fewer bounding boxes, only 98 bounding boxes per image, while selective search requires 2000 boxes. Finally, the independent components of the model are combined into a single co-optimized model.

Developed from the past thinking all the way, the biggest feature of YOLO is direct point-to-point object detection, using the whole picture as the input of the neural network, directly predicting the coordinate position of the outer frame, the credibility of the object contained in the outer frame, and the object's belonging category.

YOLO v2 can detect more than 9,000 kinds of object classifications in real time and propose a new joint training algorithm (joint training algorithm), which uses a hierarchical view to classify objects, and uses a huge classification data set to expand the detection data set, and then Mixing two different databases, Ren et al. (2016) imported the anchor box in Faster R-CNN to no longer directly map the coordinates of the outer frame, but to predict relative parameters, and use K-Means cluster classification, In addition, the v2 version changes all the fully connected layers of v1 to convolution calculations, and instead of using dropout, it uses batch normalization for optimization, which improves convergence and eliminates normalization of other forms Dependence, by adding batch normalization to each convolutional layer of the Redmon and Farhadi (2017) network, the final mAP is increased by 2%.

Redmon et al. (2016) proposed YOLO v3 by referring to other architectures and optimizing their own models. After using the Resnet network, the new base network of v3 is changed to Darknet-53. As the number of network layers continues to deepen, the use of the Resnet structure effectively solves the problem of gradient disappearance or gradient explosion. Using the feature pyramid network (FPN) multi-level prediction network, the prediction ability of the model for small objects is improved, the feature layer is changed from a single layer to a multi-layer, and the prediction types are also increased from 5 types in a single layer to 3 types in each layer There are a total of 9 object categories. The architecture of FPN can integrate the better target position at the lower level and the better semantic features at the higher level, and make independent predictions at different feature levels, making the improvement of small object detection very significant, as shown inFig. 3, the potholes, cars, pedestrians, roadblocks, etc. in the screen, and the detected type and the intersection and union ratio during detection are displayed on the detection result. The detection of small targets usually leads to the failure of traditional model detection. The improved multi-dimensional prediction framework based on FPN can generate multi-dimensional features. Three scales can be used for point-to-point training and used in both training and testing, thus, FPN can improve the accuracy without increasing the architecture.

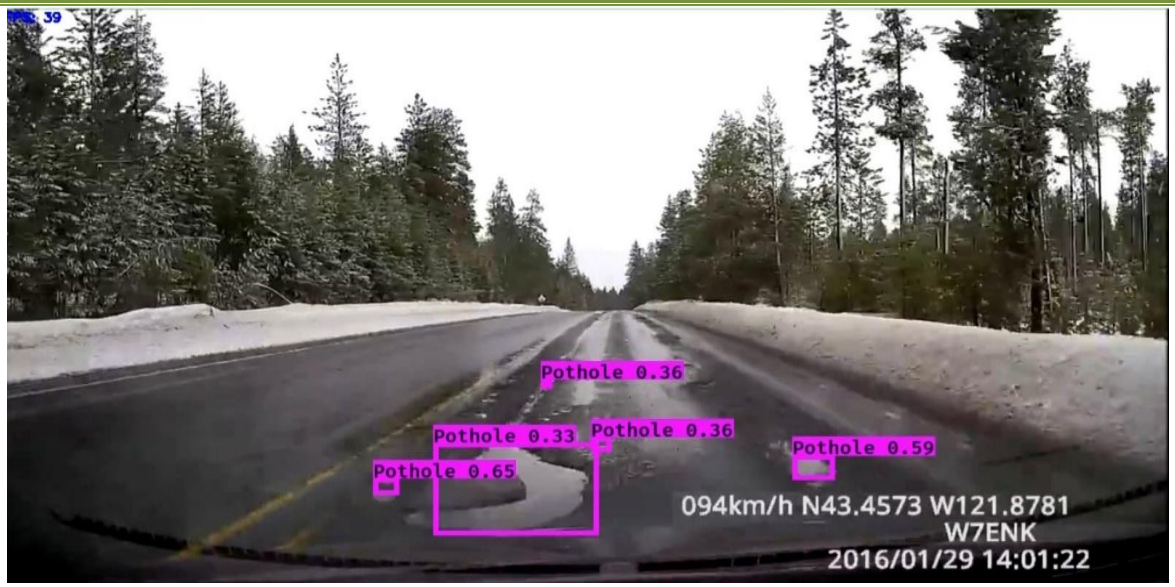

Fig. 3 The actual application of YOLO v3 to the detection screen of road information. (Image referenced from lh3.googleusercontent.com)

The introduction of YOLO v4 in Figure 4 presents a new performance. The architecture of the YOLO v4 model consists of three parts: BackBone: CSPDarknet53, Neck: SPP+PAN, HEAD: YOLO HEAD. In order to enable the network to operate quickly and optimize parallel calculations, two types of neural networks are used, a small number of groups (1~8 groups) are used in the convolutional layer, and ResNeXt50 and Darknet53 are combined with Cross Stage Partial Network (CSPNet) combined to form CSPResNeXt50 and CSPDarknet53.

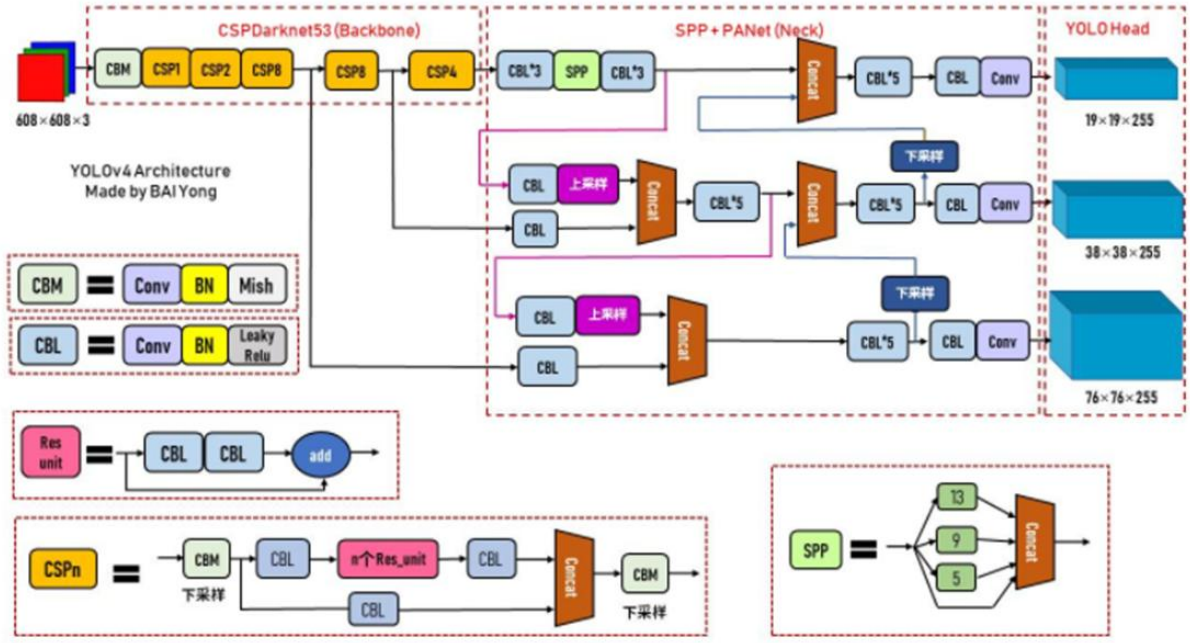

Fig. 4 The network structure of YOLO v4.(Image referenced from Yolov4)

In the target detection task, it usually runs on a small device, so a model with a low computational load is required to shorten the inference prediction time. Wang et al. (2019) believed that the problem of high prediction calculations was caused by the repetition of gradient information in network optimization, so they proposed a cross-stage local network (CSPNet). The main purpose of CSPNet is to enable the network architecture to obtain richer gradient fusion information and reduce the amount of calculation. The method is to first divide the feature map of the Base layer into two parts, and then pass through transition,concatenation, transition. Transition layer represents the transition layer, including 1x1 convolutional layer and pooling layer.

This approach enables CSPNet to solve three problems: 1. Increase the learning ability of CNN. Even if the model is lightweight, it can maintain accuracy. 2. Remove the computational power. High computing bottleneck structure, 3. Reduce memory usage.

After CSPNet combines different backbones, the accuracy of ImageNet classification remains unchanged or slightly improved, but the amount of calculation is greatly reduced. In the results of MS COCO detection, CSPNet's AP50 is significantly higher than other methods, and at the same accuracy, FPS is much faster.

The head part follows the head of YOLO v3. The part of Neck is to expand the receptive field and integrate information of feature maps of different scales (better feature fusion), PANet (Path Aggregation Network) is improved based on FPN, and the number of layers of strings is added. One more layer. In addition, the original added part is modified to merge, the effect will be better than the added one, but the disadvantage is that there will be more channels and the amount of calculation will increase.

In addition, the complete methods used by YOLO v4 include BoF for backbone, CutMix, Mosaic data augmentation, Drop Block regularization, Class label smoothing, Mish activation, Bag of Freebies (BoF) for detector, Cross mini-Batch Normalization, Self-Adversarial Training, Optimal hyper -parameters, DIoU-NMS, these methods are the improved parts of YOLO v4, so YOLO v4 greatly improves the detection accuracy of the model and reduces the requirements for hardware use. As can be seen from Figure 5, YOLO v4 has obtained an AP value of 43.5% (65.7% AP50) on the MS COCO dataset. YOLO v4 is twice as fast as EfficientDet with the same performance as Efficient Det; compared with YOLOv3, YOLO v4's AP and FPS are increased by 10% and 12%, respectively, as shown in Figure 5.

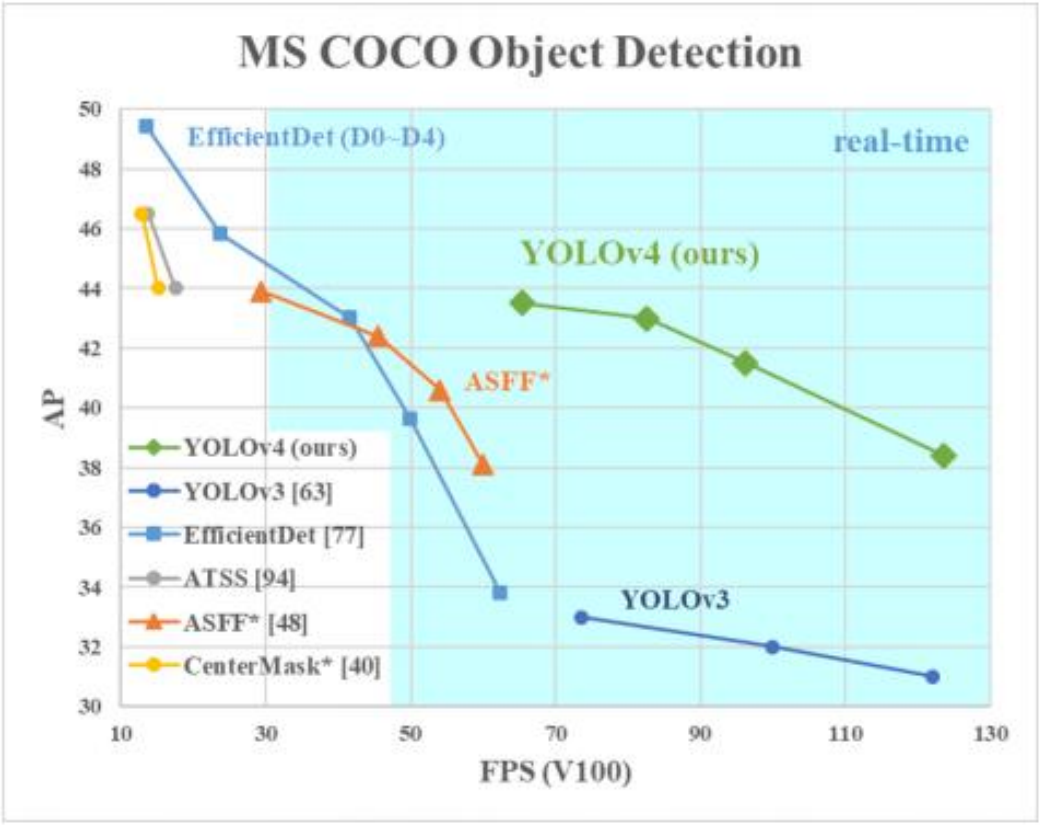

Fig. 5 Performance comparison between YOLOv3/4 and EfficientDet. (Image referenced from yolov4 darknet)

Through various versions of YOLO, this project will use the most appropriate framework for the hardware design of traffic flow recognition, which will avoid the aforementioned intrusive hardware installation or the limitations of other systems.

III. RESEARCH METHOD

1. Research process flow

The research process is shown in the Fig. 6, including SWIR image AI recognition module development, electronic circuit design and subsequent product testing.

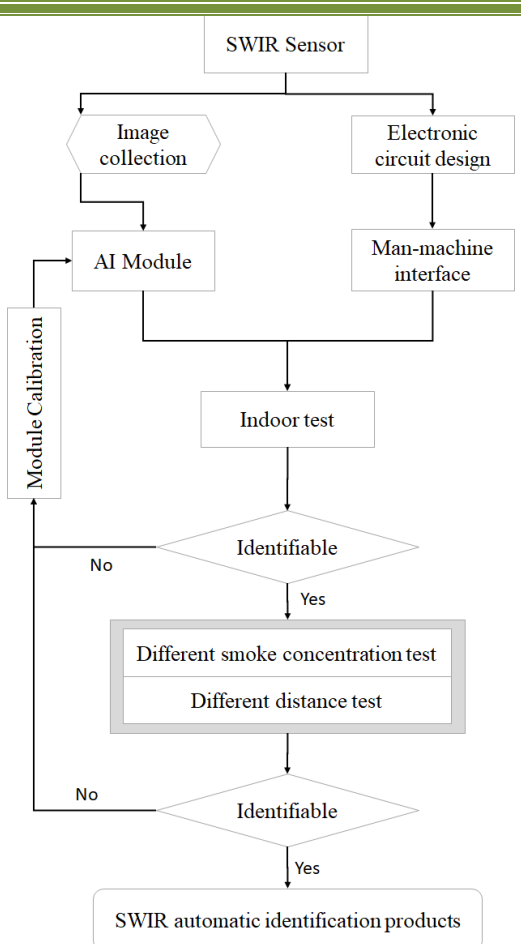

Fig. 6 Flowchart of this study

2. SWIR sensor and image collection

Since the short-wave infrared light in the optical band between 900-1,700 nm has the function of penetrating smoke, we choose the short-wave infrared camera produced in Taiwan by HUNSON Company, which can achieve a quantum efficiency of more than 85% at 1,200-1,700 nm(Shown as Fig. 7). This band is in the range of short-wave infrared light, and its related specifications are shown in

Table 1.

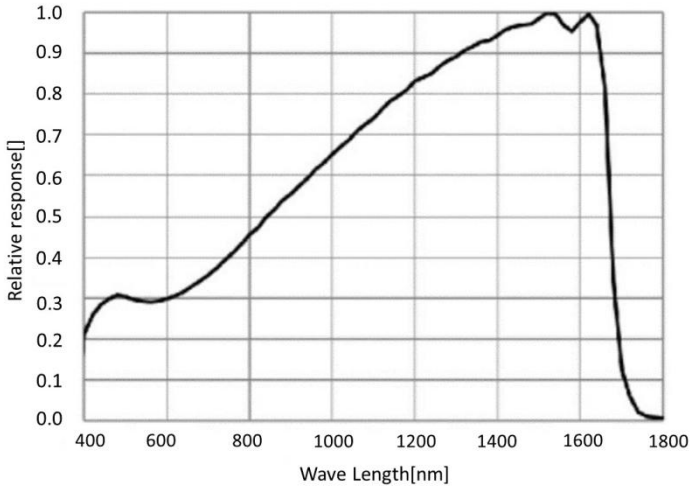

Fig. 7 The relationship between the spectral wavelength of the SWIR camera and the relative response value in this study. (Image referenced from techxpert.com)

Table 1 SWIR Camera Specifications.

Model	HN-W1-M	Sensing Unit	Sony IMX991
Sensing Unit Size	1/4”	Technology	Global shutter
Pixel resolution	H*V=640*512(0.33M)	Pixel size	5 μm
Frame Rate	240@8 bit	Output Interface	USB 3.1 Gen1
Video Output Format	8, 12 bit	Lens Connector	C Mount
Camera Size	W=58mm, H=44mm, L=50 mm	Weight	187g
Operating temperature	0 ~ 50ºC	Humidity	20 ~ 80%

Use the SWIR camera to collect several SWIR images of human and vehicles with different distances and angles for subsequent identification of sample data.

3. Electronic circuit design

The circuit of the SWIR identification system in this study is divided into two stages. First, an image capture module will be developed in response to the sensor, and the SWIR image will be streamed through the module as available information, and then imported into the AI identification module. First, identify human and vehicles in the SWIR image and display the number of identified objects; second, mark the identified image with a crosshair, and output the identification screen and identification results to the micro-display for display. The overall circuit design architecture is shown in Fig. 8.

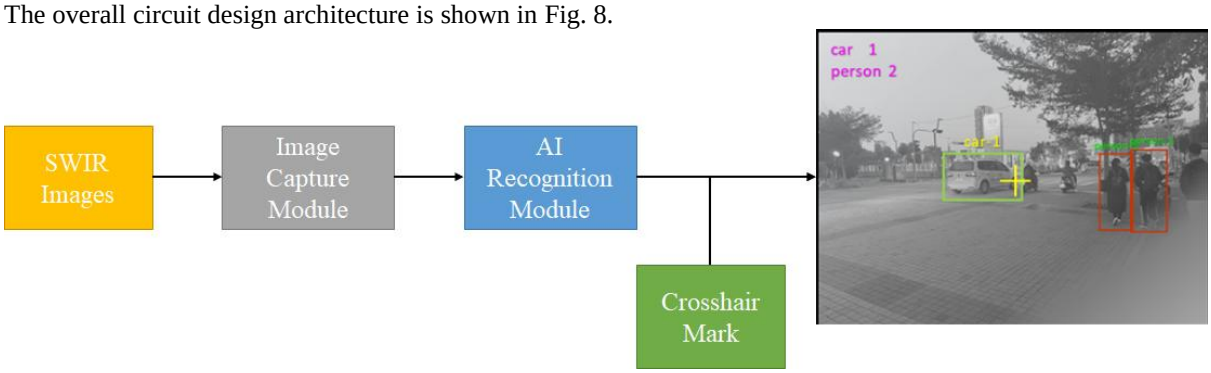

Fig. 8 The SWIR identification system circuit architecture of this research.

4. AI identification module development

Specifically, YOLO's CNN network divides the input image into S×S grids, and then each grid is responsible for detecting objects whose center points fall within the grid. Each grid predicts B bounding boxes and C conditional class probabilities Pr(Pr = 1 when the grid contains the object, otherwise Pr = 0). It is characterized by the probability that the bounding box object predicted by the grid belongs to each class, but these probability values are actually conditional probabilities under the confidence level of each bounding box. No matter how many bounding boxes a cell predicts, only one set of class probability values is predicted, which is a shortcoming of the YOLO algorithm. In later improved versions, YOLO9000 concatenates class probability predictions with bounding boxes. And each bounding box contains (x, y, w, h, S), S is the confidence level, and the result is expressed as S = Pr * IoU. Among them, the intersection over union (IoU) IoU(A,B) is the predicted result and the actual intersection divided by the union:

$$IoU(A,B)=(A\cap B)/(A\cup B)$$

The most commonly used indicator for general prediction is IoU(A,B)>0.5, which means that in a frame prediction, the calculated IoU indicates that the test result is successful. So S actually only has two values, 0 or IoU itself. The predicted value (x, y) of the center coordinates is the offset value relative to the coordinate point of the upper left corner of each cell, and the unit is relative to the cell size (equivalent to positive normalization), while the w and h of the bounding box The predicted value is relative to the ratio of the width to the height of the entire picture, so theoretically the size of the 4 elements should be in the [0, 1] range (that is, x, y is the position of the frame relative to the grid, and w, h is the relative for the entire picture). Each cell needs to predict

(B ∗ 5 + C) values. If the input image is divided into an S×S grid, then the final predicted value is a tensor of size S × S × (B ∗ 5 + C).

IV. EXPERIMENT RESULT

In this study, the images obtained through the SWIR lens are sent to YOLOv4, which is based on the double pyramid structure, for identification. The types of identification focus on human and vehicles, and then the identification results are output.

The well-developed SWIR intelligent identification system will be tested as a whole. In order not to affect the safety of on-site driving and pedestrians, this test is expected to use materials that do not affect the light transmission effect to make a 0.5 m3 smog airtight box, and put the smoke into the airtight box Inside, put the SWIR sensor in the smoke-tight box for testing, place this test system for a long time in places where vehicles and human often appear, collect image recognition data and overall test the development of the integrated software and hardware.This study divides the testing methods into two categories:

1. Different Concentration Tests

It is designed to fill the airtight box with the smoke of burning smoke cake, and set up the concentration detector. The preliminary plan is to record every 5 minutes before burning and after burning, and detect the current concentration value (unit: PPM) to show the recognition effect of different concentrations on SWIR images.

2. Test of human and vehicles at different distances

In order to understand the recognition effect of targets at different distances under the same smoke concentration (unit: PPM), the test system was designed with a distance of 5-40 m, every interval of 5 m, and distances of 100, 150 and 200 m, a total of 11 The test group of the group is tested, and the test scene is shown inFig. 9.

Fig. 9 Schematic diagram of the test scene.

Fig. 10 is the actual test image, comparing the visual conditions of SWIR image and RGB image in a simulated smoke environment. Through the comparison of the experimental results, it can be clearly seen that whether it is 20m or 40m away from the target, the clarity of the SWIR image is better than that of the traditional RGB image.This study uses the comparison method (mAP: Mean Average Precision) commonly used in AI deep learning to evaluate the accuracy.

Table 2 is the comparison table between SWIR images and RGB images at different distances from the target. We can find that when the distance is 15m~30m, the accuracy of the RGB image is close to that of the SWIR image, but as the distance increases, the recognition of the RGB image decreases. Overall, the recognition of SWIR images is better than that of traditional RGB visible light images.

Fig. 10 Tested in smoke at different distances.

Table 2 SWIR and RGB test comparison table at 15m to 200m.

Distance (m)	SWIR – mAP (%)	RGB – mAP (%)
15	95.73	93.33
20	97.90	74.17
25	95.63	76.79
30	97.50	58.86
35	94.25	52.88
40	94.72	57.00
100	80.59	27.78
150	80.16	26.03
200	81.10	13.96

V. CONCLUSION

Aiming at the smog caused by various special factors in the environment, which leads to the inability to effectively identify human and vehicles, causing safety concerns, this paper conducts research on identifying human and vehicles in smog. Therefore, the SWIR sensor is used with the computer to build the AI recognition module for human and vehicles in SWIR images, which can effectively identify human and vehicles to improve safety. This research burns smoke cakes to simulate the smoky environment, and develops a SWIR camera image capture module, which automatically captures images and streams them to the AI interpretation module for analysis. In addition to successfully identifying human and vehicles, it also uses images Fusion combines the target object into the screen marked with a crosshair. In this research, through the self-developed SWIR image AI recognition system, it can display the recognition screen and the crosshair on a small screen to realize the

finished product of the human-machine interface.

ACKNOWLEDGEMENTS

This research was funded by Defense Advanced Technology Research Program. The authors express their gratitude to the Ministry of Defense for sponsoring this research.

REFERENCES

Journal Papers:

[1] A. Bochkovskiy, C.Y. Wang, H.Y. Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection, *Computer Vision and Pattern Recognition*, 2020,arXiv:2004.10934

[2] Lin, Y. C., Chen W. H., and Kuo C.H., Implementation of Pavement Defect Detection System on Edge Computing Platform, *Applied Sciences*, 2021, Vol.11(8), pp. 1-16

[3] Redmon, J., and Farhadi A., YOLO9000:Better, Faster, Stronger, *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 6517–6525

[4] Redmon, J., Santosh D., Girshick R., and Farhadi A., You Only Look Once: Unified, Real-Time Object Detection, *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 779-788

[5] M. Bertozzi, R. I. Fedriga, A. Miron, J. L. Reverchon, Pedestrian Detection in Poor Visibility Conditions: Would SWIR Help, *International Conference on Image Analysis and Processing*, ICIAP, 2013, 229-238

[6] Ren, S., He K., Girshick R., and Sun J., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *Neural Information Processing Systems*, 2016, 1-14

[7] Wang, C. Y., Liao H. Y. M., Yeh I H., Wu Y. H., Chen P. Y., and Hsieh J. W., CSPNet: A New Backbone that can Enhance Learning Capability of CNN, *Computer Vision and Pattern Recognition*, 2019, arXiv: 1911.11929

