sensors

MDPI

Article

Using Deep Learning with Thermal Imaging for Human Detection in Heavy Smoke Scenarios

Pei-Fen Tsai, Chia-Hung Liao and Shyan-Ming Yuan *

Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan;
alicetsai.cs10@nycu.edu.tw (P.-F.T.); aiallen.cs07g@nctu.edu.tw (C.-H.L.)
* Correspondence: smyuan@nycu.edu.tw

Abstract: In this study, we propose using a thermal imaging camera (TIC) with a deep learning model as an intelligent human detection approach during emergency evacuations in a low-visibility smoky fire scenarios. We use low-wavelength infrared (LWIR) images taken by a TIC qualified with the National Fire Protection Association (NFPA) 1801 standards as input to the YOLOv4 model for real-time object detection. The model trained with a single Nvidia GeForce 2070 can achieve >95% precision for the location of people in a low-visibility smoky scenario with 30.1 frames per second (FPS). This real-time result can be reported to control centers as useful information to help provide timely rescue and provide protection to firefighters before entering dangerous smoky fire situations.

Keywords: thermal imaging camera; LWIR; infrared thermal camera; convolutional neural network; evacuation in fire; human detection; smoky fire scene; firefighter protection; human rescue; YOLO; real-time object detection

check for updates

Citation: Tsai, P.-F.; Liao, C.-H.; Yuan, S.-M. Using Deep Learning with Thermal Imaging for Human Detection in Heavy Smoke Scenarios. *Sensors* **2022**, *22*, 5351. https://doi.org/10.3390/s22145351

Academic Editors: Peng Wang, Qijie Wang and Hua Li

Received: 16 June 2022
Accepted: 8 July 2022
Published: 18 July 2022

1. Introduction

Fire is one of the biggest workplace safety threats. It is a hazard that we would want to not only prevent in advance, but also to respond to quickly in terms of emergency response. An effective fire evacuation plan is important to save human lives, protect firefighters, and minimize property loss [1]. We will focus on the application of the YOLOv4 model with a thermal camera following the National Fire Protection Association (NFPA) 1801 [2] standards to detect humans in a smoky fire scene during an emergency evacuation.

It is widely known that the most dangerous factor in a fire scene is smoke and heat, especially smoke. Smoke causes zero visibility and people can die from smoke inhalation during evacuations from buildings or from firefighter rescue [3].

In a fire situation, smoke spreads at a speed of 3–5 m/s, while humans' top movement speed is 0.5 m/s. Where the smoke arrives, the fire will follow. With this in mind, how to locate people and guide them out of a heavy smoke environment will be the key to surviving an evacuation process. Detecting a human in heavy smoke is a challenging task. The most commonly used solutions are laser detection and ranging (LADAR), 3D laser scanning, ultrasonic sensor, or infrared thermal camera [3,4].

Here, we propose to use an infrared thermal camera that follows NFPA1801 standards for good visibility in a smoky fire environment with the YOLOv4 [5] artificial intelligence human detection model as a convolutional neural network (CNN) processed by a single GPU computer to locate humans in a smokey scene, while providing information to the central control room during an emergency evacuation.

2. Background and Related Work

2.1. Thermal Spectrum and Sensor

Objects above absolute zero emit infrared radiation across a spectrum of wavelengths referred to as thermal radiation, as shown in Figure 1. The higher the thermal radiation, the shorter the wavelength in the spectrum.

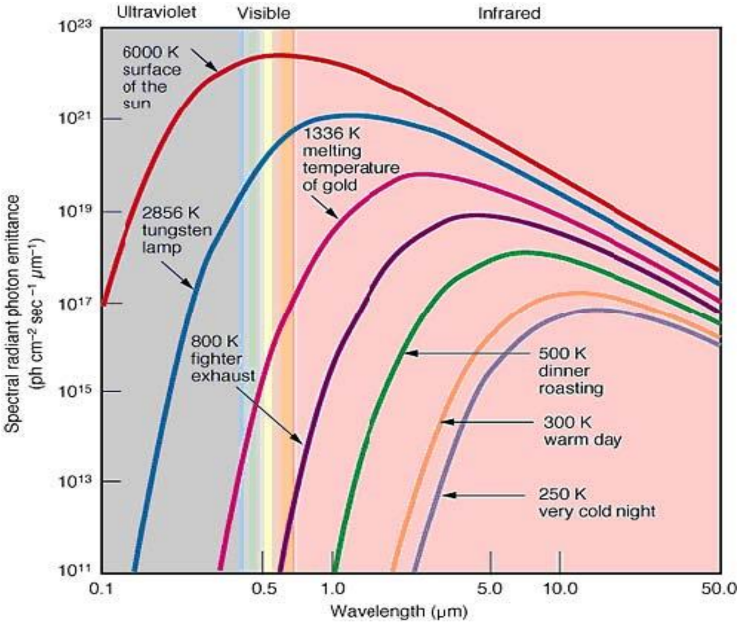


Figure 1. Thermal radiation in different temperatures (Hajebi, 2008; pp. 105–112) [6].

A thermal imaging camera (TIC) converts the thermal radiation into an electric signal to make the invisible infrared (IR) wavelength into a visible RGB or grayscale image. This is a commonly used approach to detect heat sources and to change temperatures into a visible digital image.

IR from 0.7 µm to 15 µm is invisible to the human eye. Different sensor materials can be used to detect corresponding wavelength ranges, as illustrated in Figure 2 [6]. A micro-bolometer is commonly used as a long-wavelength infrared (LWIR) TIC since it does not require low temperature cooling and provides good contrast with human images in both day and night as described in Table 1.

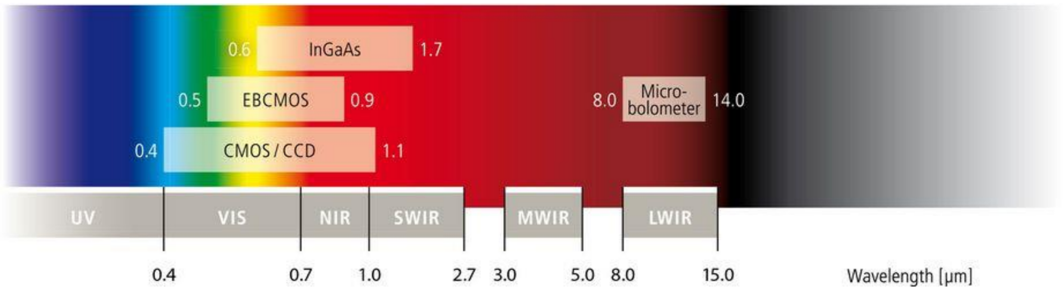


Figure 2. Camera sensor type responding to different wavelengths.

Table 1. Spectrum wavelength range and properties.

Spectrum	Wavelength Range	Property
Visible spectrum	0.4–0.7 µm	Range visible to the human eye.
Near-infrared (NIR)	0.7–1.0 µm	Corresponds to a band of high-atmosphere transmission, yielding the best imaging clarity and resolution.
Short-wave infrared (SWIR)	1–3 µm	
Mid-wave infrared (MWIR)	3–5 µm	Nearly 100% transmission, with the advantage of lower background noise.
Long-wave infrared (LWIR)	8–15 µm	Nearly 100% transmission on the 9–12 µm band. Offers excellent visibility of most terrestrial objects.
Very long-wave infrared (VLWIR)	>15 µm	

2.2. Different Sensor Used in Smoky Fire Scene

A variety of instruments have been studied that can be used in fire and smoky situations, including visible camera, kinetic depth sensor, LIDAR, night vision, IR camera, radar, and sonar. It has been demonstrated that the most efficient instruments are thermal cameras, radar, and LIDAR.

The time-of-flight technology of LIDAR and radar have the best distance measurement accuracy that is suitable for robot navigation [6–9].

Thermal cameras and radar are the best technology to penetrate heavy smoke and are less affected by heavy smoke and high temperatures.

2.3. NFPA1801 Standard of Thermal Imagers

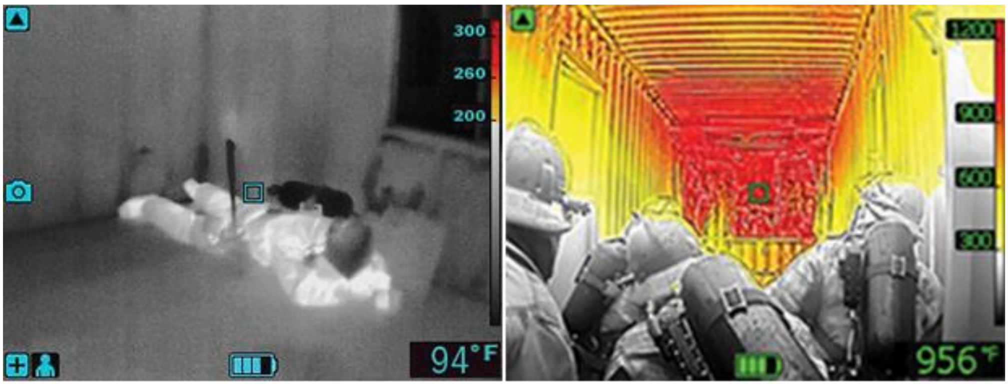
The National Fire Protection Association (NFPA) defines a consensus standard for the design and performance of thermal imaging sensors (TICs) to be used in smoky fire scenarios [10]. The NFPA defines the standard for TICs because the high-temperature and low-visibility conditions are dangerous for firefighters. It defines the criteria for TICs including “interoperability”, “durability”, and “resolution” to make it easy to operate the Tic in a smoky fire scene.

The interoperability is to allow firefighters to operate a TIC with no hesitation during a rescue to save both victims and themselves in a low-visibility scene.

The durability is to enable the TIC to operate in a high-flame or dusty environment with no malfunction.

The most important specifications for TICs on the market for various applications include:

- High resolution (>320 × 240),
- High refresh rate (>25 Hz),
- Wide field of view,
- Low temperature sensitivity (0 °F–650 °F),
- Temperature bar: gray, yellow, orange, and red (Figure 3).

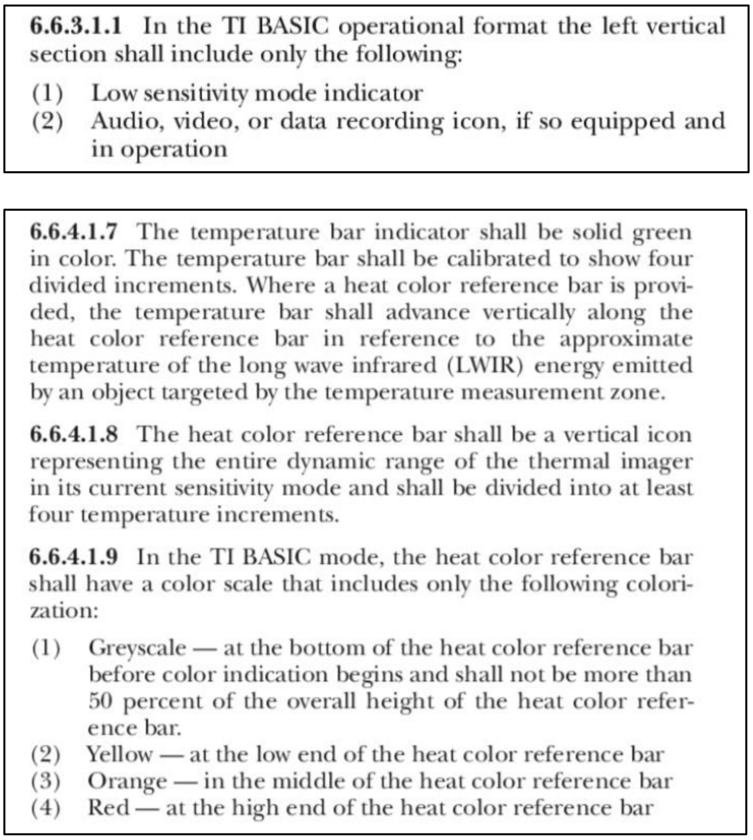


(**a**) (**b**)

Figure 3. Different sensitivity image comparison (**a**) Medium sensitivity: 0 °F–300+ °F (**b**) Low Sensitivity: 0 °F–1200 °F (Source: TechValidate. TVID: B52-065-90D) [11].

Section 6.6.3.1.1 indicates the low sensitivity mode indicator as a basic operational format. The intention of this design is to show high-heat regions in color to enable firefighters to easily distinguish dangerous areas, and to present victims in gray for easy searching, as in Figure 3.

Section 6.6.4 defines the temperature in the color bar to have <50% temperature range in gray and higher temperatures in yellow, orange, and red, respectively, as in Figure 4. TICs that follow NFPA1801(2021) on the market for firefighting purposes commonly define the 25% temperature range in gray, as in Figure 5. In this case, the 25% range for gray will be 162.5 °F (72.5 °C), meaning that humans fall within the grayscale at the 105 gray level, as in Figure 4.



6.6.3.1.1 In the TI BASIC operational format the left vertical section shall include only the following:

(1) Low sensitivity mode indicator
(2) Audio, video, or data recording icon, if so equipped and in operation

6.6.4.1.7 The temperature bar indicator shall be solid green in color. The temperature bar shall be calibrated to show four divided increments. Where a heat color reference bar is provided, the temperature bar shall advance vertically along the heat color reference bar in reference to the approximate temperature of the long wave infrared (LWIR) energy emitted by an object targeted by the temperature measurement zone.

6.6.4.1.8 The heat color reference bar shall be a vertical icon representing the entire dynamic range of the thermal imager in its current sensitivity mode and shall be divided into at least four temperature increments.

6.6.4.1.9 In the TI BASIC mode, the heat color reference bar shall have a color scale that includes only the following colorization:

(1) Greyscale — at the bottom of the heat color reference bar before color indication begins and shall not be more than 50 percent of the overall height of the heat color reference bar.
(2) Yellow — at the low end of the heat color reference bar
(3) Orange — in the middle of the heat color reference bar
(4) Red — at the high end of the heat color reference bar

Figure 4. Definition of NFPA1801 for TIC sensitivity and temperature colorization bar [10].

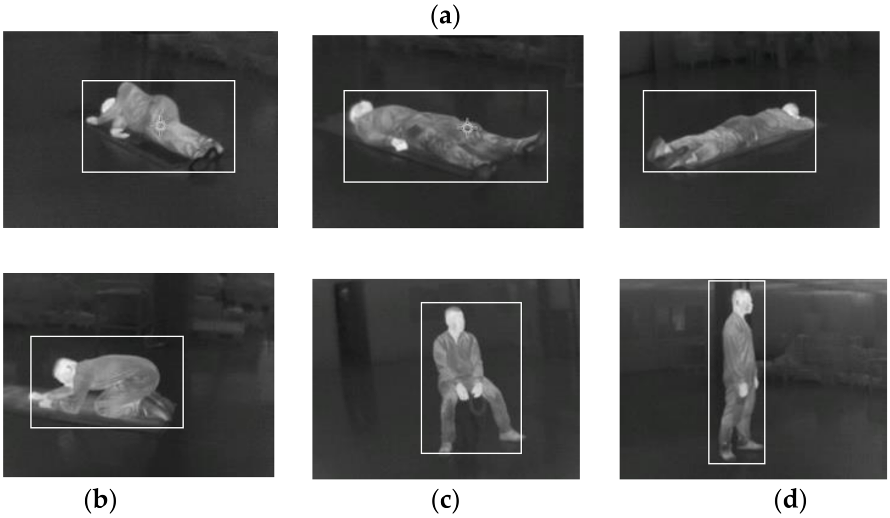


Figure 5. Flue Ti300+ self-taken images to include (**a**) side lying, lying upward, lying downward, (**b**) squatting, (**c**) sitting, and (**d**) standing postures.

NFPA code 1801: Standard on Thermal Imagers for the Fire Service, 2021 edition. In NFPA National Fire Codes Online. Retrieved from http://codesonline.nfpa.org.

2.4. Different Sensors Used in Smoky Fire Scenes

There is an increasing number of papers demonstrating the ability of firefighting robots with stereo infrared vision, radar, or LIDAR to generate fusion images or environmental measurements to identify and locate objects [8,9,12].

In low-visibility smoky scenes, robots can be implemented for environment mapping and indoor navigation. Among the studies on this topic, 3D infrared vision is always implemented into the system that can be used for distance measurement and object detection.

There is related work using FLIR imaging that followed NFPA1801 with a CNN model to help firefighters to navigate in a fire scene [13]. It shows the capability of convolutional neural network models for detecting humans with a thermal imaging camera.

2.5. Convolutional Neural Network (CNN) Object Detection

The task of object detection requires the localization of the objects of interest with coordinates or a bounding box in the image frame.

The CNN model has robust object detection results compared to the traditional computer vision model [14], because of the benefit of the feature extraction of a full image by sliding windows and the combination of multi-scale feature extraction with the increasing depth of network architecture for complex features.

Successful CNN architectures have been proposed recently for object detection tasks, such as R-CNN [15], RFCN [16], Fast R-CNN [17], Faster R-CNN, Mask R-CNN [18], and YOLO [19].

YOLOv4 has been shown to be a fast, real-time object detection technique with high accuracy, with a precision of 43.5% in MS COCO datasets and fast detection of 65 FPS (frame per second) with Nvidia GPU Tesla V100 [5].

There is also evidence of an LWIR sensor with a faster R-CNN model in military surveillance [20] achieving a mean average precision (mAP) of 87% in military operations. In outdoor pedestrian detection, a region-CNN-based model is proposed to achieve an mAP of 59.91% [21].

3. Experiment Methodology

3.1. Model Fluke Ti300+ of Thermal Imaging Camera (TIC) for Data Collection

The model specifications for the Fluke Ti300+ are listed in Table 2. This model was chosen for thermal image data collection since it follows NFPA1801 in terms of resolution, temperature sensitivity, and spectral range with an uncooled microbolometer.

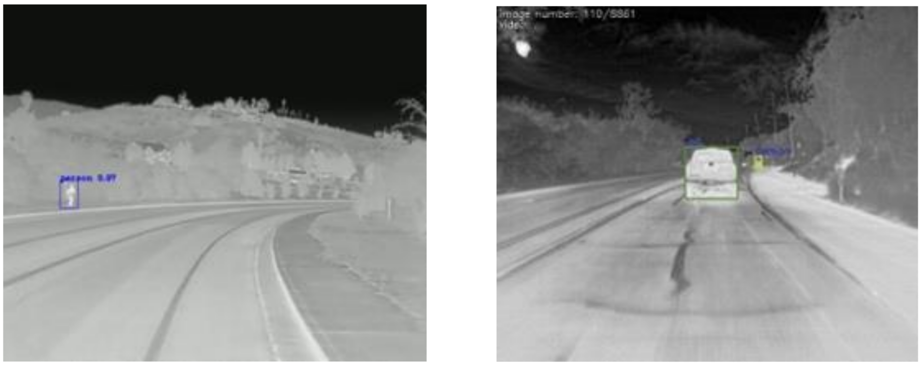
Table 2. Specification of Fluke Ti300+ with IR-fusion technology.

Functionality	Specification
Resolution	320 × 240 pixels
Temperature Range	−20 °C–650 °C (−4 °F–1212 °F)
Spectral Range	7.5 µm–14 µm

We use this TIC to generate extra postures of people falling, lying, and squatting in 360° as training data for the scenario of people requiring help in the event of an evacuation in a smoky environment, as in Figure 5. Human body temperature will correspond to a GL of 105, which is the same gray level as our self-taken human thermal image.

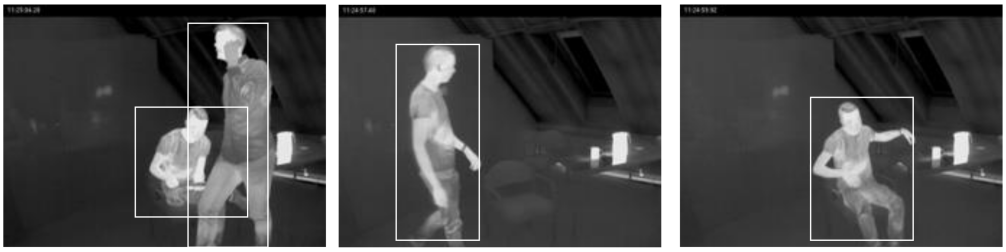
3.2. Thermal Datasets for Training

The 'V's of big data: volume, variety, velocity, and veracity are important factors for deep learning. There is a problem of lack of volume in public thermal image datasets. In this case, we combine our self-taken images, as in Figure 5, and other thermal datasets including pedestrians as in Figure 6, and the Kaggle dataset of indoor people, as in Figure 7, to increase the volume and variety of the training data in Table 3.



(**a**) (**b**)

Figure 6. FLIR ADAS thermal image for walking people and cars on the road (**a**) people walking on the road marked with blue box, (**b**) Car driving on the road with green box. Teledyne FLIR LLC (2022) [22].



(**a**) (**b**) (**c**)

Figure 7. Kaggle AAU thermal indoor image of (**a**) people sitting and standing in one scene, (**b**) walking and (**c**) sitting posture [23].

Table 3. Training dataset.

Datasets	Resolution	Objects in Scene	Main Posture	# of Frames			Environment
				With Person	w/o Person	Total	
FLIR ADAS	640 × 480	People, Car, Bicycle	Walking, Standing	70	830	900	Street
KAGGLE AAU	640 × 512	People, Table,	Sitting	450	0	400	Indoor
FLUKE TI300 SELF-TAKEN	320 × 240	People	Lying, Squatting	450	0	450	Dark Room

3.3. YOLOv4 Model

In [24], it is proven that the YOLO Darknet 2.0 is capable of detecting normal and tiny long-range thermal objects, since thermal images are normally limited to low resolution and tiny objects are only 50 pixels in area. The deep network of Darknet is capable of adding values to thermal image applications in more scenarios.

We use the YOLOv4 [5] model as a CNN-based object detector, which is the highest real-time object detection model in 2020 that fits our scenario of human detection in a hazardous environment.

The model is trained on a single conventional 2080 GeForce GPU as a real-time detector of below the backbone, neck, and head. With deep networks, the backbone is for a different scale of feature extraction, the neck with SPP is for reducing model parameters for training efficiency, and the head of YOLOv3 is for one-stage classification and localization.

- **Input:** Image
- **Backbone:** CSPDarknet53 [25]
- **Neck:**

- Additional blocks: SPP [26]
- Path-aggregation blocks: PAN [27]

- **Head:**
  - Dense Prediction(one-stage): YOLOv3 [28]
- **Selection of BoF and BoS:**
  - Activation: Mish [29]
  - Bounding box regression loss: CIoU
  - Data augmentation: DropOut
  - Normalization: Batch Normalization [30]

The input image resolution is set to 608 × 608 for better detection of small, occluded, or overlapping humans into accurate objects classification and localization.

The pre-trained weight of COCO datasets from the GitHub of AlexeyAB/darknet [31] is used as a good approach to give a good initial position and fast converge within 5000 epochs.

The data augmentation of saturation factor 1.5 is used to cover the different brightness of scenes due to limited available training data volume.

3.4. Optimal Model Selection

We include three datasets of FLIR ADAS, Kaggle AAU TIR image, and Fluke Ti300+ self-taken 360-degree thermal images, as in Figure 7, and with a self-labeled bounding box as the people class. All images are in LWIR (8–15 µm) as it has the best image visibility in the IR spectral range.

The model was trained to 5000 epochs with a learning rate of 0.0014 and a burn-in of 1000, which is less than one day in our system with default hyperparameters, since the thermal human image is of good contrast and extracting features at various scales with 53 layers of network is quite efficient.

Figure 8 shows that there was a very fast loss drop after 1000 epochs and convergence within 5000 epochs.

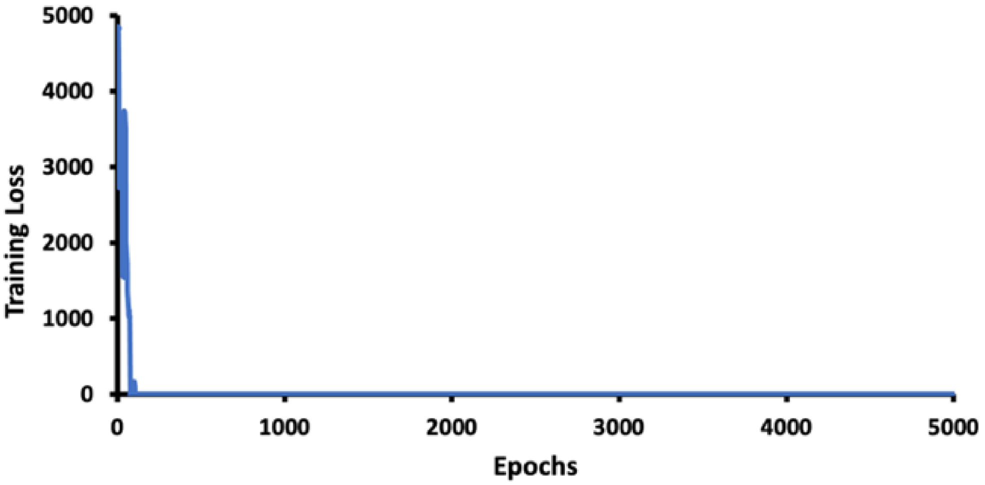


Figure 8. Training loss in epochs.

To prevent overfitting due to the small training data volume, we used validation set accuracy for optimal weights selection at 4000 epochs for our model, as shown in Figure 9. The optimal model can be trained within 24 h on a single GPU to achieve model convergence and to avoid overfitting of the training set.

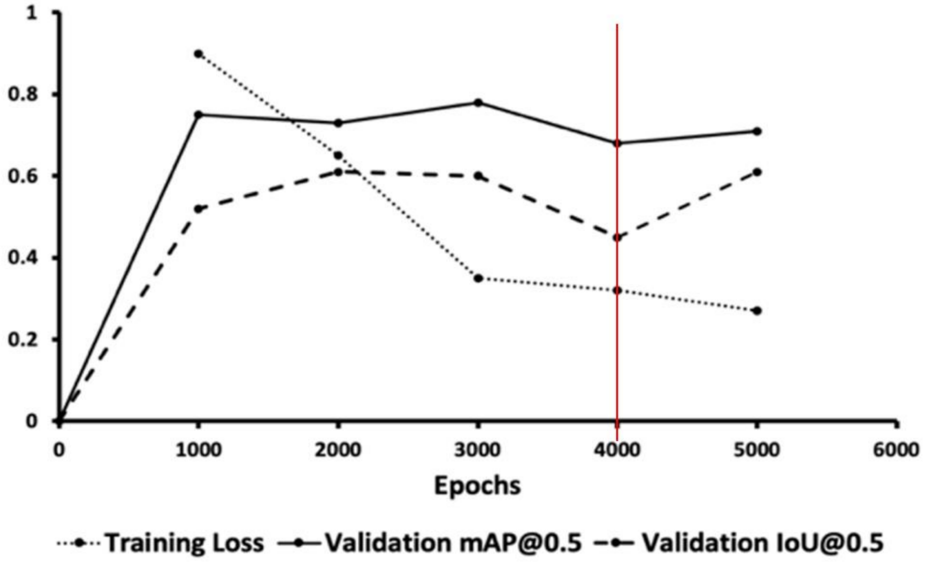


Figure 9. Graph of training loss, validation mAP (mean average precision) and IoU (intersection of union) and red line is the optimal weight with lowest training and validation loss, and highest mAP.

4. Experiment Result

4.1. Ground Truth Calculation of Occluded Objects

In our scenario, if a person is occluded, we still want to detect and count them as one independent object. However, if two people are too close to each other, and one is occluded by the other for over 50%, then this counts as one object.

For instance, as illustrated in Figure 10, the left image will count five people as ground truth (GT), while the right image will count three people as GT, since on the right side, three people are too close together to be treated as individuals.

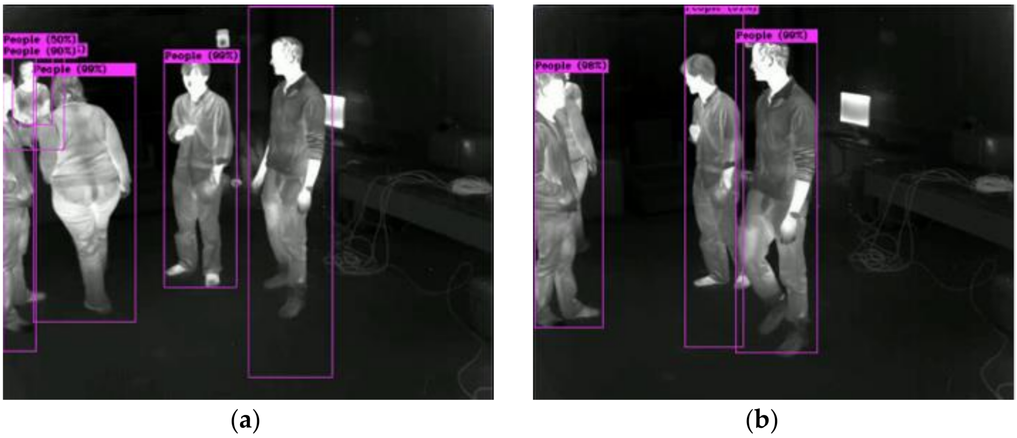


Figure 10. Ground truth for occluded objects: (**a**) five objects and (**b**) three objects.

This ground truth is used for calculating the intersection of union (IoU) of localization precision as true positive (TP) or false negative (FN). We combine the IoU and classification for the object detection index of precision, recall, and PR curve for our model efficiency.

4.2. Metric for Precision, Recall, and Accuracy

We followed the metric of PASCAL VOC for object detection accuracy definition as IoU $\geq$ 50% and correct classification as true positive, IoU < 50% missing detection as false negative, and bounding other objects as human as false positive, as in Table 4.

Table 4. Precision and recall definition.

		Real Label		
		People Labeled	People Not Labeled	
Predicted	People Labeled	TP GT with IoU [1] ≥ 0.5	FP BG is labeled as People	Precision = TP/# Predicted
	People Not Labeled	FN GT with IoU [1] < 0.5	TN BG is not labeled as People	
		Recall = TP/# GT		

[1] IoU: intersection of union.

4.3. Precision, Recall, and PR Curve in Test Datasets

We chose weights at 4000 epochs as the best weights for our model with the lowest training loss, but high accuracy and object localization with the test dataset described in Table 5.

Table 5. Test dataset.

Datasets	**Resolution**	**Objects in Scene**	**Main Posture**	**# of Frames**			**Environment**
				With Person	**w/o Person**	**Total**	
PTB-TIR CLASSROOM	640 × 480	People, Monitor	Walking, Standing	450	103	553	Classroom
KAGGLE AAU	640 × 512	People, Table,	Sitting	100	40	140	Indoor
FLUKE TI300 SELF-TAKEN	320 × 240	People	Lying, Squatting	369	0	369	Dark Room

We achieve a detection of >97% in both precision and recall in Table 6, and the curve in Figure 11 shows an ROC > 98% with various kinds of human postures and viewing angles. This is the power of 53 layers of deep CNN network modeling, resulting in a precise bounding box of >50% IoU.

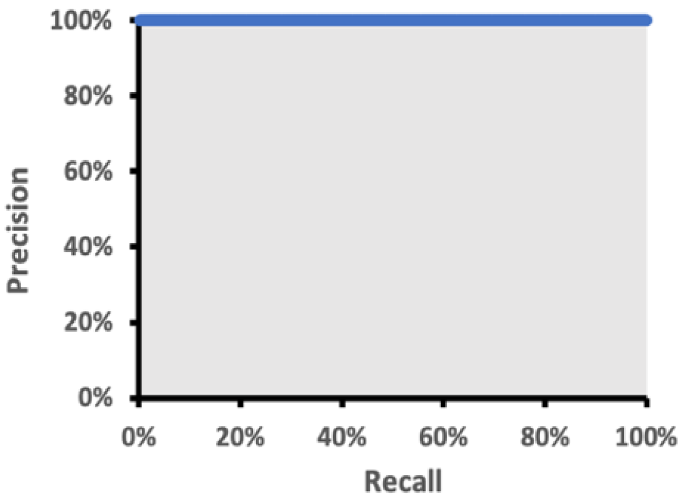


Figure 11. Precision to recall curve.

Table 6. Precision and recall in test dataset.

Property	Standing	Sitting	Lying	Overall
Recall	98.05%	100%	98.64%	98.52%
Precision	100%	100%	100%	100%

4.4. Detection of People in Different Postures

4.4.1. Standing

For the standing posture, our model achieves very high precision, since the FLIR training data have small and large sizes of pedestrians and the PAN implementation in the neck structure for feature aggregation. In Figure 12, the infrared images from CVC [18] demonstrate the ability for large and small human detection. Moreover, the detection of a person who is 50% occluded gives an IoU > 50%, and will be useful to replace human eye searching in the scene.

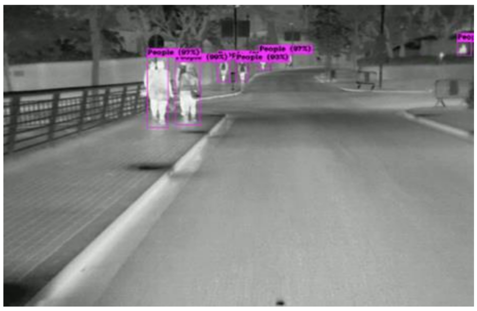


Figure 12. Detection of walking people in CVC dataset.

4.4.2. Sitting

From the testing data from Kaggle AAU Scene 3 in Figure 13 as an indoor sitting test image, our model can detect the ready-to-sit and sitting posture with over 90% confidence; even an occluded sitting human can be detected with 99% confidence. In addition, an image of a sitting person with no face can be detected and labeled correctly with no problems.

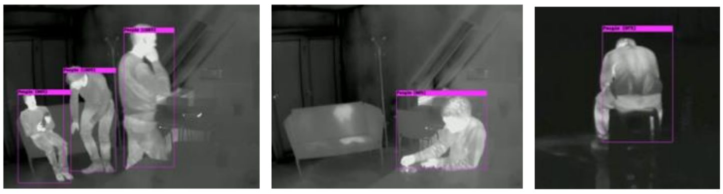


Figure 13. Detection of ready-to-sit, sitting, occluded sitting human, and human sitting backwards.

4.4.3. Lying

A lying posture is the most challenging task, despite being the most important posture that we want to detect in an emergency evacuation scenario. In Figure 14, our model shows >80% confidence of detection, even with the human body occluded.

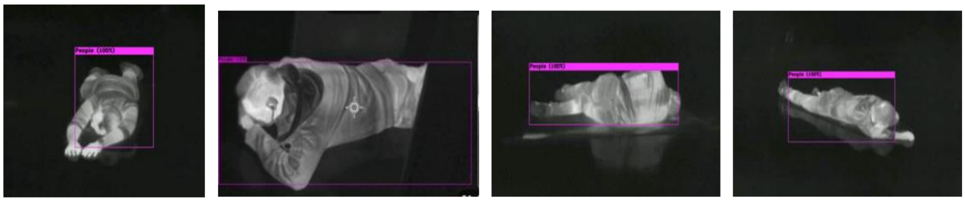


Figure 14. Detection of human lying on the floor from different viewing angles.

In a scenario in which the person has fainted and is lying on the floor, it means they are not able to leave the hazard by themselves and require help from others. People lying down on the floor should be the top priority for rescue.

4.4.4. Squatting

The squatting posture of a human, as shown in Figure 15, including left side, backside, and right side, can be detected successfully, while the confidence level will drop 5% and the bounding box shift 10%. This gives a hint that training datasets are not large enough and need to try some data augmentations or more squatting images to improve the dataset variants.

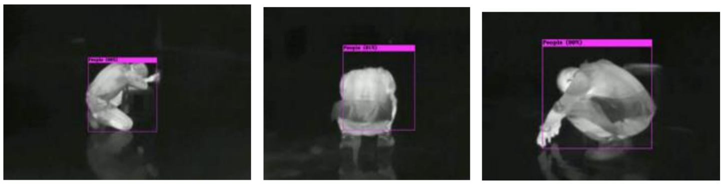


Figure 15. Detection of squatting human from different viewing angles.

4.5. FPS (Frame per Second) for Real-Time Test

The accurate localization and real-time detection of humans are two key factors for rescuing people in danger. This is the reason that we chose one stage object detection of YOLOv4 as our model proposal. In [5], YOLOv4 achieves a state-of-the-art result of 65FPS and 43.5% AP (65.7%AP50) in MS COCO datasets with Tesla V100, published in 2020.

This paper demonstrates the capability for the real-time object detection of a deep learning model on an embedded system of NVIDIA Jetson as a firefighter's protective equipment (PPE) [32]. The benefit of an embedded system is small volume and low power consumption.

We compared the FDS with same YOLOv4 model on an NVIDA GeForce series GPU and Jetson embedded system. On the NVIDIA GeForce RTX2070 with 8GB memory, YOLOv4 achieves 40FPS with an input resolution of 608 × 608, which is a similar FPS to this paper of 55FPS with 416 × 416 input resolution. The Jetson Nano has only 1.1 FPS, as shown in Table 7, which is a similar FPS to the Jetson TX2 in paper [33]. We found that there is still a gap in terms of conducting real-time detection on a Jetson Nano-embedded system.

Table 7. Comparison FPS [1] of Nvidia GeForce and Jetson series.

	Nvidia GeForce RTX 2080Ti	**Nvidia GeForce RTX 2070**	**Nvidia Jetson Nano**	**Nvidia Jetson TX2**
Computation Capability	72 TFLPS [1]	52 TFLPS	0.5 TFLPS	6 TFLPS
Inference Efficiency	40 FPS [2] [33]	30.4 FPS	1.1 FPS	2.1 FPS [33]
Power Consumption	285 W	175 W	5–10 W	10–15 W

[1] FLPS: floating-point operation per second; [2] FPS: frames per second.

5. Discussion

One of the ways to achieve the real-time detection of YOLOv4 on a GeForce GPU will rely on a firefighting robot. There are some studies related to firefighting robots [7,9,12] with thermal imaging cameras to enable robots to build the map and search for people in the building during an evacuation, and send back the information regarding the locations of people to provide timely help.

The other way is to reduce the model size to fit the relatively low computation capability of the embedded system. In this way, the trade-off of a shallower model is less accuracy for human localization.

6. Conclusions

We successfully demonstrate the YOLOv4 deep learning model's capability and benefit for detecting humans in heavy smoke using a thermal imaging camera that follows NFPA1801, with high resolution and low temperature sensitivity, that sufficiently enhances humans as grayscale in high-temperature and low-visibility fire scenes.

Using MS COCO pretrained weights and default parameters, the model can converge within 4000 epochs in one day's training on single Nvidia GeForce 2070 GPU. The accuracy with an IoU of 50% can achieve >95% for the standing, sitting, lying, and squatting postures of humans; even those occluded by 50% can be detected as individual objects. This would be useful for evacuation scenes with people in crowds for counting and searching for people who require help. Real-time detection with 30.1 frames per second is achieved by a trained YOLOv4 model.

In the future, this approach can be combined with robot fusion detection [7,9,12] to allow robots to build the environment map and carry out the search for humans in buildings under evacuation, and send back the information of human locations to provide timely help. The heat source location can also be sent back to the information center to provide useful information before firefighters enter dangerous situations.

Author Contributions: Conceptualization, P.-F.T. and S.-M.Y.; methodology, P.-F.T.; validation, P.-F.T. and C.-H.L.; formal analysis, P.-F.T.; investigation, P.-F.T.; resources, P.-F.T.; data curation, P.-F.T.; writing—original draft preparation, P.-F.T. and C.-H.L.; writing—review and editing, P.-F.T. and C.-H.L.; visualization, P.-F.T.; supervision, S.-M.Y.; project administration, S.-M.Y.; funding acquisition, S.-M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology of Taiwan, grant number 111-2410-H-A49-070-MY2.

Informed Consent Statement: Patient consent was waived due to limited to 1st author only and for indoor room image shooting and no requirement sought for approval.

Data Availability Statement: Not applicable.

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

References

1. Jaradat, F.; Valles, D. A Human Detection Approach for Burning Building Sites Using Deep Learning Techniques. In Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, Las Vegas, NV, USA, 12–14 December 2018; pp. 1434–1435.
2. *NFPA 1801*; Standard on Thermal Imagers for the Fire Service. NFPA: Quincy, MA, USA, 2020.
3. Gaur, A.; Singh, A.; Kumar, A.; Kulkarni, K.S.; Lala, S.; Kapoor, K.; Srivastava, V.; Kumar, A.; Mukhopadhyay, S.C. Fire sensing technologies: A review. *IEEE Sens. J.* **2019**, *19*, 3191–3202. [CrossRef]
4. Aathithya, S.; Kavya, S.; Malavika, J.; Raveena, R.; Durga, E. Detection of Human Existence Using Thermal Imaging for Automated Fire Extinguisher. In *Proceedings of the International Conference on Emerging Current Trends in Computing and Expert Technology, Chennai, India, 22–23 March 2019*; Springer: Chennai, India, 2019; pp. 279–287.
5. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
6. Hajebi, K.; Zelek, J.S. Structure from infrared stereo images. In Proceedings of the 2008 Canadian Conference on Computer and Robot Vision, IEEE, Windsor, ON, Canada, 28–30 May 2008; pp. 105–112.
7. Fritsche, P.; Zeise, B.; Hemme, P.; Wagner, B. Fusion of radar, LiDAR and thermal information for hazard detection in low visibility environments. In Proceedings of the 2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR), IEEE, Shanghai, China, 11–13 October 2017; pp. 96–101.
8. Kim, J.-H.; Lattimer, B.Y. Real-time probabilistic classification of fire and smoke using thermal imagery for intelligent firefighting robot. *Fire Saf. J.* **2015**, *72*, 40–49. [CrossRef]
9. Kim, J.-H.; Starr, J.W.; Lattimer, B.Y. Firefighting robot stereo infrared vision and radar sensor fusion for imaging through smoke. *Fire Technol.* **2015**, *51*, 823–845. [CrossRef]

10. *NFPA 1801*; Standard on Selection, Care, and Maintenance of Open-Circuit Self-Contained Breathing Apparatus (SCBA), 2021 Edition. NFPA National Fire Codes Online. NFPA: Quincy, MA, USA, 2021. Available online: http://codesonline.nfpa.org (accessed on 1 February 2022).
11. Teledyne FLIR LLC. 2022. Available online: https://www.flir.com/instruments/firefighting/ (accessed on 1 February 2020).
12. Kim, J.-H.; Keller, B.; Lattimer, B.Y. Sensor fusion based seek-and-find fire algorithm for intelligent firefighting robot. In Proceedings of the 2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, IEEE, Wollongong, NSW, Australia, 9–12 July 2013; pp. 1482–1486.
13. Bhattarai, M.; Martinez-Ramon, M. A deep learning framework for detection of targets in thermal images to improve firefighting. *IEEE Access* **2020**, *8*, 88308–88321. [CrossRef]
14. Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
15. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
16. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* 2016.
17. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
18. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Breckon, T.P.; Gaszczak, A.; Han, J.; Eichner, M.L.; Barnes, S.E. Multi-modal target detection for autonomous wide area search and surveillance. In Proceedings of the Emerging Technologies in Security and Defence and Quantum Security II and Unmanned Sensor Systems X, International Society for Optics and Photonics, Dresden, Germany, 22 October 2013; Volume 8899, p. 889913.
21. Devaguptapu, C.; Akolekar, N.; Sharma, M.M.; Balasubramanian, V.N. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
22. Teledyne FLIR LLC (2022). FLIR ADAS Dataset. Available online: https://www.flir.in/oem/adas/adas-dataset-form/ (accessed on 1 February 2020).
23. Kaggle AAU VAP Trimodal People Segmentation Dataset. Available online: https://www.kaggle.com/datasets/aalborguniversity/trimodal-people-segmentation (accessed on 1 February 2020).
24. Ghenescu, V.; Barnoviciu, E.; Carata, S.-V.; Ghenescu, M.; Mihaescu, R.; Chindea, M. Object recognition on long range thermal image using state of the art dnn. In Proceedings of the 2018 Conference Grid, Cloud & High Performance Computing in Science (ROLCG), IEEE, Cluj-Napoca, Romania, 17–19 October 2018; pp. 1–4.
25. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. 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, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
27. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
28. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
29. Misra, D. Mish: A self regularized non-monotonic neural activation function. *arXiv* **2019**, arXiv:1908.08681.
30. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 448–456.
31. Simon, M.; Rodner, E.; Denzler, J. Imagenet pre-trained models with batch normalization. *arXiv* **2016**, arXiv:1612.01452.
32. Bhattarai, M.; Jensen-Curtis, A.R.; Martínez-Ramón, M. An embedded deep learning system for augmented reality in firefighting applications. In Proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, Miami, FL, USA, 14–17 December 2020; pp. 1224–1230.
33. Sung, J.-Y.; Yu, S.-B. Real-time Automatic License Plate Recognition System using YOLOv4. In Proceedings of the 2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), IEEE, Seoul, South Korea, 1–3 November 2020; pp. 1–3.