WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n2.03
Journal homepage: https://wepub.org/index.php/IJCSIT/index

UAV Infrared Image Human and Vehicle Object Detection Based on YOLOv5

Leihao Zhao

School of Electronics and Information Engineering at Southwest Minzu University, China

ABSTRACT

Currently, the primary methods for searching for missing persons and vehicles in the wilderness involve extensive manpower for ground-based grid searches. Occasionally, a limited number of helicopters are deployed to assist in the search and rescue operations, and in some cases, UAVs equipped with visible light imaging cameras are used for manual image recognition. However, these search methods are inefficient and require significant human and material resources. To address these issues, this paper proposes a search and rescue method that combines multi-rotor UAVs with image target recognition technology. The YOLOv5-based object detection method is employed to identify humans and vehicles in UAV infrared images, achieving a mean Average Precision (mAPs) of 99%, with a mAP of 98.5% for human recognition and 99.5% for vehicle recognition. This object recognition algorithm model has profound implications for the automation of multi-rotor UAV wilderness search and rescue operations.

KEYWORDS

YOLOv5; Object Detection; Infrared Images; Human Object Recognition; Vehicle Object Recognition

1. INTRODUCTION

With the advancement of urbanization, people have fewer opportunities to interact with natural environments in their daily lives. As a result, an increasing number of individuals are seeking a more natural experience during their leisure time away from familiar surroundings. Outdoor activities, such as forest hiking and off-road driving in uninhabited areas, have become increasingly popular. However, most participants lack professional training and basic outdoor survival knowledge, leading to a significant rise in the number of incidents where individuals go missing in the wild. The complex terrain of natural environments poses a critical challenge to quickly locating distressed individuals, which is essential for initiating rescue operations.

Currently, conventional rescue methods involve extensive ground personnel searches or utilizing UAVs equipped with aerial cameras to capture images for manual identification of potential targets [1]. These methods aim to quickly locate distressed individuals and initiate rescue operations. However, many of the areas where individuals go missing are undeveloped natural environments with complex terrains, making rapid search and rescue operations difficult and posing challenges for rescuers. Conventional rescue methods rely heavily on visual observation, which requires considerable physical endurance and subjective judgment from search personnel. This approach is prone to oversight, especially during prolonged search operations. Additionally, conventional rescue methods require substantial human, material, and financial resources, and are inefficient.

This paper investigates human and vehicle target detection technology in UAV infrared images [2], aiming to replace visual observation with automated recognition. This approach allows for the quick

and accurate identification of human and vehicle targets, overcoming the issue of missed detections due to rescuers' subjective factors during search and rescue operations. Infrared image recognition also mitigates the problem of insufficient lighting [3], facilitating nighttime search and rescue operations. By integrating image recognition technology with UAV autonomous flight control technology [4], UAVs can autonomously search for missing individuals in the wilderness [5]. This approach addresses the deficiencies of traditional search methods, enabling the efficient acquisition of the geographical locations of missing targets and providing valuable time for subsequent rescue efforts.

2. RELATED WORK

Object detection and recognition is a fundamental topic in the field of image processing [6]. Therefore, numerous studies have been conducted by scholars both domestically and internationally, and these methods have been applied in practice. However, over the past 20 years, a significant amount of research on human target detection has focused on visible light images. Current research on human and vehicle target detection in UAV video images based on visible light has achieved certain results [7]. However, due to the imaging principles and characteristics of thermal infrared images, existing algorithms often fail to achieve the desired outcomes.

Renowned international universities and research institutions such as Stanford University, Cambridge University, and Google's AI Research Center have dedicated groups for thermal infrared image processing, making significant progress in thermal infrared image target recognition. Meanwhile, Chinese universities have also made positive strides in this field. Institutions such as Nanjing University of Aeronautics and Astronautics, Tsinghua University, Shanghai Jiao Tong University, and the University of Electronic Science and Technology of China have proposed numerous high-quality recognition algorithms, greatly contributing to the development of thermal infrared image recognition. Among the existing thermal infrared target recognition algorithms, the most representative ones are filter-based methods, human visual system-based methods, and deep learning-based methods.

Filter-based methods primarily process the input image using filters and then set a threshold to identify regions where pixel values exceed this threshold as target objects. These methods are simple in terms of algorithms and computationally efficient, but they perform poorly and are only suitable for specific scenarios, failing to address thermal infrared target recognition in complex backgrounds. Human visual system-based methods, inspired by biological vision systems, first extract local feature information from images and then use component analysis or mathematical transformations to merge these local pieces into global features, thereby identifying target objects as salient regions. The key to these algorithms is the acquisition of salient maps. Utilizing salient maps as tools, human visual system-based methods show significant promise and value in enhancing the speed and simplifying the process of thermal infrared image target recognition, warranting further exploration and research.

With the rapid development of deep learning, a plethora of deep learning-based object detection models have emerged, typically categorized into one-stage and two-stage detection models. Representative one-stage models include YOLO, RetinaNet, and FCOS [8][9][10], while two-stage models include R-CNN, Faster R-CNN, R-FCN, and Mask R-CNN [11][12][13][14]. Many scholars have applied deep learning methods to infrared image target detection with notable success. Minglei Li and Xingke Zhao used BASNet to extract salient maps from thermal infrared images, obtaining clear boundary target images and replacing RGB channels with these salient maps. They employed the lightweight network MobileNetv2 to replace the traditional feature extraction module in YOLOv3, resulting in the "ComNet" model, which achieved excellent performance in recognizing humans and vehicles in infrared images [15]. Xinyu Jia et al. utilized a binocular camera to capture images and obtain infrared and visible light images with the same spatial-temporal and field of view by cropping. They labeled pedestrians in the images to establish a fusion dataset and used infrared images to

address pedestrian recognition under low-light conditions. This approach validated the significant role of infrared images in pedestrian recognition under weak light conditions [3].

3. METHOD

Our object detection algorithm utilizes YOLOv5s for detecting human and vehicle targets in infrared images. YOLOv5 is an anchor-based detection method belonging to the single-stage object detection category. Compared to YOLOv4, YOLOv5 offers faster speed and higher accuracy, making it one of the leading object detection algorithms in the industry.

YOLOv5 is based on the one-stage approach in object detection algorithms. Its primary concept is to divide the entire image into multiple grids, with each grid predicting the type and location of objects within it. The predictions are then filtered based on the IoU value between the predicted boxes and the ground truth boxes, ultimately outputting the class and location information of the predicted boxes. This method is characterized by high efficiency, high accuracy, and strong usability.

YOLOv5 comprises five versions: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. YOLOv5n is the smallest version, while YOLOv5x is the largest. The differences among them lie in the depth, width, and number of parameters of the network. We chose the YOLOv5s version due to its lower parameter count, high precision, and faster detection speed.

The YOLOv5s model is primarily composed of three parts: Backbone, Neck, and Head. The network architecture is illustrated in Figure 3.1. The Backbone is responsible for feature extraction from the input image. Its main function is to transform the original input image into multiple layers of feature maps for subsequent object detection tasks. YOLOv5 uses either the CSPDarknet53 or ResNet backbone networks, which are relatively lightweight and ensure high detection accuracy while minimizing computational load and memory usage. The main structures in the Backbone include the Conv module, C3 module, and SPPF module.

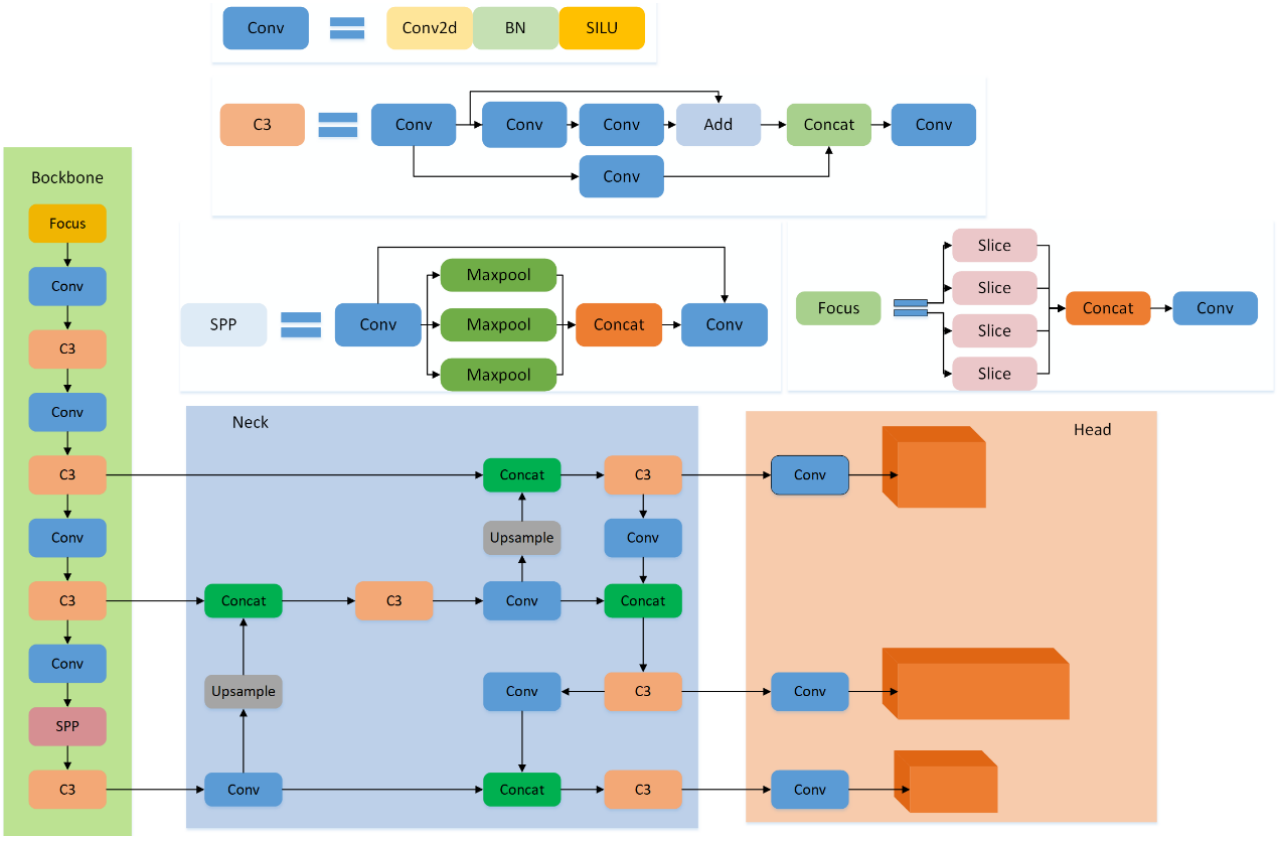

Figure 3.1. YOLOv5 Network

The Neck is responsible for multi-scale feature fusion and passing these features to the prediction layers. Since the size and position of objects in an image are uncertain, a mechanism is needed to handle targets of different scales and sizes. The feature pyramid is a technique used for multi-scale object detection, which can be implemented by adding feature layers of different scales to the backbone network. In YOLOv5, the FPN (Feature Pyramid Network) structure is used. This structure combines different levels of feature maps through upsampling and downsampling operations to create a multi-scale feature pyramid. The top-down part mainly achieves feature fusion of different levels through upsampling and merging with coarser feature maps, while the bottom-up part merges feature maps from different levels using a convolutional layer. In object detection algorithms, the Neck module is usually used to combine feature maps from different levels to generate multi-scale feature maps, thereby improving the accuracy of object detection. In YOLOv5, a feature fusion module called PANet is used as the Neck module.

The Head performs the final regression prediction. The object detection head is the part used for object detection on the feature pyramid. It includes several convolutional layers, pooling layers, and fully connected layers. In the YOLOv5 model, the detection head module is primarily responsible for multi-scale object detection on the feature maps extracted by the backbone network. This module mainly includes three parts: Anchors, Classification, and Regression. Anchors are used to define object boxes of different sizes and aspect ratios. They are usually obtained by clustering the target boxes of the training set using K-means clustering, which can be calculated before model training and stored in the model for generating detection boxes during prediction. Classification is used to classify each detection box, determining whether it is a target object, usually employing a fully connected layer with a Softmax function for classification. Regression is used to regress each detection box to obtain its position and size, usually employing a fully connected layer for regression. Additionally, YOLOv5 uses several techniques to further enhance detection accuracy, such as GIoU loss, Mish activation function, and multi-scale training.

4. EXPERIMENTS

4.1. Dataset

In this study, we utilized the publicly available Infrared Person Vehicle Detection Images dataset for our experiments [15]. This dataset was collected using a DJI M600 PRO drone equipped with a FLIR thermal infrared camera Vue Pro. The drone's flight altitude ranged from 20 meters to 40 meters, and the image resolution was 640×512 pixels. The original images captured by the thermal infrared camera were single-channel grayscale images representing brightness levels. For ease of pedestrian and vehicle detection research, the received thermal infrared images were converted to RGB format three-channel pseudo-color images through temperature mapping. In the single-channel grayscale images, a pixel value of 0 was mapped to blue, a pixel value of 255 was mapped to red, and the values in between were smoothly interpolated, using color temperature to indicate low and high temperature areas.

The dataset includes outdoor thermal infrared images collected during both daytime and nighttime, covering various scenes such as playgrounds, highways, and squares. The training set comprises 2434 images, containing 3555 pedestrian instances and 3189 vehicle instances. The test set consists of 271 images with 618 pedestrian instances. The validation set includes 270 images with 595 pedestrian instances and 667 vehicle instances. An image annotation tool, Labelme, was used to annotate the bounding boxes of the targets. An example of the annotated images is shown in Figure 4.1.

Figure 4.1. Example of Dataset Annotations

4.2. Model Training

We implemented our model using the Pytorch framework. The model training was conducted using four NVIDIA GeForce GTX 1080Ti GPUs, and the model testing was performed on a single NVIDIA GeForce GTX 1080Ti GPU. The server for model training ran the Linux operating system Ubuntu 18.04.6 LTS, with CUDA version 12.1, Python environment 3.8.5, and Pytorch version 1.8.0. The training process involved 100 epochs with a batch size of 40.

The loss in YOLOv5 is primarily composed of three parts: Classes loss, Objectness loss, and Location loss. The Classes loss, which is the classification loss, uses BCE loss and is calculated only for positive samples. During training and validation, this is referred to as cls_loss. The Objectness loss, which pertains to the detection of objects, also uses BCE loss and is calculated for all samples, named obj_loss in the training and validation phases. The Location loss, which is the localization loss, employs CIoU loss and is calculated only for positive samples, referred to as box_loss during training and validation. The decline curves of the training loss functions for the training and validation sets are shown in Figure 4.2.

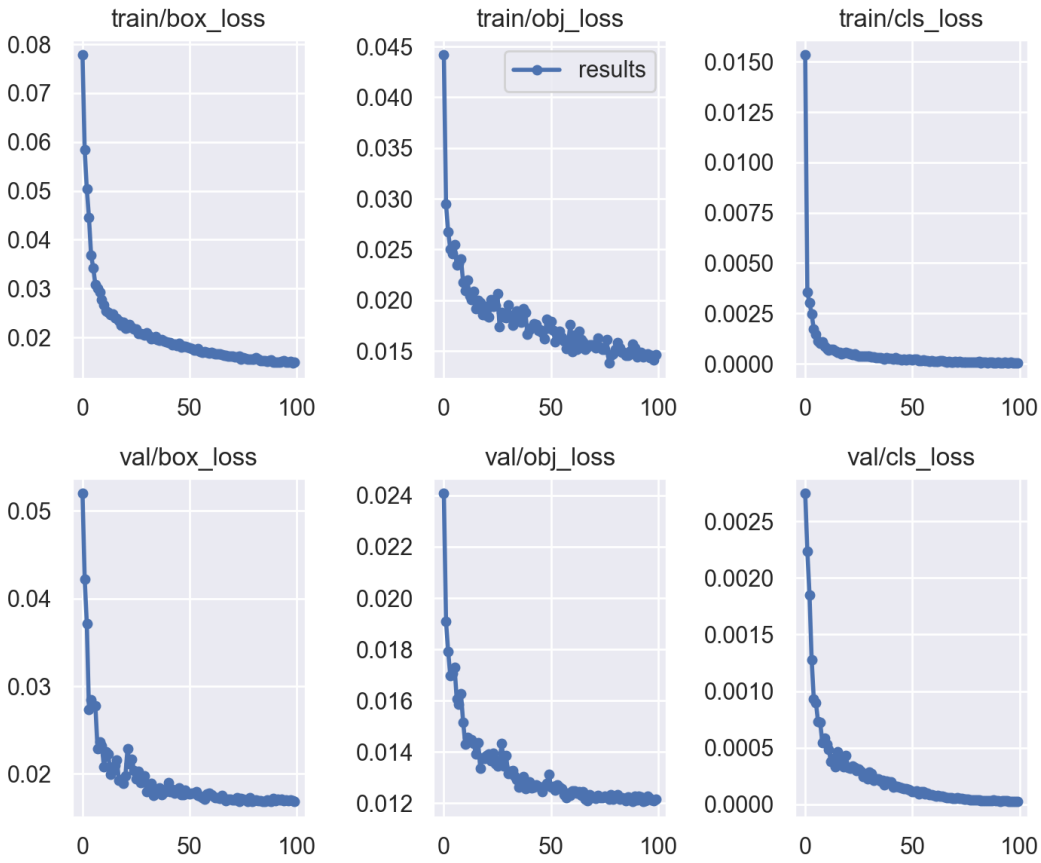

Figure 4.2. Decline Curves of Training Loss Functions

As shown in the figure above, after 100 epochs of training, the localization loss, object detection loss, and classification loss for both the training set and validation set each converge to a fixed value, achieving the training objective.

4.3. Experimental Metrics

In this experiment, we utilized precision (P), recall (R), and mean Average Precision (mAP) as evaluation metrics. The calculation formulas are as follows,where TP represents the number of true positives, FP represents the number of false positives, and FN represents the number of false negatives.

After 100 epochs of training, the experimental results showed a precision of 99.29%, a recall of 97.83%, a mean Average Precision (IoU=0.5) of 98.93%, and a mean Average Precision (IoU=0.5:0.95) of 81.49%. The curves depicting the changes in precision, recall, and mean Average Precision during the training process are shown in Figure 4.3.

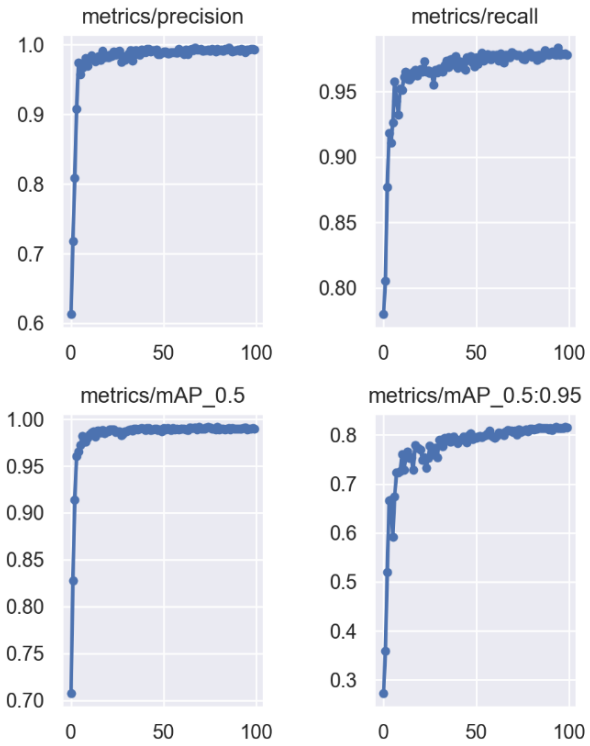

Figure 4.3. Curves of Evaluation Metrics

4.4. Experimental Analysis

Using mAP as the evaluation metric, we compared our detection results with those obtained by other methods. The comparison includes results from experiments conducted by the authors using YOLOv3 and YOLOv3-MobileNetv2 combined with image enhancement techniques on the same dataset. The comparison results are shown in Table 4.1. YOLOv3-MobileNetv2 is an improved version of the YOLOv3 network, utilizing an efficient depthwise separable convolution method to enhance computational speed and incorporating inverse residuals and linear bottleneck modules for more efficient basic modules.

Table 4.1. Comparison of Detection Results

Image enhancement method	Models	mAP(Pedestrian)	mAP(Vehicle)
\	YOLOv5	0.985	0.995
\	YOLOv3	0.836	0.873
\	YOLOv3-MobileNetv2	0.792	0.826
Saliency maps	YOLOv3	0.771	0.820
Saliency maps	YOLOv3-MobileNetv2	0.719	0.761
Fusion B-G-S	YOLOv3	0.927	0.932
Fusion B-G-S	YOLOv3-MobileNetv2	0.880	0.889
Fusion R-B-S	YOLOv3	0.938	0.956
Fusion R-B-S	YOLOv3-MobileNetv2	0.881	0.899
Fusion R-G-S	YOLOv3	0.905	0.972
Fusion R-G-S	YOLOv3-MobileNetv2	0.857	0.925
Weighted fusion	YOLOv3	0.944	0.978
Weighted fusion	YOLOv3-MobileNetv2	0.903	0.930

From the comparative analysis in the table, it is evident that the mAP achieved using the YOLOv5 model is significantly higher than that obtained by other methods. Additionally, we selected some

images to test the model's target detection performance, with results shown in Figure 4.4. Despite the complexities such as occlusion by branches and incomplete body captures, the model successfully identified and framed the targets with high confidence, providing very high prediction probabilities. This demonstrates the model's robust capability in identifying human and vehicle targets in infrared images.

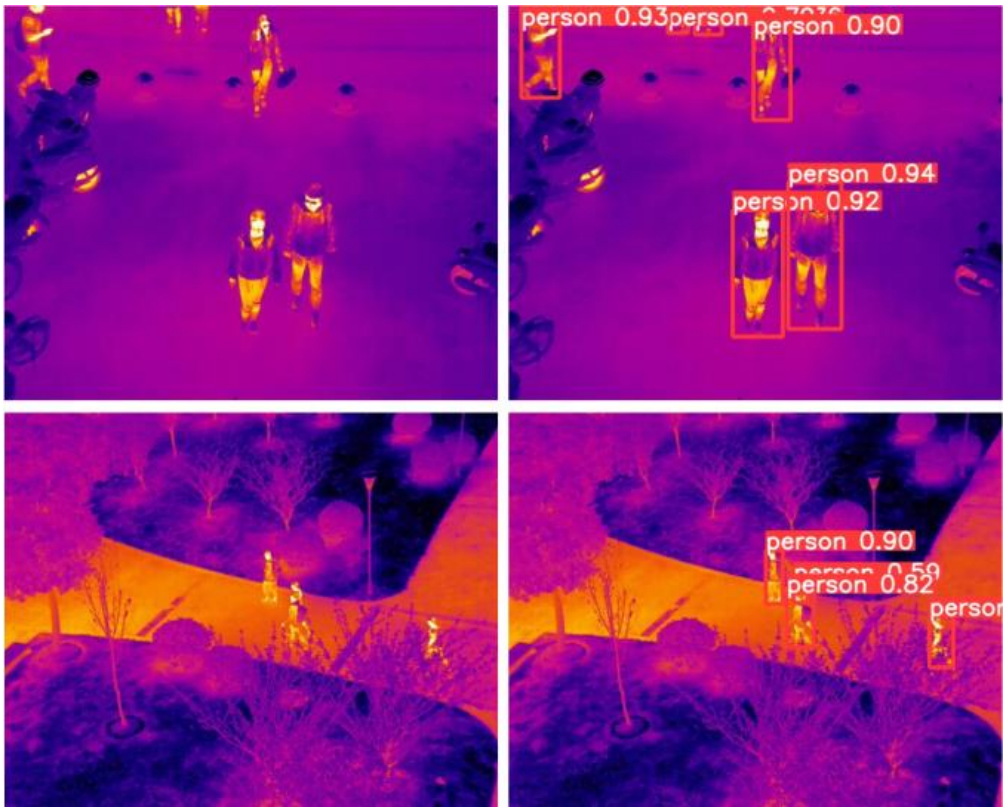

Figure 4.4. Human Target Detection Test Images

5. CONCLUSION

Our model achieved commendable results in the experimental environment, demonstrating the efficacy of YOLOv5 in detecting human and vehicle targets in UAV infrared images. This target recognition algorithm model holds significant promise for the automation of field rescue missions using multirotor UAVs. Moving forward, we plan to optimize the model for lighter weight and test its detection speed, ultimately deploying it on UAVs for practical testing. Given the complex environments encountered in field rescues, the scenes in our dataset might be insufficient. In deep learning, the quality of the model is often constrained by the training dataset. Therefore, we will also work on further refining and expanding the dataset.

ACKNOWLEDGEMENTS

We extend our gratitude for the financial support provided by the Southwest Minzu University Graduate Innovation Research Project No. 320022450144. We also thank the Artificial Intelligence Laboratory of the School of Electronics and Information Engineering at Southwest Minzu University for providing the GPU equipment.

REFERENCES

[1] Fan, Bangkui, et al. "Review on the technological development and application of UAV systems." Chinese Journal of Electronics 29.2 (2020): 199-207.

[2] Portmann, Jan, et al. "People detection and tracking from aerial thermal views." 2014 IEEE international conference on robotics and automation (ICRA). IEEE, 2014.

[3] Jia, Xinyu, et al. "LLVIP: A visible-infrared paired dataset for low-light vision." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[4] Gündoğan, Meryem Mine, et al. "IR Reasoner: Real-time Infrared Object Detection by Visual Reasoning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[5] Bravo, Raissa Zurli Bittencourt, Adriana Leiras, and Fernando Luiz Cyrino Oliveira. "The use of UAVs in humanitarian relief: An application of POMDP-based methodology for finding victims." Production and Operations Management 28.2 (2019): 421-440.

[6] Zou, Zhengxia, et al. "Object detection in 20 years: A survey." Proceedings of the IEEE 111.3 (2023): 257-276.

[7] Maes, W., A. Huete, and K. Steppe. "Optimizing the processing of UAV-based thermal imagery, Remote Sens., 9, 476." (2017).

[8] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[9] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.

[10] Tian, Zhi, et al. "FCOS: A simple and strong anchor-free object detector." IEEE Transactions on Pattern Analysis and Machine Intelligence 44.4 (2020): 1922-1933.

[11] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[12] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." IEEE transactions on pattern analysis and machine intelligence 39.6 (2016): 1137-1149.

[13] Dai, Jifeng, et al. "R-fcn: Object detection via region-based fully convolutional networks." Advances in neural information processing systems 29 (2016).

[14] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[15] Li, Minglei, et al. "ComNet: Combinational neural network for object detection in UAV-borne thermal images." IEEE Transactions on Geoscience and Remote Sensing 59.8 (2020): 6662-6673.