A Comprehensive Study on Deep Learning Architectures for Robust Object Identification in UAV-based Thermal Imaging

Sadman Sadik Khan[1*], Mahtab Chowdhury[1], Md Shahriar Mannan Prottoy[1], Kazi Zakiul Haque[1], Al Momit[1] and Tasnuva Arfin Janisa[1]

[1] Department of Computer Science & Engineering, Daffodil International University, Dhaka, Bangladesh

sadman15-13696@diu.edu.bd*, {chowdhury22205101076, prottoy2205101490, kazi15-4101, momit22205101083, janisa22205101098}@diu.edu.bd

Abstract. In recent years, various industries have been experiencing a notable increase in the use of machine learning for object recognition, resulting in the development of different methods and technologies. Thermal detection technology is one such aspect that uses thermal images captured from objects to determine them based on their heat signatures. This has been particularly useful in situations like surveillance, search and rescue missions, and nighttime operations where conventional imaging techniques are limited by poor visibility or darkness. The advent of unmanned aerial vehicles (UAVs) has even widened the scope of practical application of thermal imaging. In situations such as monitoring wildfires or carrying out search and rescue operations, UAVs fitted with thermographic cameras can cover large areas within a short time. The combination of UAVs and thermal detection technology has greatly improved thermal data acquisition expanding its reach into fields such as disaster relief, animal tracking among others including industry inspections. In this study, we used the HIT-UAV dataset which has much popularity in object identification research to investigate the efficiency of the latest YOLOv8 model more accurately the YOLOv8m and YOLOv8s by comparing these models to its previous version YOLOv5 models. Through our implementation of YOLOv8m on the HIT-UAV dataset we were able to achieve precision of 87.6% and mean average score of 81.9% and with YOLOv8s we were able to achieve precision of 89.1% and mean average score of 82.3%.

Keywords: Computer Vision, Deep Learning, Object Detection, YOLO, UAV thermal Image.

1 Introduction

The introduction and utilization of unmanned aerial vehicles (UAVs) and thermal imaging has transformed multiple industries, one frequent use of this combination is for military reconnaissance. Unfortunately accurate object detection in a variety of

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_44

environmental situations is one of the main issues in object detection but using infrared thermal data has shown to be a potential option in this area. The significance of infrared thermal datasets in UAV-assisted object detection is the main focus of this paper.

Feature detection and classification of objects are necessary for feature extraction therefore the method is suitable for applications such as video tracking, motion detection and 3D object recognition. There are existing simple methods that depend on intensity based features but it is well-known that maybe the similar advanced methods will be needed for proper feature recognition in thermal images in comparison with regular ones [1]. Detecting people's thermal characteristics is a complex task. Precise and fast detection is still a major problem, even with recent improvements in visual and thermal imagery pedestrian detection [2]. Regardless, thermal images are still better than other forms of images as by detecting the thermal energy that objects release, it easier to distinguish and identify objects in a variety of lighting and weather scenarios [3]. Furthermore, Intelligent video surveillance and driver assistance systems are becoming more interested in automatic pedestrian recognition through thermal infrared imagery.[4]. In recent times the You Only Look Once approach or most famously known as the YOLO method has gained popularity recently for detection, object detection and object recognition with drones or UAVs, with multiples researches showing it outperforming other methods in the UAV object detection such as DPM and R-CNN methods because of it real time detection accuracy, we have also seen YOLO being combined with other models such as YOLO-VIT but that makes the architecture more complex [5][6][8][9].

YOLO is a one-shot object that uses a complete modular neural network to process images simultaneously, enabling real-time recognition. Although it may not be accurate for small features, its end-to-end processing and simultaneous predictions of bounding box and class probabilities have greatly improved feature recognition and this implementation has led to a variety of versions [10]. YOLO started with YOLOv1, which focused on speed over accuracy. YOLOv2 optimized architecture and data development to analyze over 9000 studies. YOLOv3 improved small feature detection using a feature pyramid network and Darknet-53. YOLOv4 further increased speed and accuracy by using SPP, CSP-Darknet53, and Path-Aggregation Network. YOLOv5 uses PAnet and export formats. YOLOv6, EfficientRep increased performance for industrial applications. YOLOv7 improves speed and accuracy in the EELAN system. The latest, YOLOv8, released by Ultralytics in 2023, features an anchorless model, fast NMS, and mosaic development, enabling it to achieve 53.9% AP in the MS-COCO test-dev 2017 split [11][12][13].

In earlier works, the SSD, RetinaNet showed better performance than YOLOv3 [14] and SSD-512, achieved better accuracy on the HIT-UAV dataset than YOLOv4 and Faster-RCNN [15]. Furthermore, even though it was quicker but YOLOv4 had trouble handling little objects [16]. This study proposes the implementation of the latest entry of YOLO, the YOLOv8 model, to showcase YOLO’s most recent improvements in accuracy and speed for object detection in UAV imagery.

2 Literature Review

Object detection using thermal images is pivotal for security and surveillance applications. This paper explores the latest developments in utilizing deep learning methods to improve detection precision in low-light environments. We identify promising directions in this important topic by reviewing the body of existing literature.

According to Alexander et al [1], two object detection algorithms: those that include scanning and filtering, and those that involve object detection itself. Algorithms for scanning and filtering can accurately identify items in photos with complicated backgrounds, irrespective of their sizes and shapes. Li et al [2] show an improved Histogram of Oriented Gradient (HOG) algorithm for human detection in thermal images. For binary classification, they fed the resulting vectors into a linear Support Vector Machine (SVM).

Teju et al [3] introduced the necessity for accurate human identification for security sectors and brought up the shortcomings of existing biometric methods, such as facial recognition, especially when it comes to managing fluctuations in illumination. The suggested solution aims for a multi-system that puts together thermal and regular imaging. Non-linear image transfer functions are also used to improve the image quality, and a hybrid algorithm is applied for object detection and classification which improves accuracy. The OFSA algorithm is the main focus of this hybrid algorithm. Another method by San-Biagio et al [4] for object detection is segmentation. Using this method, background regions are removed until the civilian is found by dividing the image recursively,and pedestrians and non-pedestrians are then distinguished by classification.

Redmon et al [5] were the first to show a completely new way for object identification based on the YOLO model. In a single evaluation, his method uses a single neural network to predict bounding boxes and class probabilities straight from the full image. Notably, YOLO outperforms R-CNN in real-time image processing, achieving 45 frames per second but localization problems and geographical restrictions that restrict the amount of adjacent items this model can identify are the YOLO model's drawbacks. Shen et al [6] showed his object detection approach for the detection of flames in different situations.

Gomez et al [7] proposed the use of LWIR thermal images which could be used in public places to count the number of people in that place for example, a beach, it is based on small CNNs that can be run on a limited-memory low-power platform. Zhao et al [8] suggested YOLO-ViT, a YOLOv7-based technique that allows UAVs to detect infrared vehicle targets. It uses C3-PANet with CARAFE upsampling for improved recognition accuracy, incorporates MobileViT for feature extraction, and uses K-means++ clustering to optimise anchor box sizes. Based on the HIT-UAV dataset, the experimental results demonstrate improvement in mean average accuracy (mAP).

Jia et at[9] suggested a method for detecting forest fires using photos taken by UAVs and processed using the YOLO family of technologies. The YOLOv8, YOLOv7, and

YOLOv5 models are thoroughly compared, with the conclusion being that YOLOv8 gives the optimum balance between speed and accuracy. They used the nano-sized YOLOv8 algorithm which showed excellent fire detection accuracy and presents a viable way to reduce the harm caused by forest fires. Etik et at[11] did a comparison of YOLO models using the DOTA dataset. Their research analyzes YOLOv2 and YOLOv3's performance, where YOLOv2 performed better in 5 categories but for speed and small object detection, YOLOv3 is superior.

Ippalapally et at [20] showed in their study, object recognition models for thermal images are compared, and their performance is evaluated using the COCO metric and the FLIR dataset. Fang et al [21] showed the use of Tinier-YOLO; it contains fire modules and dense connections intended to increase both the detection accuracy and real-time functionality. However, it achieves a smaller weight compared with its predecessors that propel newer generations of embedded devices faster.

3 Methodology

The proposed methodology for object identification in UAV thermal imaging follows these stages: dataset collection, image preprocessing, model selection, experimental design and model training, and evaluation stages. The system utilizes YOLOv5 and YOLOv8 deep learning models to extract features for object identification.

3.1 Data Collection

The HIT-UAV dataset [17] contains 2898 infrared thermal images extracted from 43470 frames, captured by UAV from different scenes such as schools, parking lots, roads, etc. The original data was stored in two formats: COCO and PASCAL VOC, but in this version, they have transformed the data into the YOLO format. The dataset has five classes namely, Person, Car, Bicycle, Other Vehicle and Don't Care. There are a total of 2866 images with each image having labels.

Fig. 1. Sample of thermal Images in the dataset

3.2 Data Preprocessing

For data processing to get the training images and associated labels from the dataset's training, validation, and testing directories, the preprocessing steps included loading them. A custom function was developed resulting in an image file path gathering for each file. These coordinates were changed into standard bounding box format that ensures visualization of images and further processing of data for analysis. In addition to this, image-label pairs were shuffled so that randomness is assured during training. Finally, a bounding box was drawn around a subset of the training data used so that the accuracy of the annotations could be verified. Therefore, these preprocessing steps led to a dataset that was good enough for model training and evaluation.

3.3 Model Selection

Yolov8: YOLOv8, also known as You Only Look Once version 8, is the fastest and most accurate object detection model that is highly suitable for any real-time application. The efficient training and advanced network design have contributed to its superiority over the previous versions of YOLO. In this sense, YOLOv8 is especially appropriate since it focuses on thermal object detection using Unmanned Aerial Vehicles (UAVs) due to its real-time monitoring aspect. Moreover, the main advantage of multi-scale feature extraction offered by YOLOv8 is that it allows for to detection of details in thermal images where objects are faint or completely invisible. It is very effective for tasks like surveillance, search-and-rescue operations, and monitoring with helicopters featuring infrared sensors.

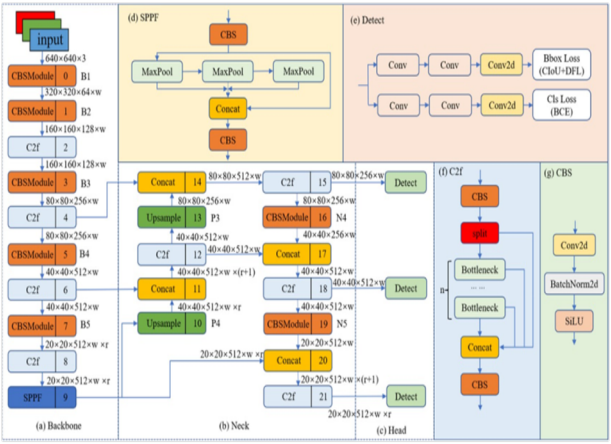

Fig. 2. Architecture of YOLOv8

Yolov5: YOLOv5 is a very popular object detection model that is recognized for its rapidity and precision. It extends upon the capacities of previous YOLO models by enhancing network architecture, leading to an efficient real-time detection. On the other hand, YOLOv5 has a broad range of applications such as detecting objects in both images and video,s including tiny objects that are complicated. Furthermore, it is lightweight and easy to install or use in numerous places, making it fit for autonomous

driving, security monitoring as well as industrial observation, where quick and precise identification of objects is critical.

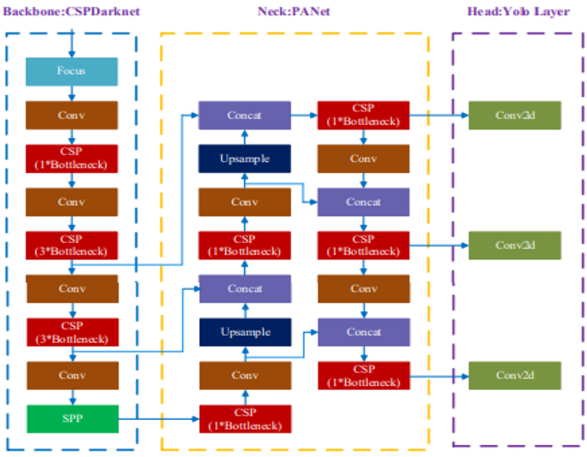

Fig. 3. Architecture of YOLOv5

3.4 Experimental Design and Model Training

We used the YOLOv8 model on the HIT-UAV dataset to investigate the object detection capabilities of this latest model of YOLO. The YOLOv8 network structure is shown in Figure 2, primarily made up of a head, neck, and backbone. Then we used the YOLOv5 model on the same dataset to compare with YOLOv8. The original images were of 640x512 size, which we reduced to 512x512. For the training, we used 100 epochs and a batch size of 32, as shown in Table 1. Our model was trained on the Google Colab platform using their T4 GPU.

Table 1. Parameter Values

Parameter Used	Value Taken
Image size	512
Epoch	100
Batch size	32

3.5 Model Evaluation

This research evaluates the proposed model using commonly used metrics such as precision (P), recall (R), and mean average precision (mAP). The recall of a model suggests its capability to identify every real appearance of an object inside an image.

Recall is calculated by dividing the sum of real objects by the number of real objects detected. A greater value of recall suggests that even though the model may show a large number of false positives, it is less likely that it will skip over real objects.

Precision is a measurement of detection accuracy. It is calculated by dividing the total number of positive detections by the number of actual detections. High precision means that while the model may miss some real objects, it is less likely to produce a lot of false positives.

Mean average precision (mAP) is a statistic used to judge item detection ability by taking accuracy at different recall levels. It evaluates the model's item detection accuracy across a fixed range of recall limits. By calculating the area under the Precision-Recall curve and averaging these areas across all the classes in the dataset. Finally, by the end result, being mAP is determined. Since it offers a better representation of the model's performance across various classes and levels of complexity in object identification, mAP is one of the better metrics. A better comprehensive understanding of the model's object identification performance can be observed by using these three measures: precision, recall [18] and mean average percentage [19].

4 Result Evaluation

In this section, our YOLOv8 model is evaluated for object detection using the HIT-UAV dataset. The YOLOv8 model has various configurations: nano (n), small (s), medium (m), large (l), and extra-large (x).

4.1 Model Performance Evaluation

Table 2. Performances of the models

Model	Precision	Recall	map50	map95
YOLOv8m	0.876	0.764	0.819	0.54
YOLOv5m	0.885	0.724	0.791	0.519
YOLOv8s	0.891	0.716	0.823	0.534
YOLOv5s	0.922	0.734	0.838	0.563

Table 2 represents the performance result of the model in comparison to the older YOLOv5 models. As shown in the table, the YOLOv8m model achieved the mAP0.5 score of 81.9%, a precision of 87.6% and a recall of 76.4% better than the YOLOv5m model by a big margin. On the other hand, the YOLOv8s model falls short of the YOLOv5s model but with little difference, hence showcasing the overall better output of the YOLOv8 model.

In YOLOv8, the training process includes an early stopping mechanism that stops training if there is no improvement in performance within 50 epochs. This ensures

efficient use of resources. To complete 100 epochs, the model took around 1 hour 80 minutes while using the T4 GPU on Google Colab. The training duration highlights YOLOv8's efficiency and suitability for rapid model development, facilitating quicker deployment of high-performing detection models into production or practical applications.

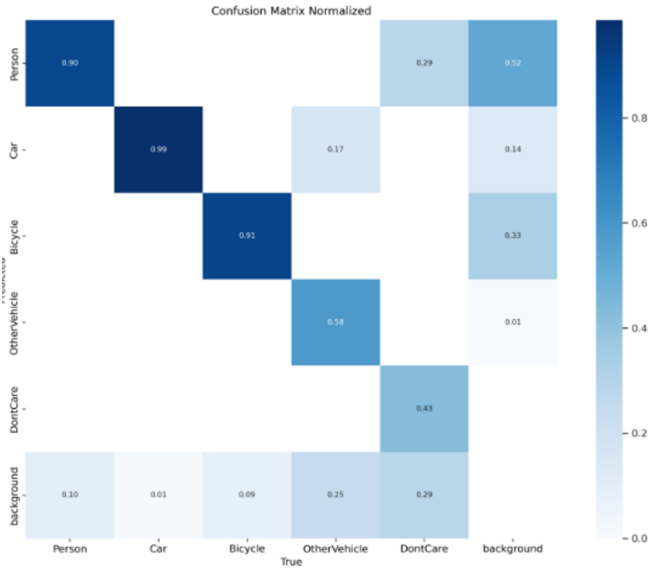

Fig. 4. Confusion Matrix of YOLOv8m

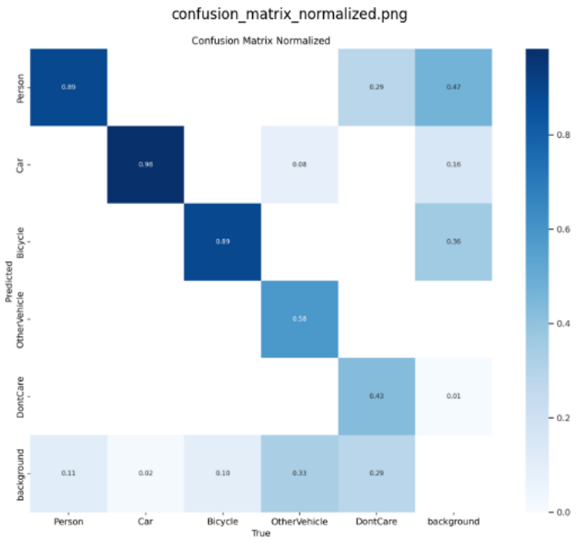

Fig. 5. Confusion Matrix of YOLOv8s

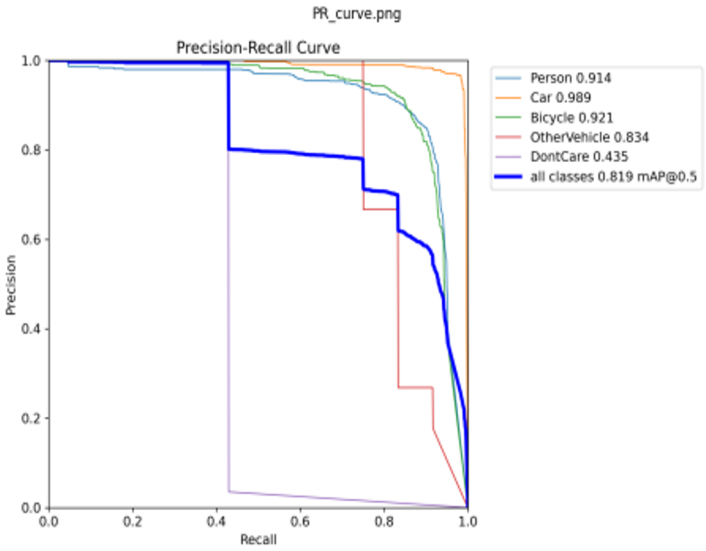

Fig. 6. YOLOv8m model's Precision Recall Curve

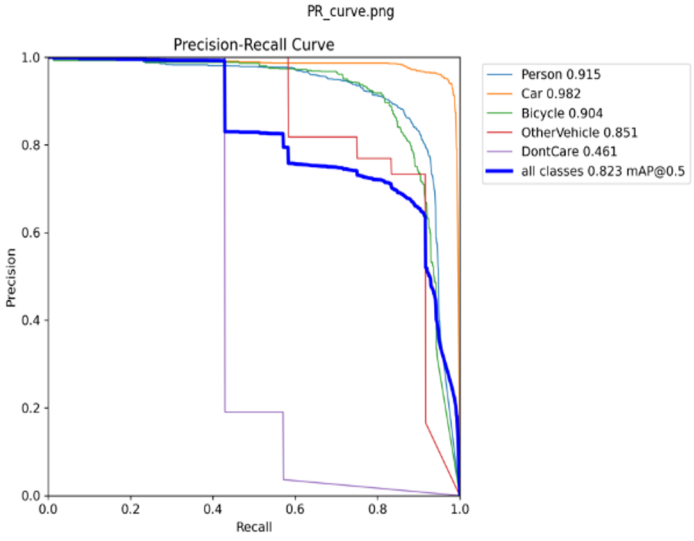

Fig. 7. YOLOv8s model's Precision-Recall Curve

Fig. 8. Prediction result for YOLOv8m

Fig. 9. Prediction result for YOLOv8s

Fig. 4 and 5 shows the confusion matrix for the YOLOv8 model. In the matrix, there are still a few error detections, the highest of which we see in people being detected as background. This issue arises because thermal imagery of humans is very small and very likely to blend in with other white backgrounds. We also observe similar but fewer issues with the background for the other classes. YOLOv8 model, although effective, still struggles with these misclassifications. This confusion can be because of the nature of thermal images, which capture heat signatures instead of visible light. The heat radiation profiles of different objects may not be clear and distinct enough, especially when temperature differences between the objects and the background are minimal.

The precision-recall curve is shown in Fig. 6 and 7 for YOLOv8 models. In Fig. 8 and 9, the model's detection capabilities are shown. While the model demonstrates good performance, there are still challenges in accurately identifying objects. Even small changes in detection performance can have a huge impact on practical applications. For

example, in security and surveillance, the inaccuracy of object detection can critically affect the effectiveness of the system. Misclassifications or missed detections may compromise the reliability of the system. Additionally, in resource-constrained environments, models that offer similar performance but with lower computational requirements might be preferred.

5 Conclusion and Future Work

The object detection using UAVs' thermal images are high in demand for both speed and accuracy. YOLO is a suitable model for real-time processing, but previous research has shown its accuracy to be lacking. This study proposes using the latest version of YOLO. Specifically, the YOLOv8 models to be tested on the HIT-UAV dataset, which contains thermal images of various objects such as Cars, Persons, Bicycles, and Other Vehicles. Experimental results show that YOLOv8m achieved mAP0.5 scores of 0.819. The performance of YOLOv8 is comparable to other frequently used models in the object detection sector, making it a viable option for real-time thermal detection. Though there is much more to be desired of the outcome, our model was tested on Google Colab with the T4 GPU, using a better GPU, like for example the NVidia 4090 and using more epochs than we did might have the possibility of producing better results on both precision and completion time.

References

1. A. Alexander and Meher Madhu Dharmana, “Object detection algorithm for segregating similar coloured objects and database formation,” Apr. 2017, doi: https://doi.org/10.1109/iccpct.2017.8074332.
2. W. Li, D. Zheng, T. Zhao, and M. Yang, “An effective approach to pedestrian detection in thermal imagery,” May 2012, doi: https://doi.org/10.1109/icnc.2012.6234621.
3. V. Teju and D. Bhavana, “An efficient object detection using OFSA for thermal imaging,” The International Journal of Electrical Engineering & Education, p. 002072092094443, Aug. 2020, doi: https://doi.org/10.1177/0020720920944434.
4. Marco San Biagio, M. Crocco, and M. Cristani, “Recursive segmentation based on higher order statistics in thermal imaging pedestrian detection,” May 2012, doi: https://doi.org/10.1109/isccsp.2012.6217877.
5. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, 2016, doi: https://doi.org/10.1109/cvpr.2016.91.
6. D. Shen, X. Chen, M. Nguyen, and W. Q. Yan, “Flame detection using deep learning,” IEEE Xplore, Apr. 01, 2018. https://ieeexplore.ieee.org/document/8384711 (accessed Oct. 14, 2021).
7. A. Gomez, F. Conti, and L. Benini, “Thermal image-based CNN’s for ultra-low power people recognition,” May 2018, doi: https://doi.org/10.1145/3203217.3204465.
8. X. Zhao, Y. Xia, W. Zhang, K. Chen, and Z. Zhang, “YOLO-ViT-Based Method for Unmanned Aerial Vehicle Infrared Vehicle Target Detection,” Remote Sensing, vol. 15, no. 15, pp. 3778–3778, Jul. 2023, doi: https://doi.org/10.3390/rs15153778.

9. X. Jia, Y. Wang, and T. Chen, “Forest Fire Detection and Recognition Using YOLOv8 Algorithms from UAVs Images,” Jul. 2023, doi: https://doi.org/10.1109/icpics58376.2023.10235675.
10. R. Kundu, “YOLO: Real-Time Object Detection Explained,” www.v7labs.com, Jan. 17, 2023. https://www.v7labs.com/blog/yolo-object-detection.
11. M. E. ATİK, Z. DURAN, and R. ÖZGÜNLÜK, “Comparison of YOLO Versions for Object Detection from Aerial Images,” International Journal of Environment and Geoinformatics, vol. 9, no. 2, pp. 87–93, Jun. 2022, doi: https://doi.org/10.30897/ijegeo.1010741.
12. I. S. Gillani et al., “Yolov5, Yolo-x, Yolo-r, Yolov7 Performance Comparison: A Survey,” Computer Science & Information Technology (CS & IT), vol. 12, no. 12, p. 17, Oct. 2022, doi: https://doi.org/10.5121/csit.2022.121602.
13. Uddagiri Sirisha, S Praveen, Parvathaneni Naga Srinivasu, Paolo Barsocchi, and Akash Kumar Bhoi, “Statistical Analysis of Design Aspects of Various YOLO-Based Deep Learning Models for Object Detection,” International Journal of Computational Intelligence Systems, vol. 16, no. 1, Aug. 2023, doi: https://doi.org/10.1007/s44196-023-00302-w.
14. L. Tan, T. Huangfu, L. Wu, and W. Chen, “Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identification,” BMC Medical Informatics and Decision Making, vol. 21, no. 1, Nov. 2021, doi: https://doi.org/10.1186/s12911-021-01691-8.
15. J. Suo, T. Wang, X. Zhang, H. Chen, W. Zhou, and W. Shi, “HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection,” Scientific Data, vol. 10, no. 1, p. 227, Apr. 2023, doi: https://doi.org/10.1038/s41597-023-02066-6.
16. H. Liu, K. Fan, Q. Ouyang, and N. Li, “Real-Time Small Drones Detection Based on Pruned YOLOv4,” Sensors, vol. 21, no. 10, p. 3374, May 2021, doi: https://doi.org/10.3390/s21103374.
17. https://www.kaggle.com/datasets/pandrii000/hituav-a-highaltitude-infrared-thermal-dataset?rvi=1.
18. C. J. van Rijsbergen, "Information retrieval," Butterworth-Heinemann, 1979.
19. M. Everingham et al., "The Pascal Visual Object Classes (VOC) Challenge," International Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, 2010..
20. R. Ippalapally, S. H. Mudumba, M. Adkay, and N. V. H. R., “Object Detection Using Thermal Imaging,” IEEE Xplore, Dec. 01, 2020. https://ieeexplore.ieee.org/document/9342179.
21. W. Fang, L. Wang, and P. Ren, “Tinier-YOLO: A Real-time Object Detection Method for Constrained Environments,” IEEE Access, pp. 1–1, 2019, doi: https://doi.org/10.1109/access.2019.2961959.