Journal of Computer Science

Research Article

An Enhanced Algorithm for Small Object Detection based on Thermal Imaging using YOLOv8-EPB

[1]Ravina Gupta, [1]Sarika Jain and [2,3]Manoj Kumar

[1]AIIT, Amity University, Noida, India
[2]School of Computer Science, University of Wollongong in Dubai, Dubai Knowledge Park, Dubai, United Arab Emirates
[3]MEU Research Unit, Middle East University, Amman, Jordan

Article history
Received: 25-09-2024
Revised: 01-03-2025
Accepted: 24-03-2025

Corresponding Author:
Ravina Gupta
AIIT, Amity University, Noida, India
Email: mahawar.ravina@gmail.com

Abstract: Object detection is one of the most important and challenging problems in the computer vision domain. Using the power of deep models, researchers have carefully explored and made significant contributions to increasing the effectiveness of object detection and related tasks, such as object identification, localization, and segmentation. This progress is due to the rapid progress of deep learning in the past decade. However, object detection in thermal imaging has certain challenges and has potential uses in areas like autonomous driving, security, and surveillance. When applying several popular object detection algorithms to ground-based thermal imaging, the main obstacles include the small size of the targeted object, low-quality images, obstruction, and varying illuminating conditions. In this study, to address this problem enhanced version of YOLOv8 termed as YOLOv8-EPB algorithm has been proposed to target small-size objects in ground-based thermal images. Initially replacing the CSPDarknet53 backbone with EfficientNet-B4 reduces model parameter's computational complexity and increases inference speed. In addition, a new compact target-detecting layer and head have been created to reduce noise in thermal imaging. Lastly, adding a Bidirectional Feature Pyramid Network (BiFPN) to the neck section improves model generalization by lowering detection errors caused by scale deviations and complex situations. The study evaluates a proposed algorithm through ablation experiments and comparisons with other algorithms, focusing on detection performance. The algorithm obtained a mean Average Precision of 92.3% in a self-made thermal imaging dataset, with an accuracy increase of 4.7% compared to regular YOLOv8 models and outperforming other leading-edge detection algorithms.

Keywords: Small Object Detection, Thermal Imaging, YOLOv8-EPB, BiFPN, Accuracy

Introduction

Object detection techniques are quintessential in computer vision, which makes it possible to identify, interpret and analyze objects within images or even video streams. These techniques examine an input image or video frame to determine the presence of distinct objects and then precisely outline their location and other required details. Object recognition has been extensively carried out on visual images as well as aerial, ground and space-borne remote sensing images. These methods may be used in a variety of computer vision applications, including object tracking, finding, video surveillance, captioning of images, image segmentation, clinical imaging and numerous other domains (Budzier & Gerlach, 2019). Due to the strong, low-noise characteristics of visible spectrum images, such as edges, colour and texture, most object-detection systems perform well with this type of imaging. However, existing algorithms are ineffective in detecting objects under challenging weather conditions such as fog, rain, night, noise, limited visibility, low contrast and instances where the foreground colour matches the background. This is because images captured by visible cameras often have weak contrast in low or highlight conditions, which reduces the effectiveness of traditional object detection algorithms.

These problems can effectively be addressed through Thermal Imaging. To obtain information about objects,

thermal imaging uses infrared radiation and thermal energy, which fall between 0.7 and 300 μm wavelengths on the electromagnetic spectrum range. This technology proved to be highly valuable for obtaining image data in environments with limited visibility. Thermal imaging is an appropriate night-vision approach that uses infrared radiation instead of visible light, making it suitable for operating in total darkness. Additionally, it can operate in difficult weather conditions such as haze, smog and smoke. It is a quick, accurate and radiation-free detecting method that creates images by assessing an object's surface temperature without requiring contact or intrusion (Gupta *et al.*, 2023; Pathmanaban *et al.*, 2019; Usamentiaga *et al.*, 2014). Thermal imaging has greater recognition due to its improved performance in various fields, such as public health, security, transportation monitoring and body temperature detection. The availability of high-processing resources has led to an increase in the range of applications of object detection techniques based on thermal imaging. These techniques are now being used in COVID-19 prevention surveillance, search and rescue operations and autonomous driving (Yaqoob *et al.*, 2021).

Deep learning has powerful features in learning capabilities, which has led to its wide application in object detection and image processing. The most used deep learning framework is the Convolutional Neural Network (CNN). It is known for its popularity and effectiveness across different domains (Zhong *et al.*, 2020). For classical object detection, numerous deep learning-based frameworks have been developed over the years. Two-stage detectors, such as Regions with CNN features (RCNN) (Girshick *et al.*, 2014), Spatial Pyramid Pooling Networks (SPPNet) (He *et al.*, 2015), Fast RCNN (Girshick, 2015), Faster RCNN (Ren *et al.*, 2017) and Feature Pyramid Networks (FPN) (Ren *et al.*, 2017) and one-stage detectors, such as You Only Look Once (YOLO) (Redmon *et al.*, 2016), Single Shot MultiBox Detector (SSD) (Liu *et al.*, 2016) and RetinaNet (Lin *et al.*, 2017), are examples of CNN-based object detection techniques. The development of these effective detectors allowed for the tracking and detection of objects in optical images (Li *et al.*, 2020). Among them, one-stage detector YOLO models are particularly advantageous for real-time optical image detection (Liu *et al.*, 2020). The new YOLOv8 algorithm, which was launched in 2023 and achieved extremely high accuracy, is part of the fastest-growing YOLO series of algorithms. YOLO is effective in detecting full-sized targets, but it may not perform as well as some small-sized objects are to be detected across different scenes.

Despite the advancement in current methods, detecting small targets and objects across multiple scales remains a challenge. In thermal imaging, objects within the same category often exhibit significant size variations. To provide a solution to these challenges, this study proposes a novel algorithm for the detection and classification of objects in thermal imaging. This algorithm is based on a modified version of the yolov8 model. The algorithm significantly improves the recognition of small targets within complicated scenes and has consistent enhancement in detecting normal-scale targets. It aims to improve accuracy in detecting small objects while maintaining modest improvement for normal-scale objects. The model has been trained and validated over six classes of small objects with customized thermal imaging datasets. The main contributions of this algorithm are as follows:

1. Efficient net-B4 is utilized as a feature extraction network that balances computational efficiency and the ability to capture detailed features
2. BiFPN is used to fuse multi-scale features, which improves the precision of small target recognition. This improves the integration of features from different levels, resulting in better detection performance
3. A new solution has been created to overcome the issue of complex recognition in thermal images due to noise. This involves the development of a small target detection layer and detection head

Related Work

This part covers the background study of object detection techniques in thermal imaging.

Deep learning-based object recognition techniques have demonstrated impressive results in thermal imaging. However, detecting objects at multiple scales and identifying small objects continue to pose significant challenges. Researchers are actively contributing to addressing these challenges and improving object detection. Ghenescu *et al.* (2019) developed a technique to address the challenging task of detecting distant objects in thermal imaging. The inadequate resolution that comes with thermal imaging is the cause of this problem. They addressed this issue by creating, training and testing a vast number (2640) of modifications to YOLO Darknet, a leading algorithm for object detection in visible camera images, to improve its performance in thermal imaging. Han *et al.* (2020) developed the Ghost module to create efficient designs of neural networks. They used this module to construct GhostNet, which achieved an acceptable balance between effectiveness and precision. Ma *et al.* (2020) found that the HRNet feature extraction network outperformed Faster R-CNN in improving the identification of moderate and small-sized animals in large-scale images. This network improved the ability to recognize small objects. To retrieve information from ground-based thermal images and videos, Jiang *et al.* (2020) designed a Neural Network-based You Only Look Once (YOLO) model architecture. An evaluation metrics approach was used to establish the most effective algorithms and then the proposed algorithm was applied to detect objects on thermal video streams by UAVs. The maximum detection

speed reached 50 frames in one second and the YOLOv5-s was observed to be the most compact model of all tested models. Lai *et al.* (2023) developed a feature acquisition module that integrates convolutional neural networks (CNNs) with multi-head attention, facilitating the STC-YOLO approach and expanding the receptive field. The Normalized Gaussian Wasserstein Distance (NWD) metric was also employed to improve sensitivity to positional variations in small objects. To make it more effective a lightweight YOLO network (GCL_YOLO) with a GhostConv-based backbone has been proposed by Cao *et al.* (2023). Initially, the network creates a narrow backbone network using ghost convolutions and an insignificant number of parameters. The large-object prediction head that is now in use for objects in natural scenes is thereafter to be replaced with an entirely novel small-object prediction head. The network's localization loss is finally the focus-effective intersection over union (Focus-EIOU) loss. To find small targets and locate objects at a variety of scales, Wu and Dong (2023) proposed YOLO-SE, a cutting-edge YOLOv8-based network. The network's parameter count is decreased by employing a lightweight convolution SEConv in place of regular convolutions, which expedites the detection process. The study proposes the SEF module, an improvement based on SEConv to focus on multi-scale object detection. Lou *et al.* (2023) introduced a module that predicts quality-aware factor maps for each modality. This illustrates the reliable nature of every modality and demonstrates locations where the appearance of small objects is most likely. Lyu *et al.* (2024) have developed a small object detection algorithm called DC-YOLOv8. A novel network module is developed to achieve effective performance and accuracy. The accuracy achieved on the PASCAL VOC2007 dataset is 0.5% higher than that of the original YOLOv8. To recognize tiny humans from UAV images, Lin *et al.* (2017) improved YOLOv3 by combining it with two ResNet. It builds multi-scale feature maps and extracts features using a residual network-based Feature Pyramid Network (FPN). In addition, multiple dimensions feature map selection criteria and small-scale anchor boxes are included in the method for increased object detection accuracy. Table (1) illustrates a thorough summary of various methods for identifying objects in thermal imaging.

Table 1: Survey of Object-Detection Methods in Thermal Images

Article	Approach	Utilized Dataset	Accuracy	Conclusion
Ghenescu *et al*. (2019)	YOLO Darknet	Self-created thermal dataset	68.75%	Detect extremely small objects up to 50 pixels
Han *et al*. (2020)	GhostNet	CIFAR-10 dataset	75.7%	reached a state of equilibrium between accuracy and efficiency
Ma *et al*. (2022)	HRNet feature extraction network	UAV images	92.2%	A model was developed to identify medium and small-sized animals in large images and it outperformed Faster CNN.
Jiang *et al*. (2022).	YOLOv5	UAV TIR Images	86.75%	Archived the fastest detection speed using a small model size
Lai *et al*. (2023)	STC-YOLO	TT100K	90%	The Normalized Gaussian Wasserstein Distance (NWD) metric was introduced to improve the ability to detect minor variations in the location of objects.
Cao *et al*. (2023).	YOLO network (GCL_YOLO)	VisDrone-DET2021 & UAVDT dataset	-	Reduced parameters in the proposed network and improved accuracy
Wu. *et al*. (2023)	YOLO-SE	SIMD dataset	0.91%	Conclude that YOLO-SE is a compelling solution for multi-scale object detection.
Lou *et al*. (2023)	Quality-aware RGBT Fusion Detector (QFDet)	VTUAV tracking dataset	57.43%	The proposed network can predict the accuracy of localization and classification for each modality.
Lu *et al*. (2023).	DC-YOLOv8	PASCAL VOC2007	83.5	Achieved 0.5% higher accuracy than the original YOLOv8
Lyu *et al*. (2024)	Improved Yolov3	UAV-viewed Human dataset	80.59%	The proposed method performs well on small object detection

The Network Structure of Yolov8

YOLOv8 employs a backbone similar to YOLOv5 but introduces modifications in the CSPlayer, which has been updated to the C2f module. This new C2f module integrates contextual information with high-level features through a two-stage partial bottleneck that utilizes convolution. YOLOv8 incorporates a decoupled head model for handling object detection, classification and regression tasks. This architecture enhances overall accuracy by allowing each branch to specialize in its specific function. The model uses the sigmoid function to activate object scores, which indicate the likelihood of an object being present within the bounding box. Class probabilities, which denote the likelihood of an object belonging to each potential class, are determined using the SoftMax function.

YOLOv8 employs binary cross-entropy for classification loss and uses the Complete Intersection over Union Loss (CIoU) (Zheng *et al.*, 2020) and Distribution Focal Loss (DFL) (Li *et al.*, 2020) for bounding box loss. These loss functions improve the model's object detection capabilities, especially for small objects. The architecture of YOLOv8 includes three

primary components: the backbone network, the neck network and the prediction output head. The backbone network is essential for extracting features from thermal images. The neck network, located between the backbone and the prediction output head, is responsible for processing and integrating these features. YOLOv8 typically utilizes a Feature Pyramid Network (FPN) in the neck network to effectively merge features from different scales, providing a more detailed representation.

The YOLOv8 model's prediction output head is designed to identify and locate various objects within images. It achieves this by using multiple detectors to determine both the positions and classifications of objects. To address objects of different sizes, the model incorporates three sets of detectors, each tailored to distinct scales. Figure (1) illustrates the network architecture of YOLOv8.

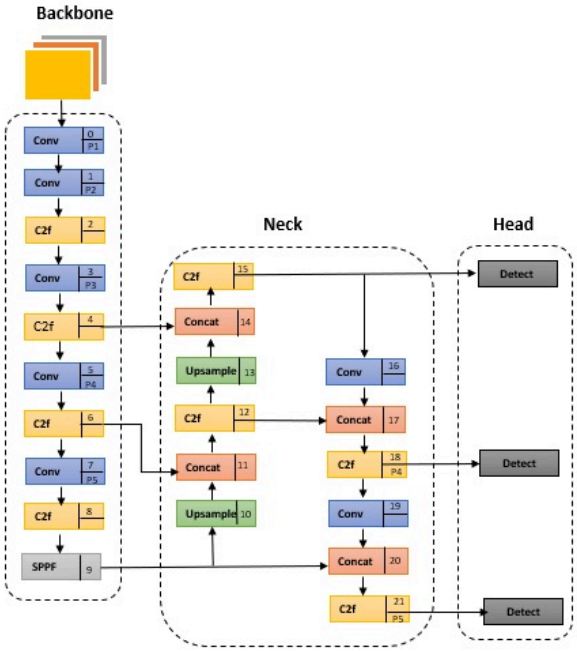


Fig. 1: Network architecture of YOLOv8

Enhanced Version of Yolov8

The YOLOv8 model has three components - backbone, neck and head. The backbone extracts the original image features using convolution layers, the neck enhances and combines them using multi-scale feature aggregation techniques and the head produces the final detection output. All these components work together to enable accurate and efficient object detection. YOLOv8 has been very effective in all aspects but faces several limitations when it comes to detecting small objects in thermal imaging. The main reasons for small target detection errors are a) During the neural network feature extraction process, Small targets can be confused with larger ones and deep-level features lack important small target details. As a result, small targets are neglected throughout the learning process, leading to poor detection performance. b) Compared to normal-sized objects, small objects are more likely to coincide with others and can be easily partially blocked by larger objects. This overlap and occlusion make distinguishing and accurately locating small objects within the image difficult. A new detection technique has been proposed to address this issue in thermal imaging, without affecting the ability to identify normal-sized targets.

Backbone Network

EfficientNet B4 is a network model that efficiently detects objects in thermal imaging. It requires less computing, fewer parameters and shorter inference times compared to traditional networks, making it suitable for embedded devices with limited power and storage. It has balanced and efficient architecture that works well in low-contrast and high-interference environments. It uses the Squeeze-and-Excitation technique to improve its ability to identify small items from the background, making it suitable for thermal imaging, where identifying small things is crucial. The features mentioned are crucial in achieving exceptional performance in object detection, image classification and semantic segmentation tasks.

As shown in Figure (2), the MBConv block takes input dimensions of H x W x 4C, where H and W are the height and breadth of the feature map and 4C represents the Number of channels. This expands the channel count allowing the network to capture more complex features in the expanded space. The feature maps are processed independently through depthwise convolution with a kernel size of k x k. Batch normalization and Swish activation are then applied. The Squeeze-and-Excitation (SE) layer recalibrates the feature maps by reducing each channel to a single value through global average pooling and then scaling each channel's relevance through a small neural network. After the SE layer, there is another 1x1 convolution to get the channels back to 4C. To prevent overfitting, dropouts are applied to feature maps. A residual connection is used to append the original input to the processed feature maps, improving gradient flow and making it easier to train deeper networks. The input dimensions of HxWx4C are retained in the final output of the MBConv block.

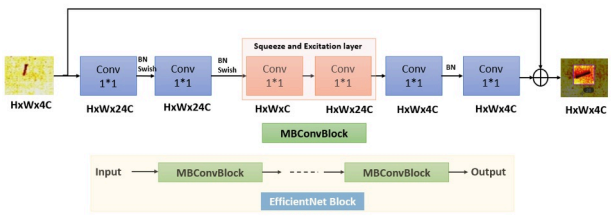


Fig. 2: The block structure of EfficientNet B4

Neck Structure

Bi-direction feature pyramid network: The Feature Pyramid Network (FPN) shown in Figure 3(a) integrates multi-scale attributes from levels 1 to 3(P1 to P3) to enhance target identification at various stages. Nevertheless, it has computational challenges and needs a lot of time for training and inference. Its unidirectional information flow restricts its adaptability. To overcome this, the feature fusion technique is being improved. Rather than relying solely on the FPN, the Path Aggregation Network (PAN) adds another top-down path aggregation network (Liu *et al.*, 2018). This enhancement helps retain detailed information in low-resolution feature maps, thereby improving detection accuracy. However, this enhancement also increases the computational burden, as Figure 3(b) illustrates. YOLOv8 then takes an instruction from PAN in Figure 3(c), simplifying the network to increase the speed of detection. YOLOv8 removes nodes with insufficient feature fusion, optimizing the feature pyramid network. Nevertheless, all feature fusion techniques have limitations in localizing and identifying small targets. This is because the network tends to overlook subtle information during feature extraction, making small targets more susceptible to interference from larger ones. Consequently, the information available about small targets decreases, leading to suboptimal target detection. The BiFPN approach evaluates the significance of various input properties using learnable weights during feature fusion and extraction. This approach addresses redundancy and information deterioration by allowing feature information to flow both ways. The method enhances feature fusion and exploitation at various sizes by integrating horizontal and vertical characteristics during the iterative top-down and bottom-up multi-scale feature fusion process. As seen in Figure 3(d), the progressive integration of horizontal and vertical links together with top- and bottom-sampled feature maps enhances feature fusion and exploitation at different scales. BiFPN is effective in handling complicated scenarios with scale fluctuations and occlusions.

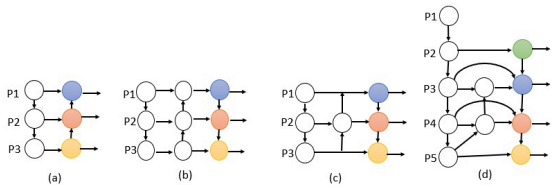


Fig. 3: Feature Fusion networks design. (a) FPN; (b) PAN; (c) YOLOv8; (d) BiFPN

Detection Module (Head)

This study introduces a detection head and small target detection layer to address the challenges in recognizing complicated targets due to large variations in the thermal imaging scale. The YOLOv8 network layout includes down samples at 8x, 16x and 32x, producing output maps of 80x80, 40x40 and 20x20, respectively. Smaller feature maps detect large targets, while larger feature maps detect small targets more accurately and effectively due to their bigger receptive fields and more semantic information (Zhang *et al.*, 2023). To enhance the network's capacity to identify small targets, the proposal proposes to add 4x down-sampling with 160 × 160 output maps to the current structure. Moreover, a larger-scale feature map is integrated into the FPN + PAN structure neck, as illustrated in Figure (4), to optimize the network structure.

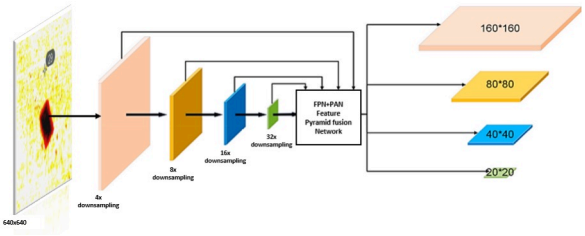


Fig. 4: Implement a dedicated module for detecting

After implementing the improvements mentioned above, the learning ability of the new network has been significantly enhanced. The optimized version of the YOLOv8 network structure is illustrated in Figure (5).

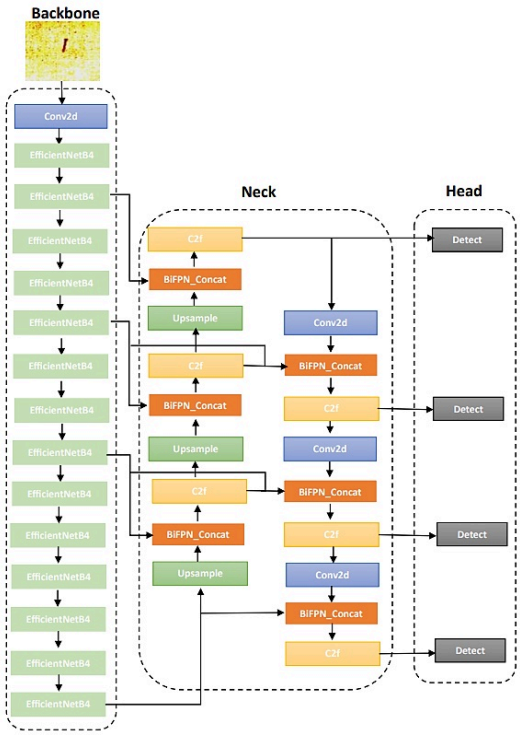


Fig. 5: Network structure diagram of YOLOv8-EPB

Materials and Experiments

Thermal Imaging Dataset

The real-time thermal imaging dataset of small objects has been collected using the “Shot Thermal

Imaging" camera. The camera, along with "seek Fusion" technology, produces a 36◦ angle of view. The thermal imaging camera has a fixed focus lens with a refresh rate of about 9Hz frame. The parameters of the acquired thermal imaging dataset are as follows: Maximum resolution of the captured thermal image is 648 x 480, wherein the approximate camera distance from the captured object is approximately 120 to 250 cm, which can be maximized or minimized depending on the size of the target object. The total number of objects in the dataset is 10706, belonging to four different classes: key, coin, cap, piece, bolt and matchbox. There are roughly two objects on average per image and the median class of object size determined by pixel count is almost 241. The percentage of the image where objects overlap is, on average, 0.26%. Table 2 provides the specifics of the parameters.

Table 2: Parameters Details of Custom Thermal Dataset

Parameter	Custom thermal dataset
Resolution for color thermal images	648 x 480
Number of images	5347
Camera distance	120-250 cm
Total number of objects	10706
Object count per image	2
Median object area (in sq. pixels)	346
Median overlap area of object and image regions	0.26%

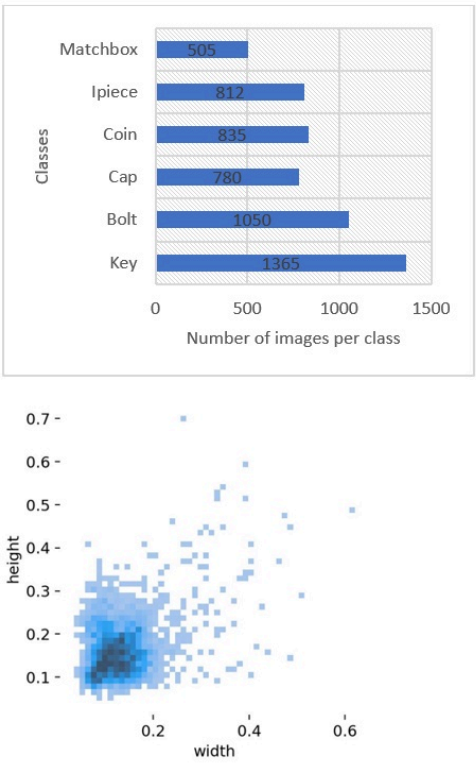


Fig. 6: Target Distribution in the Thermal Dataset: (a) Class Count Distribution; (b) Target Size Distribution

The thermal imaging dataset contains 5347 thermal images categorized into six different classes. The images are segregated as per their classes, as demonstrated in Figure (6). As Figure (6a) illustrates, the "key" class has more than 1300 thermal images, while some other classes, like "matchbox", have just 500 images. The thermal dataset was randomly divided into training, validation and testing subsets with a split ratio of 0.7:0.2:0.1. This equates to 70% of the data for training, 20% for validation and 10% for testing purposes (Usamentiaga *et al.*, 2014). Figure 6(b) shows the range of target widths and heights, with a color gradient from light to dark blue indicating areas of higher frequency. This helps to emphasize the variety and frequency of target sizes in the dataset.

The training dataset comprises 3,743 images, while the validation dataset contains 1,069 thermal images taken under various weather and lighting conditions, both during the day and at night. In addition to the 5,347 images, an extra 535 thermal images have been utilized to determine the robustness of the suggested models. The 'keep aspect ratio resizer' method is used to resize images to 224x224 pixel dimensions. To maintain the original aspect ratio, this mechanism scales the image so that the longer side is 224 pixels and adjusts the shorter side to 224 pixels if the longer side is greater than 224 pixels. Sample thermal images of small objects are shown in Figure (7).

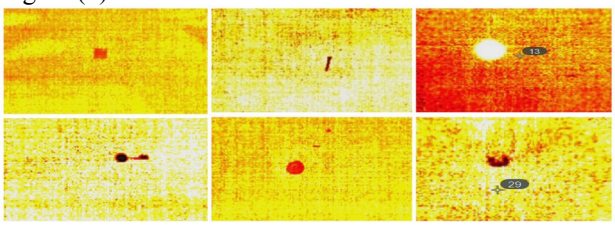


Fig. 7: Sample thermal images of small objects from the dataset

Data Annotation

Data annotation is a process of labelling or marking data to provide accurate information for machine learning model training. Annotation is a type of data annotation where objects and regions in images are marked by creating bounding boxes around objects, polygon segmentation, or labelling key points to help algorithms learn from the labelled data. The Number of labels applied to an image depends on the project requirements. An image may have a single label or multiple labels for specific items, areas, or landmarks within the image. There are various annotation tools available for labeling the data for image annotation tools such as LabeImg, VGG Image Annotator (VIA), LabelBox and COCO annotator are commonly used. In this proposed study, the RoboFlow annotation tool has been used to label small objects in the thermal imaging dataset. For labelling the objects, the thermal imaging

DOI: 10.3844/jcssp.2025.1391.1403

dataset was first uploaded to the Roboflow annotator tool and all the uploaded images were sent to the annotator for annotation. Roboflow has the ability to assign image annotation tasks to all the members involved in the respective project by just adding their credentials like mail ID or allowing the entire annotation to be sent to a single person. After the assignment of the annotation task annotator chooses the annotation method and performs manual or one-by-one annotation of each image. One of the annotation methods is a bounding box.

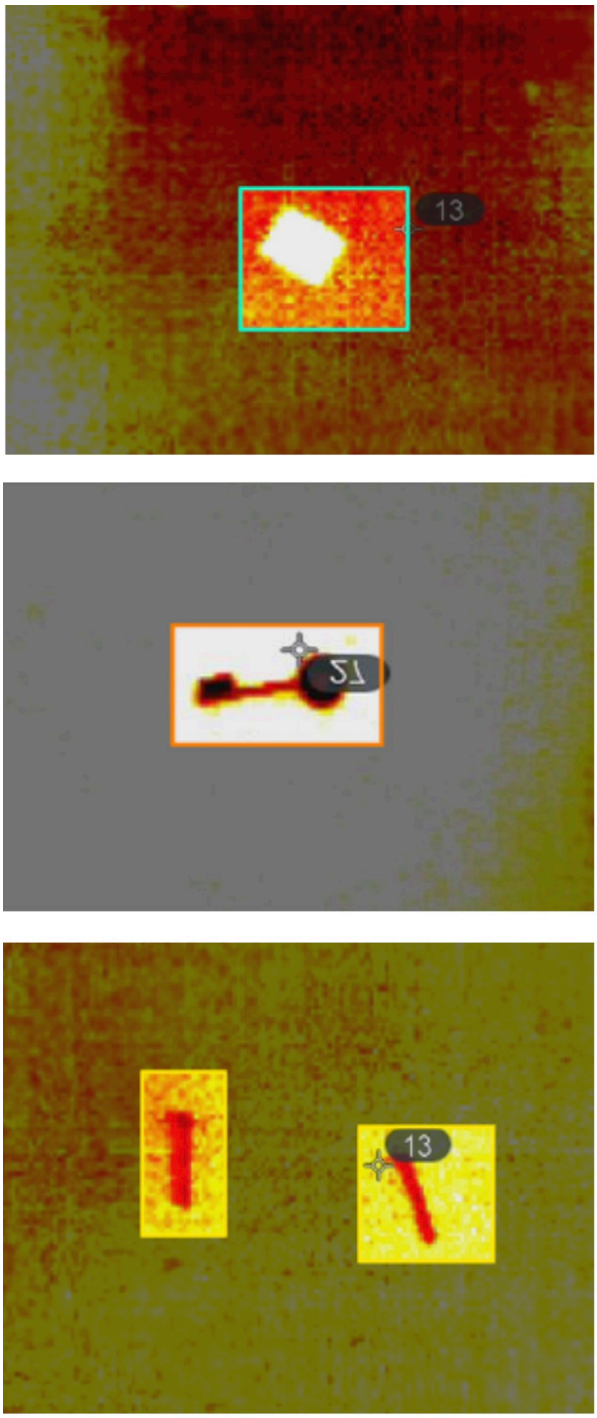


Fig. 8: Bounding box sample images

Bounding Box

The Roboflow annotator tool supports different annotation types such as bounding-box annotation, polygon annotation and keypoint annotation, etc. Bounding box annotation is a method used in computer vision to identify and represent regions of interest in an image. A bounding box has four sides (top, bottom, left and right) that enclose the target object in an image. This helps the model to determine the position and size of the object. It involves drawing rectangles around specific objects or characteristics to precisely locate them. This technique is frequently used in image segmentation, object localization and identification. To format bounding box coordinates for the YOLOv8 dataset, convert absolute pixel values to relative values based on the image's width and height. YOLOv8 requires bounding boxes in the [class x_center y_center width height] format, where:

1. Class refers to the integer representing the object class. x_center and y_center are the coordinates of the bounding box centre, normalized relative to the image's width and height
2. Width and height represent the dimensions of the bounding box and they are normalized relative to the image's width and height. Figure (8) shows the sample thermal images along with bounding boxes

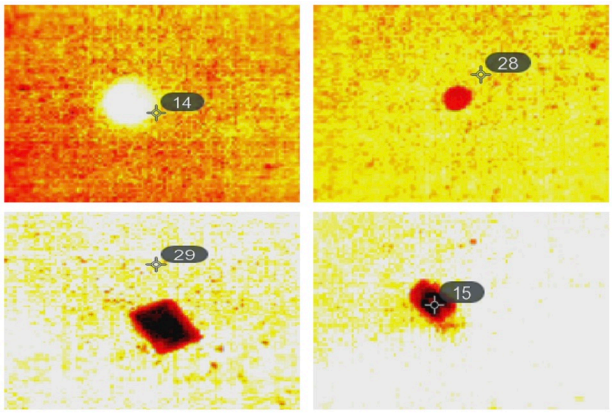


Fig. 9: The different objects were labeled. Top left: coin, top right: cap, bottom left: matchbox, bottom right: piece

Labelling

Proper labeling of each image is essential for using the extracted foreground object in supervised learning. Based on the visual field of an object, one of six labels was assigned to it: key, cap, coin, piece, bolt and matchbox. Manual labeling is applied to each image individually, checked for accuracy and adjusted if found to be incorrect. The labels are referred to as classes and for ease of identification, each class is represented by a unique colored bounding box. As a result, a manual categorization was performed by examining the shape of each object visible in the sequence of images, considering that the coin was identified as precisely

round, while the shape of the cap was described as more elliptical and inconsistent. Similarly, matchboxes are rectangular and pieces are similar in shape but different in shape. See Figure (9). For example, the images include a coin, a cap, a matchbox and an ipiece.

Experimental Platform

The paper used a Windows 11 computer with high-end hardware, including an Intel i5-1035G1 CPU, NVIDIA MX130 GPU and 16GB RAM, to conduct experiments. The tests were carried out using torch 1.12.1 software controlled by the Anaconda framework and supported by CUDA version 11.3. The system had sufficient computing capacity and resources for accurate and productive experimental procedures.

Evaluation Matrix

Mean average precision (mAP) is a standard evaluation metric used in computer vision research to assess the effectiveness of object detection methods. Object detection encompasses both localization and classification tasks: localization involves determining the bounding box coordinates, while classification involves identifying the object labels. mAP is a commonly used evaluation metric for popular object detectors such as Faster R-CNN, SSD and YOLO. To determine the mean average precision (mAP) for object detection, it is essential to calculate recall, precision and intersection over union (IoU). Recall measures the proportion of actual positives that were detected, while precision indicates the Number of true positive predictions among all positive predictions. IoU assesses the overlap between two bounding boxes and is derived from the Jaccard Index. Equations (1-3) demonstrate the equations of precision, recall and IoU, respectively:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2}$$

$$\text{IoU} = \frac{\text{area}(\text{Bp} \cap \text{Bgt})}{\text{area}(\text{Bp} \cup \text{Bgt})} \tag{3}$$

To assess object detection, precision and recall are calculated using the Intersection Over Union (IoU) between the predicted Bounding box (Bp) and the ground-truth Bounding box (Bgt). True Positives (TP), False Negatives (FN) and False Positives (FP) are utilized in these metrics. The average precision is obtained by measuring the area under the precision-recall curve, as outlined in Eq. (4).

$$\text{AP} = \int_0^1 p(r)\, dr \tag{4}$$

Average Precision (AP) is a metric used to measure the accuracy of a model, while mean Average Precision (mAP) is the average of all AP values. The calculation of mAP involves determining the average precision for each class and across all IoU thresholds and the formula for mAP calculation is shown in Eq. (5).

$$\text{mAP} = \sum_{i=1}^{c} \frac{APi}{c} \times 100\% \tag{5}$$

The main precision evaluation metric in this work is mAP@0.5. For convenience, mAP@0.5 will be referred to as AP0.5 in the remaining parts of this study.

Methods

The YOLOv8-EPB model was trained on the custom thermal dataset. The Training utilized the Adam optimizer with a learning rate of 0.001 and weight decay of 0.0005. The model was trained for 100 epochs with a batch size of 32 and early stopping was employed based on validation performance to avoid overfitting. The loss functions used included Complete Intersection over Union (CIoU) loss for bounding box regression, binary cross-entropy for objectless prediction and cross-entropy loss for classification. Hyperparameters, including learning rate and batch size, were tuned through a combination of grid search and manual adjustment to ensure optimal convergence and training stability.

Results Analysis

Experimental Results

The thermal imaging dataset was evaluated for small object detection and achieved an impressive Average Precision of 91.5% at a detection threshold of mAP0.5. The YOLOv8-EPB algorithm was used to obtain detection outcomes for different categories and the results are presented in Table (3).

Table 3: Detection results on thermal imaging dataset

Categories	Precision %	Recall %	mAP0.5
Key	85.7	91.3	97.1
Coin	92.0	93.9	87.8
Cap	91.3	99.2	90.2
Matchbox	87.1	97.2	91.6
Bolt	93.3	96.6	95.4
iPiece	82.0	70.4	94.1
All	91.5	93.4	92.7

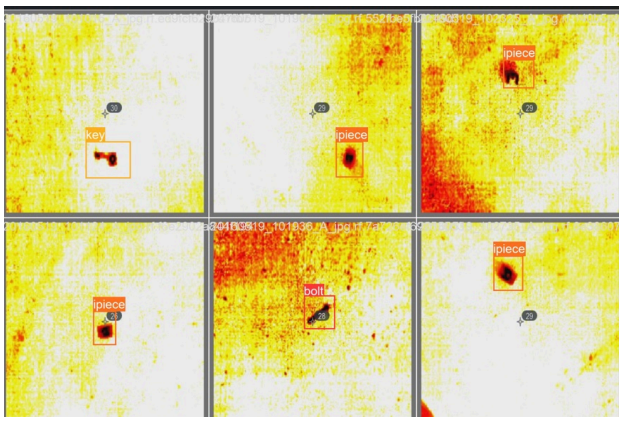


Fig. 10: Detection Results on the Thermal Imaging Dataset

Figure 10 shows how the proposed methodology was applied to a thermal imaging dataset to detect small

objects. The figures demonstrate that the algorithm is proficient in handling the challenges that come with detecting objects of different sizes and dealing with environmental noise. The algorithm performed well in accurately detecting small objects, which shows that it is robust and can work effectively in various real-world scenarios.

The graph in Figure (11) shows how well the YOLOv8-EPB algorithm performs on various classes in the thermal imaging dataset. Some categories like 'key,' 'bolt,' and 'coin' have high precision with average precisions above 90.5%, while the 'other' category has low precision with an average of 84.5%. The 'other' category includes diverse objects that lack consistent features for the network to learn from.

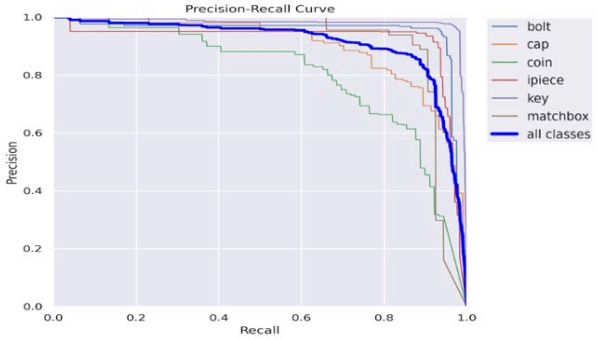


Fig. 11: Precision-Recall curve of YOLOv8-EPB

The dynamic changes in several metrics throughout the validation and training stages have been demonstrated in Figure (12). These measures include post-epoch evaluations like accuracy and recall, as well as box loss, object loss and class loss. The figure provides a comprehensive overview of how these metrics have changed during training and validation, revealing information about the model's performance at various phases.

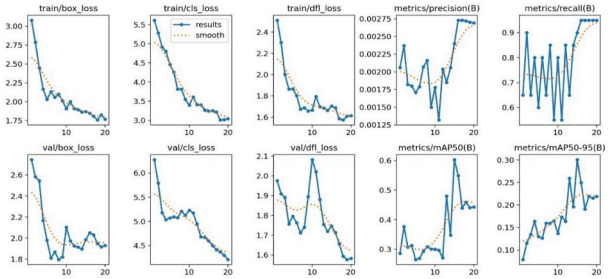


Fig. 12: Training and Validation losses matrices for YOLOv8-EPB

Experiment Comparison

A series of ablation experiments have been conducted using thermal imaging datasets to verify the effective detection of the advanced method on small-sized targets at each stage. The performance of the improved network was compared with the original YOLOv8m to evaluate the effectiveness of their modifications. In the ablation experiments, all parameters remained consistent except for those associated with the added modules. This encompasses pertinent hyperparameters, the training methodology and the experimental setting. YOLOv8-E in this study refers to the YOLOv8 module with the CSPDarknet-53 backbone network (Wang *et al.*, 2020), which EfficientNet-B4 now replaces. The P2 detecting head is added to the YOLOv8 module, which is now known as YOLOv8-P. Furthermore, the YOLOv8 module that integrates the BiFPN feature fusion network is called YOLOv8-B. The experiment's authenticity is demonstrated using mAP0.5 and mAP0.5:0.95 as evaluation indexes, with the results displayed in Table (4).

Table 4: Algorithm comparison at every stage

Methodologies	Module			Results			
	YOLO v8-E	YOLO v8-P	YOLO v8-B	mAP0.5	mAP0.5:0.95	P	R
YOLOv8m				86.6	63	78.3	79
YOLOv8-EPB	✓			86.2	64.2	79.2	81.6
YOLOv8-EPB	✓	✓		89.4	63.4	82.3	84.8
YOLOv8-EPB	✓	✓	✓	92.7	67.5	83	86.5

The YOLOv8m reference baseline achieved a mAP0.5 of 86.6% on the thermal imaging dataset for small objects, according to an analysis of the ablation experiment findings displayed in Table (4) (a) Replacing the YOLOv8 backbone with EfficientNet-B4 increased the recall rate by 2%. The design of EfficientNet-B4 requires fewer parameters and computation, making it more lightweight and practical for real-time performance (b). The modification of the detector head led to a 3.2% improvement in mAP0.5 and a 3% improvement in the recall rate. Adjusting the P2 anchor frame reduces detection errors associated with oversized anchors when identifying small objects. Combining multi-level information, especially shallow shape and size features, enhances the detection and localization of small targets. However, these improvements also result in increased computational complexity for the model. (c) Enhancing the feature fusion method effectively prevented small targets from being missed during the learning process because of inadequate information on their location. The use of this technique led to a 3.3% rise in mAP0.5, indicating that bidirectional feature information flow improves multi-level interaction as well as the fusion and enhanced usage of features across multiple scales. Experimental results indicated that improvements at each stage of the algorithm could boost the model's learning ability. Specifically, the EfficientNet-B4 network decreases the overall model size and parameter count, making deployment on embedded devices simpler. The use of BiFPN and P2 detecting head helped achieve a high mAP0.5 of 92.7%.

Figure (13) compares the experimental results of benchmark models across different categories with various improvement modules. The EfficientNet-B4 lightweight network module showed an increase in average accuracy for most categories except "Coin" and "Matchbox." The "Key" accuracy increased by 4.3%, "Cap" by 3.1% and "Ipiece" by 0.2%, while "Coin" and "Matchbox" accuracy decreased by 4.2% and 3.7%, respectively. The "Bolt" accuracy decreased slightly by 0.9%. By adding the P2 detector head module, the accuracy of certain objects improved: "Coin" by 2.3%, "Matchbox" by 4.1%, "Bolt" by 3.4% and "Ipiece" by 5.5%. "Cap" accuracy increased by 3.2% and "Key" increased by 1.5%. The BiFPN module further improved accuracies: "Key" enhanced by 4.3%, "Coin" by 4.5%, "Cap" by 4.8%, "Matchbox" by 1.4% and "Bolt" by 2.7%, with "Ipiece" remaining unchanged.

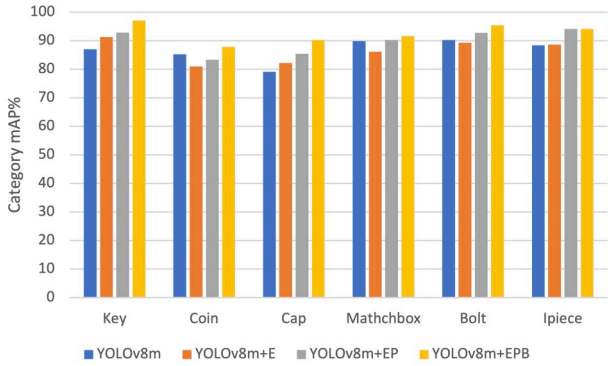


Fig. 13: Comparison of each proposed module with all the categories of thermal imaging small object dataset

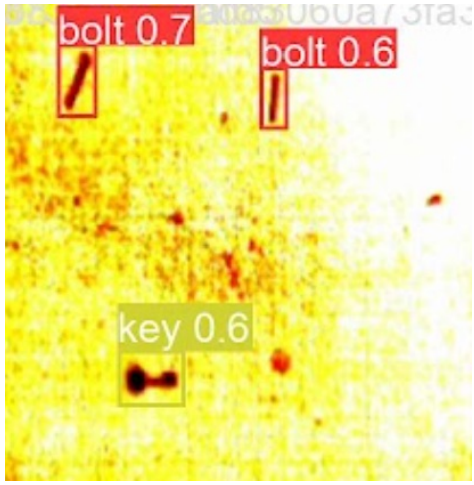
Table 5: Comparative Performance Analysis of the Proposed Algorithm and Other Target Detection Methods

Detection Algorithm	Backbone Network	mAP0.5	mP0.5:0.95
NanoDet	EfficientNet-Lite0	81.2	59.4
Faster R-CNN	ResNet	81.9	62.3
Cascade R-CNN	HRNet	82	58.4
YOLOv3	Darknet53	83.3	51.1
YOLOv5	CSPDarknet53	83.7	56.4
YOLOv7	DenseNet	86.8	57.2
YOLOv8	CSPDarknet53	87.6	63
YOLOv8-EPB (Proposed Algorithm)	EfficienNet-B4	92.3	67.5

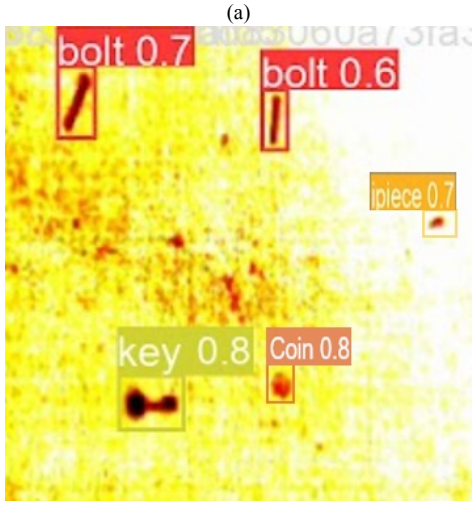
To demonstrate the enhanced model effectiveness, pertinent comparisons have been carried out with the same small object thermal imaging dataset. The comparison experiment includes both one-stage and two-stage target detection algorithms, including NanoDet, YOLOv3, YOLOv5, Faster-RCNN, Cascade RCNN, YOLOv7 and YOLOv8. Table (5) records the comparison between each algorithm that has been tested on the validation set of thermal imaging datasets. We can infer from the comparison analysis that the enhanced YOLOv8-EPB algorithm performs better than other traditional target detection techniques.

The primary reasons why the proposed algorithm excels over other mainstream algorithms are analyzed as follows:

1. The majority of conventional target detection algorithms use FAN+PAN for feature fusion, but this method might lead to information loss during the extraction of features by confusing small and normal-sized objects. In contrast, the feature fusion method in YOLOv8-EPB facilitates multi-level information interaction and improves the fusion and integration of features at different scales
2. When extracting features, small-sized target pixels will be ignored, causing reduced accuracy and irrelevant information will be automatically eliminated. On the other hand, YOLOv8-EPB uses the concept of EfficientNet-B4, which uses a compound scaling method to learn more detailed information and improve accuracy



(a)



(b)

Fig. 14: Test result comparison of YOLOv8m and YOLOv8-EPB. 14(a) Test results of YOLOv8m 14(b) Text results of YOLOv8-EPB

An image from the thermal imaging data was chosen for testing to determine the impact of YOLOv8-EPB. To compare test results, the weight files for YOLOv8 and YOLOv8-EPB have been saved. Figure (14) demonstrates the comparative performance of the two models in small object detection. In the experiment comparison, Figure 14(a-b) showed that YOLOv8 failed to detect coins and pieces. However, the algorithm proposed in this study solved the problem related to missed detection and accurately recognized all objects in the thermal images. As illustrated in Figure 14(a-b), YOLOv8-EPB outperformed YOLOv8 in terms of accuracy and Number of detected targets, indicating that YOLOv8-EPB has better abilities in detecting small objects in thermal imagery.

In summary, the YOLOv8-EPB algorithm improves the ability of the original YOLOv8 model to detect small objects in ground-based thermal imaging by optimizing feature extraction and precise target localization. YOLOv8-EPB addresses the limitations of YOLOv8, effectively reducing missed detections and increasing the accuracy of identified objects.

Discussion and Future Scope

The YOLOv8-EPB algorithm incorporates several innovative architectural advancements that distinguish it from existing object detection methods, particularly for thermal imaging. Using EfficientNet-B4 instead of CSPDarknet-53 as the backbone enables compound scaling, allowing for the extraction of finer-grained features and enhancing the detection of small objects. Additionally, integrating the BiFPN feature fusion mechanism overcomes the limitations of conventional methods, such as FAN+PAN, which may lead to information loss and confusion between features of small and normal-sized objects. Beyond these architectural improvements, YOLOv8-EPB effectively manages key challenges inherent to thermal imaging, such as thermal noise and variations in object size. EfficientNet-B4 enhances feature extraction through compound scaling, which helps mitigate thermal noise by preserving high-resolution details, ensuring clearer object representation, while BiFPN strengthens feature integration across different scales, ensuring robust detection of small objects even in cluttered or low-lighting thermal environments. These enhancements make YOLO-EPB not only more accurate but also more adaptable to real-world applications like industrial surveillance, foreign object Debris detection at airport runways and aerial monitoring, where thermal imaging plays a vital role.

Despite its strengths, the proposed YOLOv8-EPB algorithm has certain limitations that require attention. Firstly, the use of EfficientNet-B4 as the backbone, while effective in improving feature extraction, increases computational costs, making it unsuitable for deployment in real-time applications on resource-constrained devices. Additionally, the BiFPN feature fusion mechanism may struggle with very fine-grained feature differentiation in highly cluttered scenes with high noise levels. These limitations highlight the need for further research into lightweight yet robust backbone architectures and advanced noise-reduction mechanisms.

Furthermore, the use of a custom thermal imaging dataset tailored to the study's objectives may limit the algorithm's generalizability across diverse thermal imaging scenarios. The lack of diverse and large-scale annotated datasets for thermal imaging is a broader challenge in the field, which could affect the model's adaptability to different environmental conditions or object types. To overcome these limitations, future research should explore the integration of neural architecture search (NAS) to identify more computationally efficient configurations and investigate advanced denoising techniques for thermal images. Expanding the dataset to include a wider variety of thermal imaging contexts, such as aerial, underwater, or industrial environments, would also enhance the model's robustness and applicability. Additionally, hybrid approaches, such as combining attention mechanisms with BiFPN, could improve the model's ability to capture subtle variations in small objects. Addressing these technical and dataset-related challenges would extend the applicability of the YOLOv8-EPB algorithm and establish it as a robust and efficient solution for real-world object detection in thermal imaging.

Conclusion

It employs EfficientNet-B4 as the backbone network and integrates a specialized small target detection layer along with a P2 detection head to improve the ability of the network to identify small objects. Additionally, a bidirectional feature pyramid network (BiFPN) is incorporated into the neck section to enhance the generalization capabilities of the model and detection accuracy for small targets. The experiments conducted on a customized small object dataset of ground-based thermal imaging show that an enhanced algorithm has a 4.7% increase in average accuracy for object detection compared to YOLOv8m. The enhanced algorithm is suitable for practical applications that require both real-time processing and accuracy. The performance of the proposed algorithm was evaluated through ablation studies and comparisons with other existing algorithms. The investigation and testing of each optimization component proved its viability and efficacy. The YOLOv8-EPB algorithm outperformed other detectors in capturing small targets in challenging conditions, making it more accurate.

Future research will prioritize the enhancement of the computational efficiency of thermal image configurations through the implementation of Neural Architecture Search (NAS) and advanced denoising techniques. Furthermore, the integration of attention mechanisms with BiFPN could potentially improve the

model's ability to detect small objects. To ensure the model's applicability across various scenarios, the dataset will be expanded to include diverse thermal imaging contexts such as aerial, underwater and industrial environments.

Acknowledgment

We sincerely thank the publisher for supporting the publication of this research. We are grateful for the platform and resources that helped us share our work with more people. We also appreciate the efforts of the editorial team for reviewing and improving our paper. Their support means a lot and we are happy to contribute to the research community through this publication.

Funding Information

No grant from a public or private funding agency was obtained for this research.

Author's Contributions

All authors contributed equally to the research design, experimentation, data analysis and manuscript preparation.

Ethics

This piece of writing is unique and includes unreleased content. All co-authors have read and approved the article and the corresponding author attests that there are no ethical concerns.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Budzier, H., & Gerlach, G. (2019). Passive Thermography, Thermal Imaging. *Handbook of Advanced Nondestructive Evaluation"*, 1371-1400. https://doi.org/10.1007/978-3-319-26553-7_12

Cao, J., Bao, W., Shang, H., Yuan, M., & Cheng, Q. (2023). GCL-YOLO: A GhostConv-Based Lightweight YOLO Network for UAV Small Object Detection. *Remote Sens*, *15*(20), 4932. https://doi.org/10.3390/rs15204932

Ghenescu, V., Barnoviciu, E., Carata, S., Ghenescu, M., Mihaescu, R., & Chindea, M. (2019). Object Recognition on Long Range Thermal Image Using State of the Art DNN. *2018 Conference Grid, Cloud & High Performance Computing in Science (ROLCG)*, 1-4. https://doi.org/10.1109/ROLCG.2018.8572026

Girshick, R. (2015). Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile. https://doi.org/10.1109/iccv.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA. https://doi.org/10.1109/cvpr.2014.81

Gupta, R., Jain, S., & Kumar, M. (2023). Role of Thermal Images in Various Applications of Computer Vision. *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, 1-6. https://doi.org/10.1109/ICIEM59379.2023.10166103

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). Ghostnet: More features from cheap operations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1580-1589. https://doi.org/10.1109/CVPR42600.2020.00165

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904-1916. https://doi.org/10.1109/tpami.2015.2389824

Jiang, C., Ren, H., Ye, X., Zhu, J., Zeng, H., Nan, Y., Sun, M., Ren, X., & Huo, H. (2022). Object detection from UAV thermal infrared images and videos using YOLO models. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102912. https://doi.org/10.1016/j.jag.2022.102912

Lai, H., Chen, L., Liu, W., Yan, Z., & Ye, S. (2023). Small object detection network for traffic signs in complex environments. *Sensors*, *23*(11), 5307. https://doi.org/10.3390/s23115307

Li, K., Wan, G., Cheng, G., Meng, L., & Han, J. (2020). Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, *159*, 296-307. https://doi.org/10.1016/j.isprsjprs.2019.11.023

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., & Yang, J. (2020). Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 21002-21012.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936-944. https://doi.org/10.1109/cvpr.2017.106

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017 IEEE International Conference on Computer Vision (ICCV), Venice. https://doi.org/10.1109/iccv.2017.324

Liu, M., Wang, X., Zhou, A., Fu, X., Ma, Y., & Piao, C. (2020). UAV-YOLO: Small object detection on unmanned aerial vehicle perspective. *Sensors*, *20*(8), 2238. https://doi.org/10.3390/s20082238

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path Aggregation Network for Instance Segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8759-8768. https://doi.org/10.1109/CVPR.2018.00913

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. *Computer Vision - ECCV 2016*, *9905*, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

Lou, H., Duan, X., Guo, J., Liu, H., Gu, J., Bi, L., & Chen, H. (2023). *DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor*. Electronics. https://doi.org/10.3390/electronics12102323

Lyu, H., Qiu, F., An, L., Stow, D., Lewison, R., & Bohnett, E. (2024). Deer survey from drone thermal imagery using enhanced faster R-CNN based on ResNets and FPN. *Ecological Informatics*, *79*, 102383. https://doi.org/10.1016/j.ecoinf.2023.102383

Ma, J., Hu, Z., Shao, Q., Wang, Y., Zhou, Y., Liu, J., & Liu, S. (2022). Detection of large herbivores in uav images: A new method for small target recognition in large-scale images. *Diversity*, *14*(8), 624. https://doi.org/10.3390/d14080624

Pathmanaban, P., Gnanavel, B. K., & Anandan, S. S. (2019). Recent application of imaging techniques for fruit quality assessment. *Trends in Food Science & Technology*, *94*, 32-42. https://doi.org/10.1016/j.tifs.2019.10.004

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779-788. https://doi.org/10.1109/cvpr.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137-1149. https://doi.org/10.1109/tpami.2016.2577031

Usamentiaga, R., Venegas, P., Guerediaga, J., Vega, L., Molleda, J., & Bulnes, F. G. (2014). Infrared thermography for temperature measurement and non-destructive testing. *Sensors*, *14*(7), 12305-12348. https://doi.org/10.3390/s140712305

Wang, C.-Y., Mark Liao, H.-Y., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. (2020). CSPNet: A New Backbone that can Enhance Learning Capability of CNN. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1571-1580. https://doi.org/10.1109/cvprw50498.2020.00203

Wu, T., & Dong, Y. (2023). YOLO-SE: Improved YOLOv8 for Remote Sensing Object Detection and Recognition. *Applied Sciences*, *13*(24), 12977. https://doi.org/10.3390/app132412977

Yaqoob, M., Sharma, S., & Aggarwal, P. (2021). Imaging techniques in agro-industry and their applications, a review. *Journal of Food Measurement and Characterization*, *15*(3), 2329-2343. https://doi.org/10.1007/s11694-021-00809-w

Zhang, Y., Xu, X., Yang, W., & He, G. (2023). Drone-based RGBT tiny person detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, *204*(4), 61-76. https://doi.org/10.1016/j.isprsjprs.2023.08.016

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-iou loss: Faster and better learning for bounding box regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 12993-13000. https://doi.org/10.1609/aaai.v34i07.6999

Zhong, Y., Hu, X., Luo, C., Wang, Xinyu, Zhao, J., & Zhang, L. (2020). WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. *Remote Sensing of Environment*, *250*(1), 112012. https://doi.org/10.1016/j.rse.2020.112012