

Detecting and classifying objects from thermal images : A technical analysis

Ravina Gupta *

Sarika Jain [†]

Amity Institute of Information Technology

Amity University

Noida 201313

Uttar Pradesh

India

Manoj Kumar [§]

School of Computer Science

University of Wollongong

Dubai Knowledge Park

Dubai

U.A.E.

and

MEU Research Unit

Middle East University

Amman 11831

Jordan

Abstract

Object detection and classification in thermal imaging is crucial for a range that operates continuously, such as security, inspection, emergency assistance, and border tracking. When applying several state-of-the-art object detection algorithms to ground-based thermal imaging, the main obstacles include inconsistencies in target size, low-quality

* ORCID-ID: 0000-0002-4791-6361

[†] ORCID-ID: 0000-0002-4752-339X

[§] ORCID-ID: 0000-0001-5113-0639

* E-mail: mahawar.ravina@gmail.com (Corresponding Author)

[†] E-mail: ashusrika@gmail.com

[§] E-mail: wss.manojkumar@gmail.com

images, obstruction, and varying illuminating conditions. This study analyzes the ability of deep-learning algorithms to examine various objects within ground-based thermal images. The novelty of this study lies in its utilization of two network architectures, MobileNetV2 and EfficientNet-B0, which are pre-trained on ImageNet visual images. These networks are typically employed for traditional optical images, but MobileNetv2 in this research, are repurposed and trained on custom datasets of thermal images that have been created using the "Seek Shot" camera. This modification aims to test their performance and suitability in analyzing thermal images, expanding their usefulness beyond their original purpose. Standard metrics such as precision, the experiment, the 5-fold cross-validation used to address the bias of the models. The results demonstrate that the MobileNetV2 model achieves the highest detection accuracy of 96.5% for car classes. The research explores alternative deep learning models Faster RCNN and RetinaNet to compare object detection accuracy in thermal datasets. The results showed that out of all the models, MobileNetv2 had the highest mean average precision (mAP) with 92.6% and EfficientNet-B0 had better accuracy than the other two state-of-the-art models, while RetinaNet had the lowest mAP score of 6.57%.

Subject Classification: 68U10, 94A08.

Keywords: Deep learning, Thermal imaging, Custom dataset, Mobilenet-V2, EfficientNET-B0.

1. Introduction

In computer vision, object detection plays an essential role in locating and identifying objects within an image or video stream. The process involves analyzing an input image or video frame to determine the presence of one or more objects and then accurately defining and representing their positions. An extensive amount of work has been done on object detection using remote sensing images from space, the ground, and aircraft. [1]. Thermal imaging gathers information about objects by using heat energy and infrared radiation. Images can be captured by thermal sensors during the day or at night. In low-visibility conditions, this is exceptionally useful for obtaining information about the images. This technology is great for night vision, especially in scenarios with absolute darkness and difficult environmental conditions like haze, smoke, fog, and smog. [2]. Thermal imaging is widely used in human detection, autonomous transportation, monitoring systems, public security, and healthcare due to the improved performance of thermal detection systems. [3]. Recent studies have also explored thermal behavior in electronic systems, where material-level simulations revealed that electrode and separator design significantly influence heat dissipation [4].

The advancement of convolutional neural networks for image recognition classification significantly improved the performance of

finding objects in thermal imaging. Over the years, an extensive number of algorithms employing deep neural networks for the conventional determination of objects have been developed. Two categories make up the deep learning methodologies: The first category is the ROI (Region-of-Interest) based technique, which generates and classifies regions using the conventional pipeline. All of these methods use a two-stage process and get more accurate information, but they also take longer for computation to complete. The highly regarded two-stage object identification techniques include Mask R-CNN, SPP-Net, R-CNN, Fast R-CNN, and Faster R-CNN in addition to feature pyramid networks. Another category comprises regression-based algorithms, employing a single-stage pipeline for object detection, which means a reduction in execution time. A few prominent single-stage approaches are YOLO (You Only Look Once), DSSD (Deconvolutional Single Shot Detector) [5], and SSD (Single Shot Detector) [6]. A predefined set of anchor boxes is used for identifying objects in both ROI and regression-based methods. The high-definition PASCAL [7] or COCO [8] data sets are generally used to pre-train all of these models to identify objects according to distinct categories. The objects in these datasets are visually clear and of sufficient size that occupy most of the portions of the image. The major problem with the above-mentioned algorithms is that their accuracy goes down when the thermal image is noisy. This encourages us to search for a suitable method for locating objects and classification in thermal imaging, together with the additional noisy view limitations. [9]. In the proposed work, the Mobilenet v2 and Efficient-net B0 pre-trained models have been trained and validated over custom thermal datasets of different classes using K-fold cross-validation. K-fold cross-validation was used to assess the model's performance more accurately and prevent overfitting. Therefore, we fine-tuned the initial configuration of the original algorithm to make it appropriate to detect objects in a custom thermal dataset [10]. Then the detection accuracy of two models, MobileNet v2 and Efficient-Net B0, is compared to evaluating their performance. Additionally, they are compared with two advanced models, SSD MobileNet v2 and Faster RCNN.

The paper is organized into six sections. The second section of this paper discusses relevant literature, section third covers the proposed research methodology, section fourth focuses on experiment results and discussion, the fifth section presents the comparison of MobileNetv2 and EfficientNet-B0 with Faster RCNN and RetinaNet, and Section sixth concludes the entire project.

2. Literature Review

The relevant work on object detection using traditional and innovative deep learning methodologies have been discussed in this section, including deep learning one-stage detector and two-stage detector approaches.

2.1 Object Detection

Object detection has an extensive background in literature with significant advancements made in computer vision in the past few years. In addition to object detection, computer vision also offers persistent data regarding the form and positioning of the object. The literature history of object detection acknowledges two eras, the first before 2014 when traditional object detection methods were used and the emerging techniques of deep learning algorithms in early 2014 which fundamentally altered conventional procedures [11]. The continued advancement of Convolutional Neural Networks (CNN) and Deep Learning (DL), which together have become an efficient approach for feature extraction, has led to an important improvement in the performance of the detection of objects.

2.1.1 Traditional Object Detection

The conventional or traditional object identification techniques generally require the formation of Region of Interest (RoI), feature extraction, and encircling ROIs as essential steps. Most of these methods describe the object attributes using handcrafted features, such as Histogram of Oriented Gradients (HoG) [12], Local Binary Pattern (LBP) [13], and Deformable Parts Model (DPM) [14]. In their study, Teutsch *et al.* [15] technique used Maximum Stable Extremal Regions (MSER) to extract the hot regions, which were then determined by using Discrete Cosine Transform (DCT) and Random Naive Bayes classifier to identify and track pedestrians. The conventional approach to blob extraction and blob classification was employed by Ma *et al.* [16]. Blob classification was achieved via the use of Support Vector Machine (SVM) using HoG and Discrete Cosine Transform (DCT) features, while blob extraction was achieved using region-wise gradient and geometric constraints filtering. Chen, Chen, and Ni [17] introduced a novel approach for detecting and segmenting objects in thermal visualizations. To enhance the recognition of objects accuracy, this research combines the Maximum Similarity-based Region Merging (MSRM) technique with the mean shift algorithm.

Lahouli *et al.* [18] developed a framework to track pedestrians in thermal images. They used Discrete Chebyshev Moments for feature extraction and combined contrast and saliency maps to identify regions of interest. The extracted features were then fed into an SVM classifier for pedestrian tracking and detection. Li *et al.* [19] proposed a novel CNN model that uses LBP feature extraction to classify and characterize facial expressions. The researchers employed four different datasets to train and validate their model and achieved higher classification accuracy than previous research on the same datasets. Table 1 summarizes the relevant specifications, which include the author's name, the dataset utilized, the approach employed, the performance evaluation parameter, and the conclusion. A comparative review of the various methods for classical object detection is given in Table 2.

Table 1
Overview of Different Traditional Approaches to Object-Detection:

Author& year	Dataset	Approach	Performance evaluation parameter	Conclusion
Teutsch <i>et al.</i> (2014) [15]	long-wave infrared (LWIR)	Maximally Stable Extremal Regions (MSER) with modified Random Naive Bayes (RNB)	Accuracy of Detection	achieve high detection rates with low-resolution videos in real-time
Ma <i>et al.</i> (2016) [16]	UAV thermal dataset	CNN, region-wise gradient, and geometric constraints filtering	Blob extraction and classification	97.2%
Chen Y. (2016) [17]	Thermal imaging dataset	semiautomatic segmentation method	Effectiveness and robustness	Proposed novel object segmentation method
Lahouli <i>et al.</i> (2018) [18]	OSU Thermal Pedestrian Database	SVM classifier	Real-time surveillance	Track the human and their suspicious behavior
Li, J. <i>et al.</i> (2020) [19]	Thermal imaging dataset	Local Binary Pattern (LBP) feature extraction	Accuracy and Classification	Achieved higher accuracy

2.1.2 Utilizing Deep Learning for Object Detection

Object detection models in deep learning rely on two methodologies: the first being the single-stage detector, and the other being the two-stage detector. The one-stage detector perspective does not require region recommendations; instead, it makes direct predictions about the class and positioning of objects in an image. Models like SSD and YOLO employ this strategy. The two-stage detector approach first examines a set of region proposals, which are potential object locations in an input image, using a Region Proposal Network (RPN). Then, each region proposal is cropped and classified by a classifier network. Prominent models for object detection like Mask R-CNN and Faster R-CNN use this method. [20].

Mangale and Khambete, [21] conducted research on utilizing both visible and infrared images to improve moving object detection. The fusion of these images provides complementary information that can enhance accuracy and reliability. During the research, background information was removed, and moving objects were segmented using a Mixer of Gaussian (MoG). Blob analysis was used to keep the foreground intact, and image fusion was conducted with the 'AND' logical operation to enhance the object recognition's precision and dependability. Ghenescu *et al.* [22] introduced a approach 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 evaluating 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. Akula, and Sardana [23] presented a novel Nerul Network-based attributes extraction technique for identifying and recognizing objects in thermal imaging. Two already-trained deep learning networks AlexNet and VGG19 are utilized in this study. Analysis of the results demonstrates that the VGG19_fc6 architecture shows an improvement in performance with a 6% classification accuracy. In this research Qian, Ning, and Hu [24] have introduced Mobilenetv3, a compact neural network for visual image classification. The deep-learning approach has been used to attain an improved equilibrium between effectiveness and efficiency. This experiment assesses the performance of MobileNetV3 against several commonly utilized pre-initialized CNN models across diverse imaging

datasets. Civilibal, Cevik, and Bozkurt [25] is a novel study of segmentation and semantic breast cancer recognition in thermal images. The combination of transfer learning and mask RCNN is used to differentiate breast material from surrounding tissues and structures with success. The YOLOv5 framework was utilized by Farooq *et al.* [26] to generate an innovative thermal sense perception system that included adaptive moment estimation (ADAM) and stochastic gradient descent (SGD) optimizers. This study aims to evaluate and improve the most advanced object identification and classifier frameworks available today. They used three distinct testing methodologies to assess the network's performance on test data after training it using datasets that were made available to the public. To retrieve, information from ground-based thermal images and videos Jiang *et al.* [27] have been designed a Neural Network-based 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 streamed 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. Munir *et al.* [28] developed a novel domain adaptation system that effectively m to enable seamless information transfer from visible spectrum images to thermal images, hence improving adaptability across diverse domains. The method transfers low-level features through style consistency using generative adversarial networks, or GANs. Narayanan *et al.* [29] proposed an automatic pedestrian detection framework that integrates YOLO with Histogram of Oriented Gradients (HOG) for feature extraction. Their model was evaluated on the publicly available FLIR dataset and achieved a detection accuracy of 73%. As the literature review demonstrates, most existing studies have used publicly available datasets such as "The FLIR Thermal Imaging Dataset" and any other thermal dataset that might not include specific environmental information (e.g., temperature, humidity, time, etc). Only temperature masks corresponding to different sequences were set for extraction. Most datasets are recorded via UAVs and drones, so some limited viewpoints and angles can affect the model's ability to adapt to different viewpoints in real-world situations. The relevant specifications are listed in Table 2 and include the Category, Name of the Article, Purpose, Proposed Methodology, Used Dataset, and Results and Discussion.

Table 2**Overview of Different Deep Learning-based Approaches to Object-Detection**

Category	Article	Purpose	Methodology	Dataset	Results
One-stage approach	Mangale and Khambete [21]	moving object detection	Mixer of Gaussian (MoG) for background subtraction and segmentation	Fusion of visible and thermal images	Detection of moving objects with 80% Accuracy
One-stage approach	Ghenescu <i>et al.</i> [22]	Accuracy of Detection Low-resolution images	YOLO Dark-Net	Real-time dataset	Detect extremely small objects with 68.7%
Two-stage approach	Akula and Sardana [23]	Image recognition	Alex Net and VGG19	FLIR Dataset	81%
Two-stage approach	Qian, Ning, and Hu [24]	Classification and Efficiency	Mobilenetv3	Multiple image dataset	74.9%
Two-stage approach	Civilibal, Cevik, and Bozkurt [25]	Segmentation and detection	Mask R-CNN	publicly available dataset	86.8%
One-stage approach	Farooq <i>et al.</i> [26]	Enhancement	YOLOv5	Publicly available thermal imaging datasets	71.4%
One-stage approach	Jiang, <i>et al.</i> [3]	Object- Detection	YOLO V5	ground-based thermal images and videos	Archived the fastest detection speed with 86.79%
One-stage approach	Munir <i>et al.</i> [28]	Enhance Style consistency	GAN (generative adversarial network)	FLIR ADAS and KAIST Multi-Spectral	67.78%
One-stage approach	Narayanan <i>et al.</i> [29]	automatic pedestrian detection	YOLO with HOG	Publicly available FLIR Dataset	73%

3. Research Approach

Most object detection and classification models are trained on high-definition RGB datasets. The main problem with these models is that their accuracy decreases when the images are noisy. This encourages us to search for a suitable method for detecting and classifying objects in thermal imaging [30]. To evaluate the generalization capabilities of object detection models trained on RGB images when applied to thermal images, this study assesses the accuracy of pre-trained object detection models such as MobileNetv2 and Effcientnet-B0 using a custom thermal imaging dataset. Precision-recall curves as well as average accuracy are calculated for each experiment.

3.1 Research Methodology

Research methodology is a crucial part of any machine learning model. It outlines the process that was followed to develop the model and evaluate its performance. A series of steps are typically followed by the utilized model for the detection of an object in thermal imaging. Image acquisition is the first step, followed by data preprocessing, model training, model evaluation, model optimization, object detection, and post-processing of a model to estimate its accuracy. The stages are illustrated in Figure. 1 and explained more thoroughly in the next subsection [31].

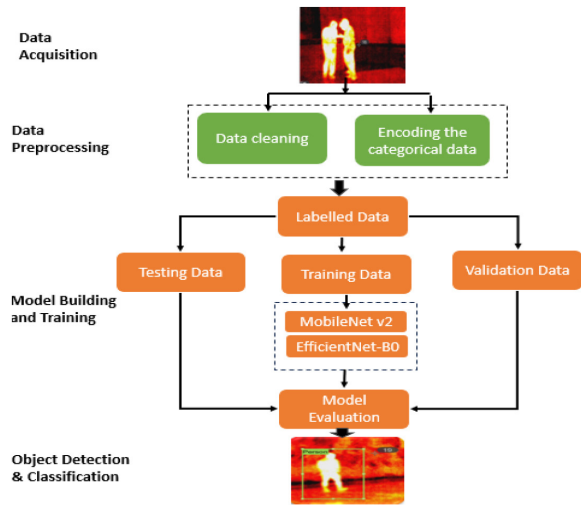


Figure 1
Object Detection Process

3.2 Data Acquisition

Real-time images are collected using a “Seek Shot” thermal imaging camera under challenging weather, different environmental scenarios, and Challenging lighting scenarios such as fog, rain, low light conditions, and environmental noise, among others. The shot thermal camera has approximately 206 x 156 pixels infrared resolution. The thermal camera has a fixed focus lens with a refresh rate of about 9Hz frame rate. The distance for taking thermal images of an object is 120 -250 cm. The dataset is divided into four categories: Person, Car, Cat, and Dog. The findings of the image data acquisition have been demonstrated in Figure. 2.



Figure 2

Sample thermal images from a custom dataset

Subsequently, the real-time thermal video is captured using a shot camera and broken into frames to expand the dataset quickly. The findings from the video data acquisition have been demonstrated in Figure 3.



Figure 3

Real-time thermal video file converted into frames

3.3 Data Preprocessing

Data preprocessing is the preparation of image datasets for further processing. Initial k- fold Cross-validation (CV) technique used to periodically split all data for training and testing. The thermal images that were obtained for this investigation were scaled to 224*224, normalized, and annotated using the MobeilNetV2 and Efficient Net deep learning models. The “VGG Image Annotation” tool is used for image annotation, which involves labeling the images. In this research, the annotation process has been carried out manually. The samples of annotated images from our custom thermal imaging dataset are displayed in Figure 4:



Figure 4

Ground truth labeled thermal infrared image

3.4 Training the Model

Two renowned pre-trained CNN models, MobileNetV2 and EfficientNet-B0, have undergone extensive testing to determine their efficacy. Utilizing a custom thermal imaging dataset that was specially curated for this study, a comparison was conducted. The goal was to distinguish and assess the rate at which each of these two models performed in the specific field of object detection inside thermal imaging.

3.4.1 MobileNet V2

CNN networks for mobile terminals have strengthened in the past several years. MobileNet is the initial mobile computer vision model developed by TensorFlow. To generate a lightweight deep neural network, deep convolutional layers were used in its architecture. In comparison to other neural networks, this resulted in an annotated reduction in the total number of parameters required. The deep learning models are designed

to be lightweight and utilized in mobile applications. The key difference between the classical CNN and In contrast to MobileNet, which arranges the convolution into 1×1 pointwise and 3×3 depth-wise convolutions, CNN uses a single 3×3 convolution layer, followed by the batch norm and a rectified linear unit (ReLU). Different versions of MobileNet models have a unique set of attributes and trade-offs between efficiency and computational accuracy. MobileNet. MobileNetV1 is the original model of MobileNet which has been designed by referencing the conventional VGG architecture and incorporating depth wise separable convolutions. The fact that it has a depth multiplier of 1 indicates that each layer has an equal number of channels [32].

MobileNetV2 [33] was introduced a year after the release of the original MobileNet model, which was demonstrated in 2018. It features a distinctive architecture with a linear bottleneck module and inverted residual blocks for further advancing the capabilities of the network. The improved version enables better performance with a smaller number of parameters. The depth multiplier in this introduced architecture is scalable, which makes it possible for the different trade-offs between efficiency and accuracy. MobileNetV2 has about 3.4 million parameters and achieves an accuracy of around 72% on the ImageNet dataset. The implemented deep neural network architecture has been described in detail. As discussed in the following paragraph the fundamental building block is a bottleneck depth-separable convolution with residuals. Table 3 displays this model's intended structure [34].

The suggested MobileNetV2 architecture consists of 19 layers, with the first layer having 32 filters and the remaining 36 bottleneck levels specified in Table 4. The intermediate layer is utilized for feature extraction, while the final layer is for classification. The layers use average pooling (avgpool) and regular convolution (conv2d), with n being the number of

Table 3

Bottleneck residual block with expansion factor t , stride s , and channel transformation from input (k) to output (k') .

Input	Operator	Output
$\text{height} \times \text{width} \times k$	1×1 conv2d, ReLU6	$\text{height} \times \text{width} \times (tk)$
$\text{height} \times \text{width} \times tk$	3×3 Depthwise conv, stride	$\text{height}/s \times \text{width}/s \times (tk)$
$\text{height}/s \times \text{width}/s \times tk$	Linear 1×1 conv2d	$\text{height}/s \times \text{width}/s \times k'$

repetitions and c being the number of output channels. Since ReLU6 is robust when combined with low-precision computing, it is used as the non-linearity in this [35]. According to the standard for current networks, the kernel size is 3×3 , and during training, dropout and batch

Table 4
Layer-wise specification of MobileNetV2 architecture.

Input	Operator	Expansion Factor	Channels	Number of Blocks	Stride
224	Conv2d	-	32	1	2
112	bottleneck	1	16	1	1
112	bottleneck	6	24	2	2
56	bottleneck	6	32	3	2
28	bottleneck	6	64	4	2
14	bottleneck	6	96	3	1
14	bottleneck	6	160	3	2
7	bottleneck	6	320	1	1
7	Conv2d	-	1280	1	1
1	Avgpool	-	-	1	-
1	Conv2d	-	-	-	-

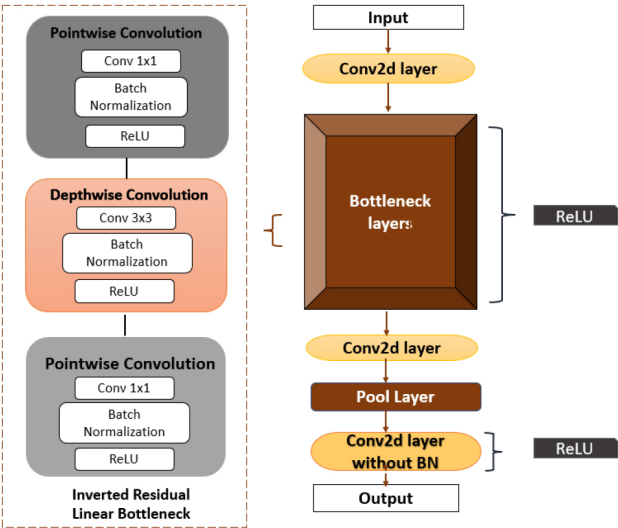


Figure 5
Mobile Net v2 Architecture

normalization are used. The MobileNetV2 architecture's layer-by-layer specification is displayed in Table 4. The overall architecture of the mobileNetv2 framework can be given in Figure. 5 [36], following the detailed layer explanation.

3.4.2 EfficientNet-B0

EfficientNet-B0 is a kind of convolutional neural network (CNN) that has been designed to provide state-of-the-art performance on several computer vision tasks, including segmentation, recognizing objects, and image classification, with the least amount of processing resources and parameters needed. The core idea behind the EfficientNet-B0 architecture is to scale up the CNNs' dimensions in a balanced manner by employing a compound scaling mechanism. Compound scaling involves simultaneously expanding the network's depth, breadth, and resolution. If $2N$ times more processing resources are required, just increase the network depth by N , breadth by βN , and envision size by N . In this instance, an additional grid search on the original model was used to find the constant coefficients α , β , γ , and. EfficientNetB0 uses compound coefficients " $\emptyset$ " to scale the depth, breadth, and resolution of the network evenly. The necessity for extra layers to expand the network's receptive area and more channels to better capture fine-grained patterns on bigger input images justifies the use of the compound scaling approach. [37].

Based on the inspiration from the inverted bottleneck residual blocks of MobileNetV2, the squeeze-and-excitation blocks comprise the foundation of the EfficientNet-B0 network. Furthermore, EfficientNets-B0, with fewer parameters, transfers well and attains the highest possible precision on CIFAR-100 (91.7%), Flowers (98.8%), and three more transfer datasets [38]. The proposed architecture of EfficientNet-B0 contains the initial fully convolution layer with 32 filters, followed by 16 MBConv1_k3x3 bottleneck layers [39]. It is a fundamental building block that combines depth-wise separable convolutions with an inverted bottleneck structure. The expansion factor of 1 denoted by the "1" in MBConv1 means that the number of input and output channels stays the same as detailed in Table 5 [40]. In addition, pooling layers are used to down-sample object feature maps, and the final classification is carried out by fully connected (FC) layers based on the acquired features. The kernel size varies between 3×3 and 5×5 according to the standard network specification.

Table 5
Baseline network of EfficientNet-B0 describes layers, input resolution, and output channels.

Stage	Operator	Resolution	#channels	#layers
1	Conv3 × 3	224 × 224	32	1
2	MBCnv1,k3 × 3	112 × 112	32	1
3	MBCnv6,k3 × 3	112 × 122	16	2
4	MBCnv6,k5 × 5	56 × 56	24	2
5	MBCnv6,k3 × 3	28 × 28	24	2
7	MBCnv6,k3 × 3	28 × 28	80	2
8	MBCnv6,k5 × 5	14 × 14	112	4
9	MBCnv6,k5 × 5	14 × 14	192	4
10	Conv1x1&pooling&fc	7 × 7	1280	1

The EfficientNet-B0 model’s architectural configuration is shown in Figure. 6, following the comprehensive layer specifications.

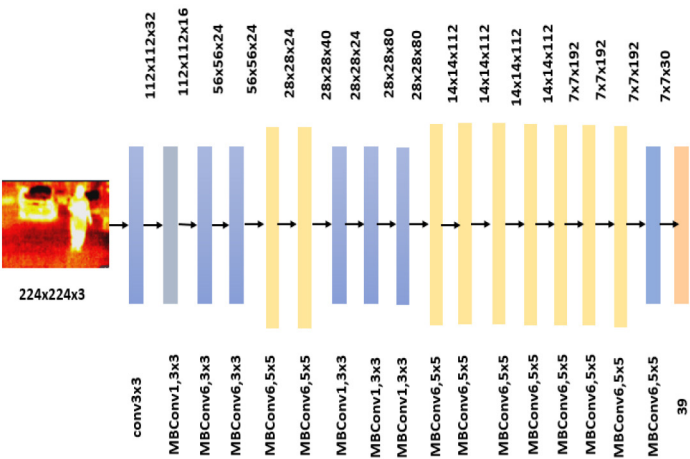


Figure 6
The architecture of EfficientNet-B0

3.5 Model Evaluation Metrics

- a. Accuracy: Accuracy is one metric used to evaluate the effectiveness of deep learning models. It counts the number of positive samples that were predicted and proved to be positive, as well as the number

of negative samples that were projected and proved to be negative. It illustrates the extent to which the created models were able to distinguish between genuine positive and negative samples across the board in the dataset [41]. The performance is assessed using the model's identification accuracy, as stated in equation (1). Formally, accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Number of Predictions}}{\text{Total Number of Predictions}}$$

To assess the accuracy of binary classification, it's important to measure the proportion of correctly predicted instances, which includes true positives and true negatives, against the total number of instances. The formula for calculating accuracy is readily available.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

In the context of this equation, TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

- b. **Cross-Entropy Loss:** In machine learning, cross-entropy is a frequently used loss function that majorly works for classification models. For the designated output classes, it calculates the discrepancy between the actual range of probability and the expected probability distribution. A probability distribution across all potential classes can be used to illustrate the outcome of a classification problem. The cross-entropy loss function pays closer attention to the utilized model when it makes a confident wrong prediction than when it makes an uncertain wrong prediction. It is commonly used in neural network models for training. The purpose is to reduce the cross-entropy loss throughout the predicted and actual labels given in eq (2).

For binary classification, cross-entropy loss defined as:

$$\text{Cross - Entropy Loss} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M p(x_{ij}) \log(q(x_{ij})) \quad (2)$$

In this context, where, N , M represents the training set's sample and class counts. $p(x_{ij})$ represents the actual probability that the i -th sample will be assigned to the j -th class. $q(x_{ij})$ is the probability of the i -th sample assigned to the j -th class produced by the model.

- c. Intersection Over Union (IoU), a prevailing classification metric that quantifies the region of the coinciding portion between the actual and the predicted boxes, is the most utilized for assessing detection accuracy. Therefore, corresponding to IoU, the L2 distance is measured between the ground truth and the model prediction. The performance metrics include average precision (P), recall (R) and F1-score as shown in the following equations (3), (4) and (5). Where precision is averaged over multiple intersection over union (IoU) values [42]. Recall evaluates the model's ability to correctly identify all positive cases or the percentage of true positive cases that are correctly identified [43]. In contrast, the F1 score is a balanced measure of model performance that combines recall and precision into one.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F - I score = \frac{TP}{TP + 1 / 2(FP + FN)} \quad (5)$$

In context of evaluation matrix, TP refers to True Positives.

TN refers to True Negatives. FP refers to False Positives, and FN refers to False Negatives.

4. Experimental Results and Discussion

This section illustrates the findings from using the deep learning-based models for object detection on the test and validation sets of the thermal dataset. In the parts that follow, the outcomes of these trials are thoroughly explained and examined, providing a summary of the efficiency with which the models perform when it comes to accurately recognizing objects in thermal images.

4.1 Dataset Description

The custom thermal imaging dataset contains 2000 thermal images categorized into four different classes. The parameters of the custom thermal dataset are as follows: Maximum resolution of the captured thermal image is 648 x 480 wherein the approximate camera distance from

Table 6
Parameters Details of Custom Thermal Dataset

Parameter	Custom Thermal Dataset
Resolution	648 x 480
Number of Images	2000
Camera Distance	120 -250 cm
Collective object count	2027
Mean objects per thermal image	2
Median pixel Area for objects	346
Median Area of object-image overlap	0.46%

the captured object is approximately 120 to 250 cm, which can be maximized or minimized depending on the size of the target object. The collection has 2027 objects in total, divided into four classes: person, car, dog, and cat. There are roughly two objects on average for each image, and the median class of object size determined by pixel count is almost 241. In the image, the average overlap area between objects is approximately 0.46%. The details of the parameters are given in Table 6.

The images are segregated as per their classes have been demonstrated in Figure. 7. The train test split has limitations as it reduces training data size and introduces bias. To address these issues, we used the Cross-validation (CV) technique by periodically splitting all data for training and testing. The dataset is divided into k equal sections at random in the cross-validation approach. While the remaining portions are used as training data, each component serves as validation data. The steps involved in conducting K-Fold cross-validation are as follows:

- a. Decide the exact amount of K.
- b. Separate the data into a validation set and a train set.
- c. Obtain the model's evaluation outcomes.
- d. Compute the Average Precision (AP)'s total mean.

In this experiment, set the value of K to be 5. Hence, the overall mean AP equation is:

$$Mean AP = \frac{1}{5} \sum_{i=1}^k AP \quad (6)$$

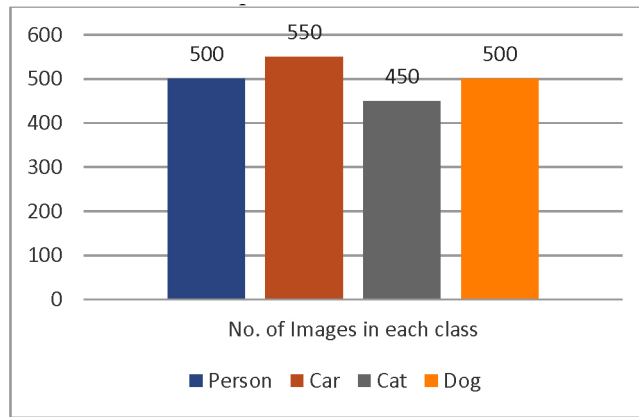


Figure 7
Number of images in each class

The process is carried out k times, with a unique collection of events used as a test set each time. The average predictions of the k test accuracy is obtained. Images with 224×224 dimensions are resized using the maintain aspect ratio resizer function. If the larger side is greater than 224, image scaling is used to make the long edge 224. If not, this function resizes the smaller image length to 224 pixels. The images maintain their aspect ratio.

4.2 Implementation Details

The MobileNetv2 and EfficientNet-B0 algorithms are trained along with their backbone architecture. The models are trained and evaluated using 2000 thermal images. Images are scaled and normalized based on the specific image size required by each network. The number of image classes was used to fine-tune the fully interconnected final layer. The MobileNet v2 networks have a learning rate of 1.66×10^{-3} , or 0.00166. With a maximum epoch of 100, a batch size of 10, a learning rate drop period of 0.2, and total iterations of 10000, it is the ideal learning rate for the MobileNetv2 model. Similarly, EfficientNet-B0 networks have their learning rate set to 0.001. A maximum epoch of 60, 10 iterations per epoch, a total of 600 iterations, a batch size of 10, and a learning rate drop period of 0.2 are among the training parameters that have been set. Grid Search, the most used technique for hyperparameter configuration, has been used in this research to fine-tune the hyperparameters. It is an exhaustive

approach that tests all combinations of hyperparameters given to the grid configuration.

The grid search algorithm can be used to identify the optimal hyperparameters. It is still regarded as state-of-the-art after years of scientific research in global optimization and other proposed hyperparameter optimization techniques. This is due to the following reasons:

- a. Implementation is simple and efficient.
- b. It uses an automated process to find a better λ value compared to manually optimizing it.
- c. It is also reliable and is of low dimensionality.

Table 7 shows the details of input resolution, learning rate, batch size and number of epochs hyperparameters for both MobilleNet v2 and EfficientNet-B0 networks.

Table 7

Hyperparameters Details of MobilleNet v2 and EfficientNet-B0 networks

Model	Input Resolution	Learning Rate	Batch Size	No. of epochs
MobilleNet v2	224x224	0.00166	10	100
EfficientNet-B0	224x224	0.001	10	60

The experiment's simulation environment uses an Intel(R) Core (TM) CPU i5-1035G1. The memory is 8GB, and the GPU is an NVIDIA MX 130. The TensorFlow framework and Python are used for the implementation and classification of the algorithm.

The results in Figure 8 show the training epoch-accuracy and epoch-loss curves for the training and validation subsets. Table 8 displays an overview of the total number of actual parameters for both models, as well as the trainable prams (Pre-training stage) and trainable prams (Fine-tuning stage).

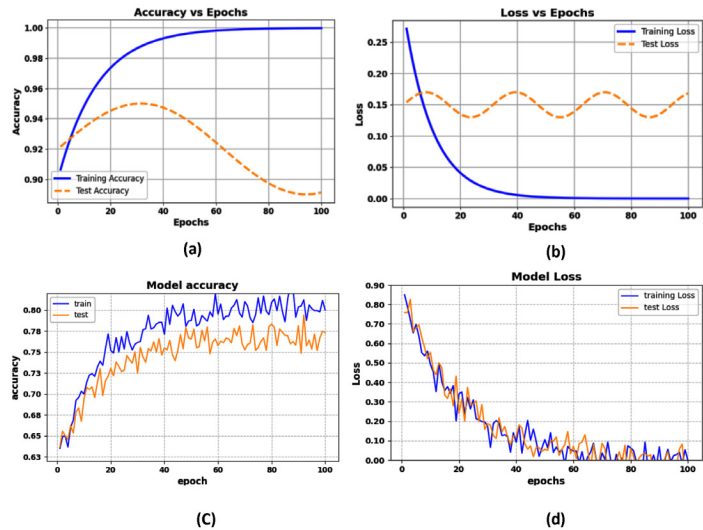


Figure 8

Epoch-wise accuracy and loss curve: (a) MobileNetV2 Epoch-Accuracy and (b) Epoch- Loss Curves (c) EfficientNetB0 Epoch-accuracy and (d) Epoch-loss curves

Table 8
Total Count of Parameters in Both Stages

Models	Total parameters	Trainable prams (Pre-training stage)	Trainable prams (Fine-tuning stage)	Dropout rate
MobileNetV2	2,626,821	3,68,837	2,231,429	0.2
EfficientNet-B0	5,334,568	295,109	295,109	0.2

Figure 8(a) evaluation findings reveal that the MobileNetv2 network architecture, when trained and tested on a proprietary thermal imaging dataset, performed with a detection accuracy of almost 100%. Figure 8(b) shows that the loss on the training and test samples is approximately 0.5%. In contrast, Figures 8(c) and (d) show that EfficientNet-B0 lost 0.2% and obtained an overall accuracy of 0.82% on the test and train datasets. The matrix depicting the confusion results for MobileNetV2 and EfficientNet B0 are given in Figure 9(a) and 9(b) respectively. The weights file of the trained MobileNetv2 model and EfficientNet-B0 are saved, and the trained model is tested using image files from a custom dataset as input images

for prediction. A qualitative evaluation of the predicted results of the models are displayed in Figure 10 and Figure 11.

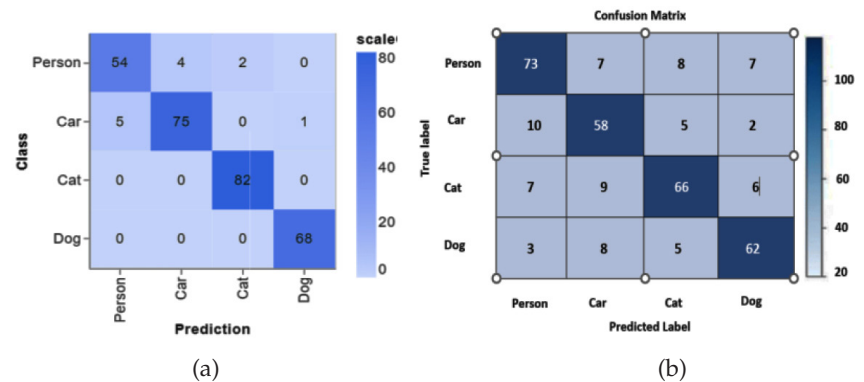


Figure 9
(a) Confusion matrix of MobileNetV2
(b) Confusion matrix of EfficientNet-B0



Figure 10
Example showcasing detection and classification outcomes of MobileNetv2 Model

The research’s findings are evaluated based on assessment metrics such as precision, recall, and fi- score, which are evaluated using the confusion matrices of the classification model from Figures 9(a) and 9(b). The higher precision value shows that positively labeled classes are inevitably positive, and the elevated recall value suggests that the class is accurately categorized. Consequently, two scenarios can be considered:

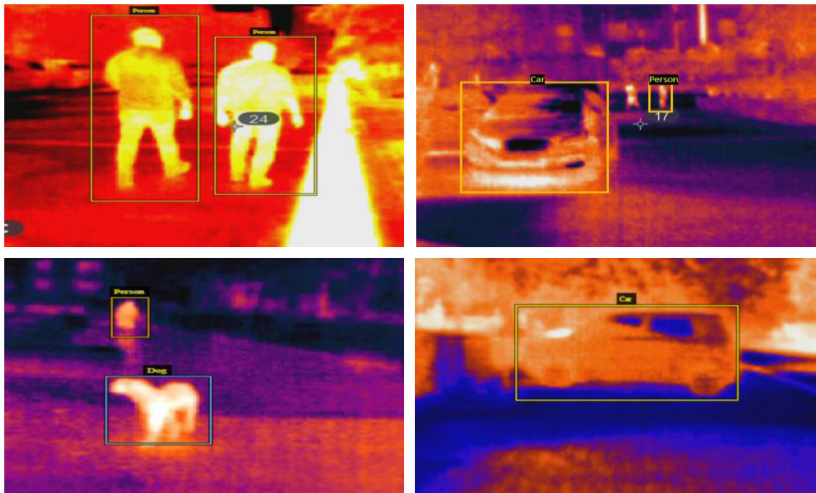


Figure 11
Example showcasing detection and classification outcomes of EfficientNet-B0 model.

- a. High recall and inadequate precision indicate that even though a substantial number of positive cases are correctly identified (low FN), there are a significant number of false positives.
- b. As was the case with EfficientNet-B0, the low recall and high precision meant that we were unable to detect a large portion of positive samples, even in indisputably positive cases (low FP) are predicted to occur.

Table 9
Performance metrics of MobileNetv2 models.

Average Precision	K-fold 1	K-fold 2	K-fold 3	K-fold 4	K-fold 5	Mean
Precision	0.865	0.823	0.886	0.898	0.905	0.875
Recall	0.895	0.916	0.948	0.985	0.980	0.945
F1-score	0.874	0.867	0.915	0.939	0.967	0.912
Person	0.912	0.896	0.935	0.946	0.958	0.920
Car	0.953	0.937	0.968	0.976	0.985	0.965
Cat	0.923	0.947	0.908	0.936	0.945	0.932
Dog	0.890	0.957	0.936	0.955	0.955	0.938

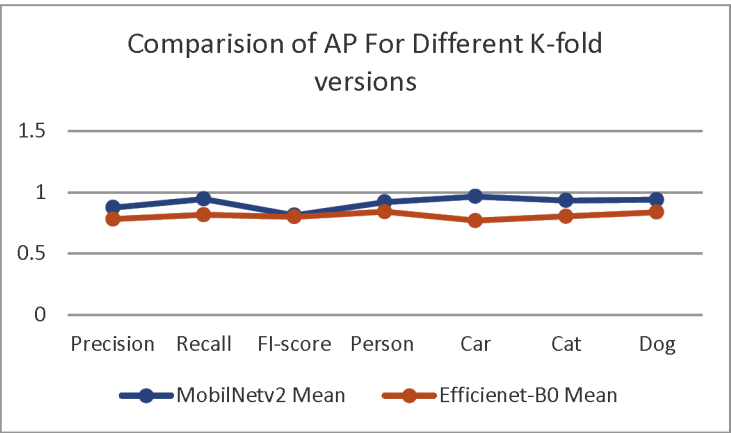


Figure 12
Comparison of Models for K-fold

Table 10
Performance metrics of Efficientnet-B0

Average Precision	K-fold 1	K-fold 2	K-fold 3	K-fold 4	K-fold 5	Mean
Precision	0.750	0.785	0.786	0.770	0.812	0.780
Recall	0.786	0.80	0.833	0.815	0.846	0.816
FI-score	0.767	0.792	0.808	0.802	0.828	0.799
Person	0.816	0.842	0.828	0.860	0.855	0.840
Car	0.760	0.772	0.745	0.774	0.789	0.768
Cat	0.816	0.822	0.787	0.786	0.806	0.803
Dog	0.880	0.865	0.883	0.892	0.903	0.837

The results in Table 9 and Table 10 show the precision of MobileNetv2 and EfficientNet-B0 their respective batch size demonstrated in Table 7 using k-fold cross-validation. Performance metrics of both models indicate that it's extremely apparent that the precision value is higher for mobileNetv2 and that AP for car classification is 96.5% in comparison to EfficientNet-B0 is 76.8%. Similarly, the accuracy values for MobileNetv2 in the other classes of human, car, and cat are 91.4%, 94.9%, and 97.6%, respectively, whereas the accuracy values for EfficientNetB0 model are 78.4%, 70.5%, and 79.4%, respectively. In the case of speed, Efficientnet-B0 achieved better execution speed than MobileNetv2. The comparison of AP displayed in Figure 12 which demonstrates that MobileNetV2 has a higher

Average Precision (AP) than EfficientNet-B0 across k-fold cross-validation. Specifically, MobileNetV2 has an AP of 96.5% for the car class, while EfficientNet-B0 has an AP of 76.8% for the same class, averaging over five folds.

5. Comparison of MobileNetv2 and EfficientNet-B0 with Faster RCNN and RetinaNet

To compare MobileNetv2 and EfficientNet-B0 with other object detection algorithms, we choose Faster RCNN and RetinaNet. RetinaNet is specifically suitable for thermal imaging object detection due to its novel focal loss function, while Faster R-CNN is a widely used algorithm that generates object proposals using a region proposal network. Both algorithms are suitable for object detection in thermal imaging. The average precision (AP) of several models was assessed using a customized thermal imaging dataset with a batch size of 10 and a learning rate of 0.001. The results in Table 11 and Figure 13, showed that MobileNetv2 achieved the highest accuracy among all the models and EfficientNet-B0 outperformed the other two advanced detection models, while RetinaNet had the lowest AP score of 0.657. In the car class, MobileNetv2 had the highest AP score of 0.965, but the mean AP of RetinaNet was lower due to its inability to identify other classes.

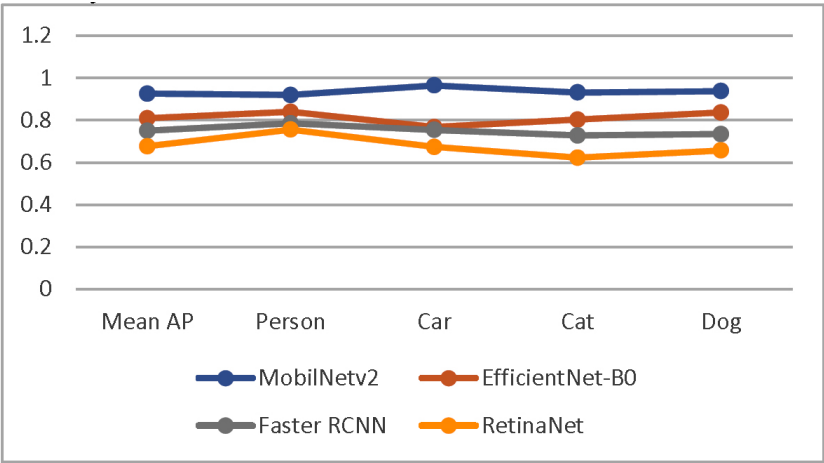


Figure 13
Model Comparison Based on Average Precision (AP) Scores

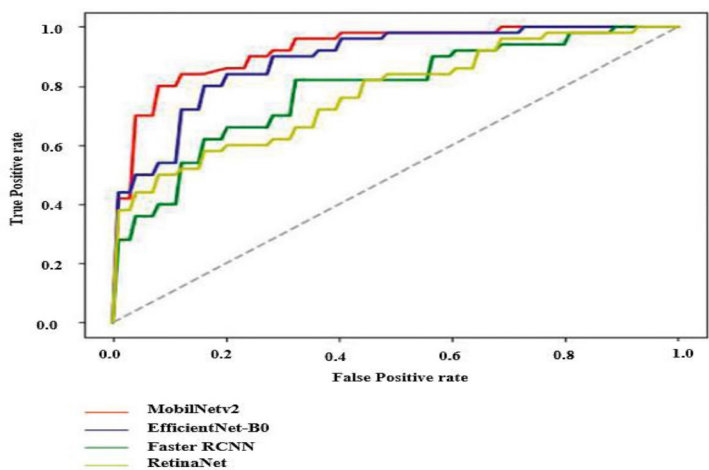


Figure 14
Comparison of Implemented Models using ROC Curve

Table 11
Comparative Analysis of AP

Average Precision	Mobile Netv2	Efficient Net-B0	Faster R-CNN	Retina Net
Mean AP	0.926	0.810	0.751	0.667
Person	0.920	0.840	0.786	0.756
Car	0.965	0.768	0.755	0.675
Cat	0.932	0.803	0.728	0.623
Dog	0.938	0.837	0.735	0.657

The considered models’ ROC curve is displayed in Figure 14. After comparing the performance of several models using the ROC curve analysis, it was discovered that MobileNetV2 had the greatest True Positive Rate (TPR) and the lowest False Positive Rate (FPR). EfficientNet-B0 followed closely with a slightly higher FPR. Faster RCNN had balanced TPR and FPR, while RetinaNet performed the least effectively with the highest FPR and lowest TPR. Overall, MobileNetV2 was the most reliable model for this dataset, followed by EfficientNet-B0, Faster RCNN, and RetinaNet.

6. Conclusion

To detect objects in thermal images, this research has analyzed and assessed the deep learning models MobileNetV2 and EfficientNet-B0. Deep neural networks are achieving exceptional results on difficult datasets, revolutionizing the object detection field. A new aspect of object detection using thermal imaging is introduced in this paper. The developed deep learning model can detect and classify objects from custom thermal imaging. The experiment is done in four different classes which include: person, car, cat, and dog. To validate and compare the efficiency of our tuning process, we ran the experiment using the custom thermal imaging dataset with cross-validation. Standard metrics such as precision, accuracy, recall, and F1-score were used to assess the efficiency of the models. The obtained results demonstrate that the MobileNetV2 model was more appealing when it came to detection accuracy for car classes by 96.5% respectively. Research explores alternative deep learning models Faster RCNN and RetinaNet to compare accuracy of object detection in the thermal dataset. The results showed that MobileNetV2 had the highest AP with 92.6% among all the models and EfficientNet-B0 had better accuracy than the other two state-of-the-art models, while RetinaNet had the lowest AP score of 0.657. The proposed method uses a limited number of thermal images for training and a small number of categories, which may affect the model's ability to generalize. Also, the experiment was not tested on other thermal datasets, raising concerns about its adaptability to diverse thermal imaging scenarios. To improve object classification tasks in various applications, future studies should focus on enhancing the size and diversity of the dataset. This will help the model to capture the nuances of thermal imagery more effectively, resulting in better performance.

Data availability statement

The corresponding author can provide the thermal datasets created for this work upon an appropriate request.

Declaration of Competing Interest

The authors state that they have no conflicts of interest to reveal.

References

- [1] T.Z. Xiang, G.S. Xia and L. Zhang, "Mini-unmanned aerial vehicle-based remote sensing: Techniques, applications and prospects," *IEEE Geoscience Remote Sensing*, vol. 7, pp. 29–63 (2019).
- [2] P. Lin, H. Yang, S. Cheng, F. Guo, L. Wang and Y. Lin, "An improved YOLOv5s method-based bruises detection on apples using cold excitation thermal images", *Postharvest Biology and Technology*, vol.199, no. 112280, pp.0925-5214 (2023).
- [3] C. Jiang, H. Ren, X. Ye, J. Zhu, H. Zeng, Y. Nan, M. Sun, X. Ren and H. Huo, "Object detection from UAV thermal infrared images and videos using YOLO models," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, pp. 1569-8432 (2022).
- [4] S. B. A. Aaditya, P. B. Karandikar, M. Jain, J. Suryawanshi, A. Bellam, D. Chavan, R. Holmukhe, P. S. Chaudhari, and A. Gandhar, "Thermal profiling of ultracapacitor modules: A layered analysis," *Journal of Information and Optimization Sciences*, vol. 45, no. 2, pp 291–303 (2024).
- [5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu and A.C. Berg, "Ssd: Single shot multibox detector," *In Proceedings of the European Conference on Computer Vision*, pp. 21–37 (2017).
- [6] C.Y. Fu, W. Liu, A. Ranga, A. Tyagi and A.C. Berg, "Dssd: Deconvolutional single shot detector," arXiv:1701.06659 (2017).
- [7] M. Everingham, L.V. Gool, C.K. Williams, J. Winn and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Mathematics*, vol. 88, pp. 303–338 (2010).
- [8] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C.L. Zitnick, "Microsoft coco: Common objects in context," *In Proceedings of the European Conference on Computer Vision*, pp. 740–755 (2014).
- [9] N.D. Nguyen, T. Do and T.D. Ngo, "An Evaluation of Deep Learning Methods for Small Object Detection" *Journal of Electrical and Computer Engineering*, 3189691 (2020).
- [10] S.S. Sumit, D.R.A. Rambli and S. Mirjalili, "Vision-Based Human Detection Techniques: A Descriptive Review," vol. 9, pp. 42724–42761 (2021).
- [11] K.R. Akshatha, A.K. Karunakar, S.B. Shenoy, A.K. Pai, N.H. Nagaraj and S.S. Rohatgi, "Human Detection in Aerial Thermal Images Using Faster R-CNN and SSD Algorithms," *Electronics*, vol. 11, no. 1151 (2022)

- [12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893 (2005).
- [13] P. Viola and M.J. Jones, "Robust real-time face detection," *International Journal of Computer Mathematics*, vol.57, pp 137–154 (2004).
- [14] P.F. Felzenszwalb, R.B. Girshick, D. McAllester and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp.1627–1645 (2010).
- [15] M. Teutsch, T. Mueller, M. Huber and J. Beyerer, "Low resolution person detection with a moving thermal infrared camera by hot spot classification," *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 23, pp. 209–216 (2014).
- [16] Y. Ma, X. Wu, G. Yu, Y. Xu and Y. Wang, "Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery," *Sensors*, vol. 16, no. 446 (2016).
- [17] Y. Chen, W. Chen, and H. Ni, "Image segmentation in thermal images," *IEEE International Conference on Industrial Technology (ICIT)*, pp. 1507-1512 (2016).
- [18] I. Lahouli, R. Haelterman, Z. Chtourou, G. De Cubber and R. Attia, "Pedestrian detection and tracking in thermal images from aerial MPEG videos," *In Proceedings of the 13th International Joint Conference on Computer Vision Imaging and Computer Graphics Theory and Applications*, vol. 5, pp. 487–495 (2018).
- [19] J. Li, k. Jin, D. Zhou, N. Kubota and Z. Ju, "Attention mechanism-based CNN for facial expression recognition," *Neurocomputing*, vol. 411, pp.340-350 (2020).
- [20] P. Poirson, P. Ammirato, C.Y. Fu, W. Liu, J. Kosecka and A.C. Berg, "Fast single shot detection and pose estimation," *Fourth International Conference on 3D Vision (3DV)*, pp. 676–684 (2017).
- [21] M. S. Mangale and M. Khambete, "Moving Object Detection using Visible Spectrum Imaging and Thermal Imaging," *International Conference on Industrial Instrumentation and Control (ICIC)*, pp. 590-593 (2015).
- [22] V. Ghenescu, E. Barnoviciu, S.V. Carata, and M. Ghenescu, R. Mihaescu, M. Chindea, "Object Recognition on Long Range Thermal Image Using State of the Art DNN," pp. 1-4 (2019).

- [23] A. Akula, and H.K. Sardana, "Deep CNN-based Feature Extractor for Target Recognition in Thermal Images," *TENCON 2019 IEEE Region 10 Conference (TENCON)*, pp. 2370-2375 (2019).
- [24] S. Qian, C. Ning, and Y. Hu, "MobileNetV3 for Image Classification. (2021)," *IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pp.490-497 (2021).
- [25] S. Civilibal, K.K. Cevik, and A. Bozkurt, "A deep learning approach for automatic detection, segmentation and classification of breast lesions from thermal images," *Expert Systems with Applications*, vol. 212, no.118774 (2023).
- [26] M. A. Farooq, P. Corcoran, C. Rotari and W. Shariff, "Object Detection in Thermal Spectrum for Advanced Driver-Assistance Systems (ADAS)," vol. 9. pp. 156465-156481 (2021).
- [27] C. Jiang, H. Ren, X. Ye, J. Zhu, H. Zeng, Y. Nan, M. Sun, X. Ren and H. Huo, "Object detection from UAV thermal infrared images and videos using YOLO models," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, no. 102912, pp. 1569-8432 (2022).
- [28] F. Munir, S. Azam, M. Rafique, A. Sheri and M. Jeon, "Exploring thermal images for object detection in underexposure regions for autonomous driving," *Applied Soft Computing*, vol. 121, no. 108793 (2022).
- [29] A. Narayanan, R.D. Kumar, R. RoselinKiruba, and T.S. Sharmila, "Study and analysis of pedestrian detection in thermal images using YOLO and SVM," In *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 431-434 (2021).
- [30] D. Rahadiyan1, S. Hartati1, M. Wahyono1, and A. Nugroho, "Percentage of Phosphorus Deficiency Estimation in Chili Leaves Images," *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, vol. 7, pp.2089-3272 (2019).
- [31] R. Gupta, S. Jain, and M. Kumar, "Enhanced Thermal Object Detection and Classification with MobileNetV3: A Cutting-Edge Deep Learning Solution," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, pp. 1-6 (2024).
- [32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510-4520 (2018).
- [33] Q. Xiang, X. Wang, R. Li, G. Zhang, J. Lai, and Q. Hu, "Fruit image classification based on mobilenetv2 with transfer learning tech-

- nique," *Proceedings of the 3rd International Conference on Computer Science and Application Engineering*, pp.1-7 (2019).
- [34] G. Andrew Z. Menglong, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR* (2017)
 - [35] S. Qian, C. Ning and Y. Hu, "MobileNetV3 for Image Classification," *IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pp. 490-497 (2021).
 - [36] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, Long Beach, CA, USA, pp. 6105–6114 (2019).
 - [37] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv*, pp.1704.04861 (2017).
 - [38] R. Gupta, S. Jain, and M. Kumar, "Role of Thermal Images in Various Applications of Computer Vision," *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom, pp. 1-6 (2023)
 - [39] M. Fatemeh, D. Javan, F. Farzaneh and S. Farhad, "Potential evaluation of visible-thermal UAV image fusion for individual tree detection based on convolutional neural network," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, no. 103011 (2022).
 - [40] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 658–666 (2019).
 - [41] X. Yang, R. Guo and H. Li, "Comparison of multimodal RGB-thermal fusion techniques for exterior wall multi-defect detection," *Journal of Infrastructure Intelligence and Resilience*. vol. 2, no. 2, pp. 2772-9915 (2023).
 - [42] R. Gupta, S. Jain, and M. Kumar, "Thermal Vision for Airfield Safety: A Dynamic FOD Detection Framework," *IAENG International Journal of Computer Science*, vol. 52, no. 4, pp.1178-1186 (2025).
 - [43] J. Yang, W. Wang, G. Lin, Q. Li and Y. Sun, "Infrared Thermal Imaging-Based Crack Detection Using Deep Learning," *IEEE Access*, vol. 7, pp. 182060-182077 (2019).

Received February, 2025