Zhuang *et al. Intell Robot* 2024;4(3):276-92
DOI: 10.20517/ir.2024.18

Intelligence & Robotics

Research Article

Open Access

An in-vehicle real-time infrared object detection system based on deep learning with resource-constrained hardware

Tingting Zhuang[1], Xunru Liang[2], Bohuan Xue[3], Xiaoyu Tang[1,2]

[1]School of Data Science and Engineering, Xingzhi College, South China Normal University, Shanwei 516600, Guangdong, China.
[2]School of Physics, South China Normal University, Guangzhou 510006, Guangdong, China.
[3]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China.

Correspondence to: Prof. Xiaoyu Tang, School of Data Science and Engineering, Xingzhi College, South China Normal University, Ma Gong Street, Shanwei 516600, Guangdong, China. E-mail: tangxy@scnu.edu.cn

How to cite this article: Zhuang T, Liang X, Xue B, Tang X. An in-vehicle real-time infrared object detection system based on deep learning with resource-constrained hardware. *Intell Robot* 2024;4(3):276-92. http://dx.doi.org/10.20517/ir.2024.18

Received: 13 May 2024 **First Decision:** 7 Aug 2024 **Revised:** 6 Sep 2024 **Accepted:** 11 Sep 2024 **Published:** 24 Sep 2024

Academic Editor: Simon X. Yang **Copy Editor:** Dong-Li Li **Production Editor:** Dong-Li Li

Abstract

Advanced driver assistance systems primarily rely on visible images for information. However, in low-visibility weather conditions, such as heavy rain or fog, visible images struggle to capture road conditions accurately. In contrast, infrared (IR) images can overcome this limitation, providing reliable information regardless of external lighting. Addressing this problem, we propose an in-vehicle IR object detection system. We optimize the you only look once (YOLO) v4 object detection algorithm by replacing its original backbone with MobileNetV3, a lightweight feature extraction network, resulting in the MobileNetV3-YOLOv4 model. Furthermore, we replace traditional pre-processing methods with an Image Enhancement Conditional Generative Adversarial Network inversion algorithm to enhance the pre-processing of the input IR images. Finally, we deploy the model on the Jetson Nano, an edge device with constrained hardware resources. Our proposed method achieves an 82.7% mean Average Precision and a frame rate of 55.9 frames per second on the FLIR dataset, surpassing state-of-the-art methods. The experimental results confirm that our approach provides outstanding real-time detection performance while maintaining high precision.

Keywords: Infrared object detection, in-vehicle system, lightweight, limited hardware resources, real-time

1. INTRODUCTION

With the speedy development of transportation systems, the increasing number of vehicles is leading to more traffic problems, and the risk of traffic accidents continues to rise[1]. According to research, around 1.35 million people die globally because of traffic accidents each year[2]. Financial costs due to traffic accidents add up to 1%-3% of the world's gross domestic product[3]. A large proportion of traffic accidents occurs under reduced visibility conditions when the view of drivers is greatly limited[4]. How to minimize the occurrence of traffic accidents has been a hot issue, and advanced driver assistance systems (ADAS) are considered a feasible way to achieve this goal[5]. As an important way to obtain information from the external environment, machine vision is considered to be the core technology of ADAS. In recent years, infrared (IR) imaging technology has been adopted to obtain information, which makes it possible to implement in-vehicle IR target detection systems[6]. IR images are not easily affected by changes in external light, which is an outstanding advantage compared with visible images. These images can provide more valuable information for intelligent automotive systems[7], especially in low-visibility weather such as heavy rain and fog. Therefore, the application of IR imaging technology is a considerable choice to make the safe driving system more reliable. With the development of technology, the research of computer vision is iterated year by year. In 2005, Dalal and Triggs proposed the Histogram of Gradient (HOG) detector[8] which became an important improvement in Scale Invariant Feature Transform and Shape Contexts at that time. Related technologies are widely used in computer vision applications and lay an important foundation for many later detection methods. In 2014, Girshick *et al.* proposed the Region with CNN features (R-CNN)[9], which selects possible object boxes from a set of object candidate boxes through the selective search algorithm and then resizes the images in these selected object boxes to a fixed-size image. Later the algorithm feeds them to the trained CNN model to extract features and finally sends the extracted features to the classifier to predict whether the image in the object box has a target to be detected. And further, predict which category the detection target belongs to. In 2016, Redmon *et al.* proposed you only look once (YOLO) v1[10], which is the first stage of the deep learning detection algorithm. Its detection speed is very fast; the idea of the algorithm is to divide the image into multiple grids, and then predict the bounding box for each grid at the same time and give the corresponding probability. Based on this idea, YOLOv1 has been continuously developed into v2, v3, v4, v5 and other versions. In 2018, Law and Deng proposed the CornerNet algorithm[11]. As the pioneer of the Anchor technology route, the network uses a new target detection method, which transforms the detection of the target bounding box by the network into a pair of key points (the upper left corner and the lower right corner).

This work proposes a method to develop an in-vehicle IR target detection system. The key contributions of this paper are:

• Building on the Image Enhancement Conditional Generative Adversarial Network (IE-CGAN) proposed by Kuang *et al.*, we introduced an innovative improvement, resulting in the IE-CGAN inversion algorithm[12]. This algorithm enhances input images, replacing traditional pre-processing methods.

• The YOLOv4 model is optimized by replacing its backbone network, CSPDarknet53, with MobileNetV3. This replacement has been shown to effectively enhance the system's real-time detection capabilities while maintaining high detection accuracy.

• The model is deployed on a Jetson Nano, an edge device with limited hardware resources, culminating in a fully integrated system that combines both hardware and software.

2. RELATED WORKS

In ADAS, detecting vehicles and pedestrians is a core task. Currently, many effective methods have been proposed, including two aspects: the YOLO series algorithm and the IR target recognition part. Among them,

YOLO series algorithms provide efficient and accurate solutions for real-time target detection; IR target recognition is usually based on IR images for target recognition, which can prove useful in night vision and adverse weather conditions.

2.1. Deep learning networks

Currently, deep learning methods for target detection are primarily categorized into two types: two-stage and one-stage detection algorithms. One-stage detection algorithms such as YOLO and Single Shot Multi-Box Detector (SSD) typically use a Fully Convolutional Network (FCN) to directly predict from the original image. While they offer fast processing speed, their accuracy in detecting small objects is relatively low. Two-stage detection algorithms, such as R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN, capture target details more effectively but operate at slower detection speeds.

YOLO is a fast and efficient target detection algorithm introduced by Redmon *et al.* in 2016[10]. Compared to traditional two-stage object detection algorithms such as R-CNN, YOLO is a single-stage detection algorithm capable of achieving real-time detection without compromising accuracy. In a pedestrian detection experiment[13], a Scale-Aware Fast (SAF) R-CNN model was introduced, using multiple subnetworks to detect pedestrians at different scales, then adaptively combining the outputs to generate the final result. Fan *et al.* proposed a data fusion CNN architecture called RoadSeg, which can extract and fuse features from RGB images and infer surface normal information for accurate free space detection[14]. In another study, a DS-Net was suggested to solve the problem that current neural networks primarily focus on single-task single-task vision scenarios[15]. The DS-Net was a multitask convolutional neural network designed for AR-HUD environment perception. Li *et al.* proposed a vision-based framework for target detection and recognition in autonomous driving, utilizing an improved YOLOv4 model that reduced the total model parameters by 74%[16]. A U-type generative adversarial network (GAN) was first developed to fuse visible and IR images. YOLOv3 combined with transfer learning is adopted using the fused images to train the model on an aerial dataset[17].

2.2. IR target detection

The studies mentioned above concentrate on obtaining information from visible images. In recent years, the research on IR technology has been more advanced. Vehicle and pedestrian target detection based on IR images is gradually becoming an attractive method.

A novel detection method for IR point targets based on eigentargets has been proposed[18]. Han *et al.* introduced the subblock-level ratio-difference joint local contrast measure (SRDLCM), which enhances real small targets while suppressing complex backgrounds[19]. A pixel-level classifier was presented for fine-grained detection of pedestrians in night-time CCTV IR images[20]. Eventually, the method maintained more than a 90% F1 score on the test. Nevertheless, the dataset used in this study lacked generality because it was acquired at a specific time and location. Cao *et al.* proposed a one-stage detector named ThermalDet based on the deep neural network[21]. A channel-wise enhancement module was used to assign weights to different channels. Besides, a dual-pass fusion block was added, which combined features from all other levels. This method reached a mean Average Precision (mAP) of 74.60% on the FLIR dataset. This article[22] proposes an anchor-free infrared pedestrian detection algorithm, which introduced a cross-scale feature fusion module and a hierarchical attention mapping module to enhance pedestrian features and suppress background noise. This algorithm integrates the anchor-free concept, which simplifies the network and improves model generalization. A CFRM_3 method[23] was provided in another work to improve the mono-spectral features with the fused multispectral features repeatedly in the network. The experimental results showed that the CFRM_3 led to substantial accuracy improvements. Du *et al.* proposed a weak and occluded vehicle detection method in complex IR environments[24]. A hard negative example mining block was added to the YOLOv4 model to depress the interference caused by complex backgrounds, and the accuracy was increased. Narayanan *et al.* presented a method for IR pedestrian detection using the HOG and the YOLOv3[25]. This work was com-

pared with the technique of using the Support Vector Machine (SVM) classifier. The results showed that the YOLOv3 reached an accuracy of 73%, which was better than that of the SVM algorithm. In the article[26], two multi-scale feature extraction and features fusion mechanisms were designed and added to a target detection model named CMF Net. One of the outstanding advantages of the CMF Net was that the final output backbone feature map contained both low-level visual features and high-level semantic features, facilitating the adaptation of this network to the multi-scale features target.

Although the above works can achieve excellent detection accuracy, there is still much room for improvement in detecting speed. The ability to process the collected road condition information in real time is paramount for the intelligent traffic system (ITS). Zhang *et al*. proposed the CDNet, which implemented real-time crosswalk detection on the Jetson nano device[27]. In another paper[28], a high inference speed framework was introduced to effectively tackle challenges inherent to traffic sign and traffic light detection. Similarly, a delicate balance of accuracy and the real-time performance requirements is considered to implement a pedestrian and vehicle detection task on resource-constrained edge devices, which is also the main focus of our study.

3. METHODS

Our study can be elaborated in two aspects. On the one hand, we find a new way of image processing that is more suitable for IR images than the traditional approaches. On the other hand, we fuse the advantages of the YOLOv4 algorithm and the MobileNetV3 network to build the MobileNetV3-YOLOv4 model. An army of experiments shows that this method performs well in both accuracy and speed.

3.1. IE-CGAN inversion algorithm

IR thermal imaging is a passive IR night vision technology based on the principle that all objects with temperatures above absolute zero (-273.15 °C) radiate IR light noise in an IR image. These can be considered non-periodic random variables that lead to low contrast and resolution of IR images. Therefore, it is indispensable to pre-process the IR images before they are input for training.

The histogram equalization algorithm is a standard method in image pre-processing. The distribution of IR image pixels is extreme, which is different from the RGB images. Consequently, the IR images are usually darker or brighter, making their contrast relatively low. The histogram equalization algorithm can extend the dynamic range of the grayscale fetch, enhance the contrast, and make the image transparent. However, this method also improves the noise in the image, which we want to avoid. The filtering algorithm is also a classical approach widely used to eliminate image noise, but it simultaneously removes some details.

The traditional methods mentioned above have obvious shortcomings, so they are not fully applicable to pre-processing IR images. Currently, deep learning is widely applied in image processing. An attractive network for image enhancement tasks should be equipped with the capabilities to enhance contrast and details while suppressing the background noise. However, existing network architectures for IR image processing, such as residual and encoder-decoder architectures, fail to produce optimal results in network performance and the range of applications. In response to this challenge, Kuang *et al*. devised a novel conditional Generative Adversarial Network (GAN)-based architecture[12]. Their innovation yielded visually captivating results characterized by enhanced contrast and sharper details, addressing the shortcomings of previous approaches. We have further improved their work to obtain a pre-processing method named IE-CGAN inversion algorithm, which is more suitable for IR images.

IE-CGAN contains a generative sub-network for contrast Enhancement and a discriminative sub-network for assistance [Figure 1], where D is a deconvolution layer, the concatenated features are restored to the original resolution using a deconvolution layer followed by a Tanh activation. The generative module first extracts input

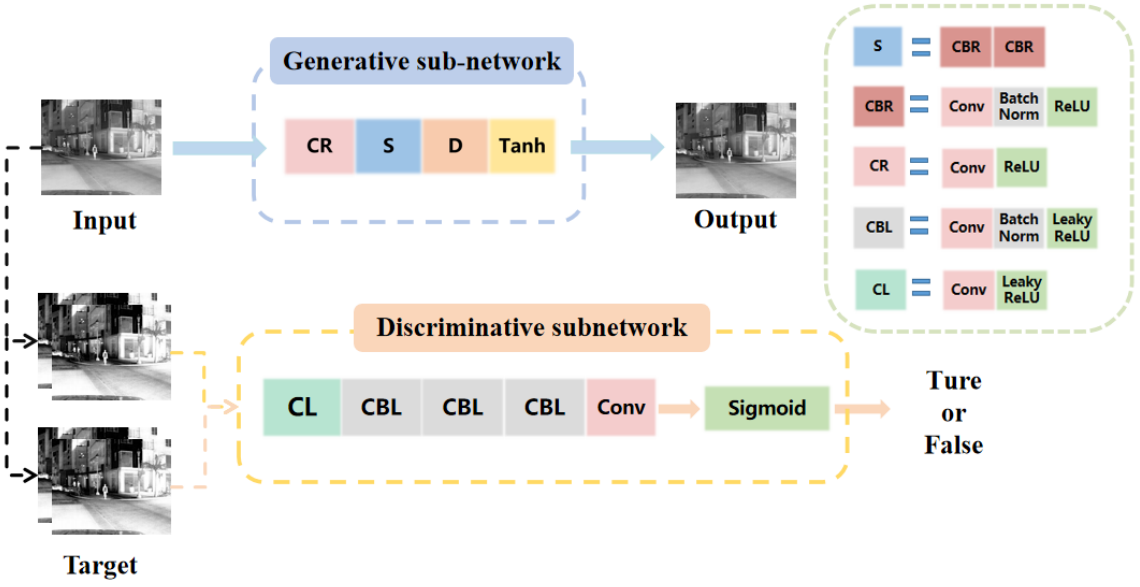

Figure 1. The structure of IE-CGAN algorithm. IE-CGAN: Image Enhancement Conditional Generative Adversarial Network.

image features and then performs a linear combination. Crucially, it integrates a skip connection between the first and third feature maps to preserve fine image details throughout the mapping process. Meanwhile, the discriminative network distinguishes improved results from labeled images, assisting the generative process in creating visually striking outputs. Image transformation into compact feature maps precedes applying a stacked sigmoid function, resulting in normalized probability scores within the range of [0,1]. In addition to producing high-quality pictures, IE-CGAN can be used to any resolution, achieving excellent results in both network performance and the range of applications.

Although the images are enhanced, they must be further processed before being fed into the training network. Generally, RGB images are taken during the daytime, when the image's background is bright and the target color is dark. Nevertheless, since IR images are radiometric, the target radiation is generally more substantial, and the background radiation is weaker, which means that the distribution of light and dark in IR images is the opposite of that in RGB images. The mainstream target detection algorithms are more suitable for RGB images than IR images. Thus, the detection accuracy can be improved if the IR images become closer to the RGB images after pre-processing. In general, images have 256 grayscales. Supposing there is an IR image whose original grayscale is denoted by x_1. After the grayscale inversion process, the grayscale is represented by x_2. Then the relationship between x_1 and x_2 is expressed as follows.

$$x_2 = 255 - x_1 \tag{1}$$

Where x_1 and x_2 are integers, taking values in [0, 255]. After the above processing, the input images are visually closer to the RGB images. We name it the IE-CGAN inversion algorithm.

The comparison in Figure 2 can demonstrate the superiority of our method. The IR image gives higher contrast after the histogram equalization, and the edges of the objects in the picture are more distinct. However, the consequent problem is more noise points in other positions. The filtering algorithms are not practical for processing IR images. They do not make the picture clearer and even blur some image details. The IE-CGAN can significantly improve the contrast of the IR images. Furthermore, the image details and edges are both enhanced. Our method has the advantages of the IE-CGAN and makes the image closer to RGB grayscale

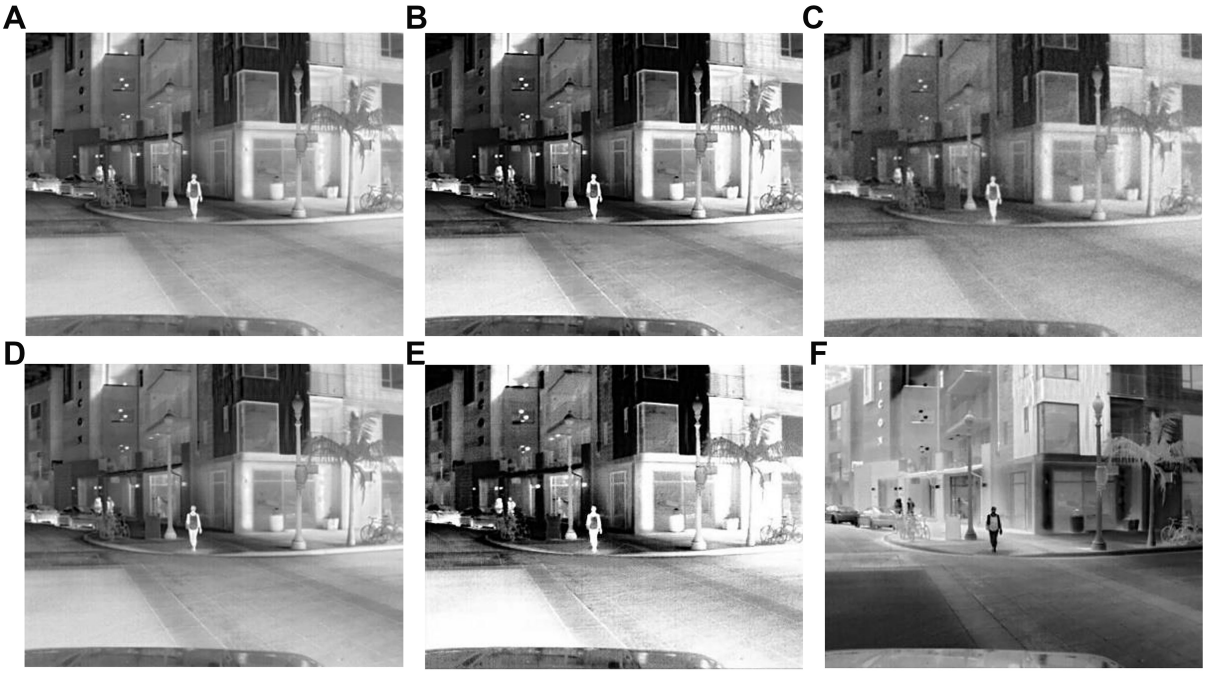

Figure 2. (A) Original IR image; (B) Histogram equalization; (C) Mean filtering; (D) Median filtering; (E) IE-CGAN; (F) IE-CGAN inversion algorithm. IE-CGAN: Image Enhancement Conditional Generative Adversarial Network.

image, which achieves satisfactory results.

3.2. MobileNetV3-YOLOv4 target detection model

The research in target detection can be broadly split into two leading schools: Two-Stage and One-Stage target detection. The Two-Stage target detection algorithm represented by Faster R-CNN[29] has an early origin but generally suffers from large models and slow operation. Redmon *et al*. proposed the pioneering One-Stage algorithm called YOLO[10] to address these drawbacks. Our approach is to improve the YOLOv4 to attain a perfect balance between detection speed and detection accuracy.

The YOLOv4[30] model proposed by Bochkovskiy *et al*. has been upgraded in many aspects compared to the previous version. Figure 3 draws its structure, which can be divided into three components: backbone, neck, and head. YOLOv4 references Cross Stage Partial Networks (CSPNet) and updates the original backbone network Darknet53 into CSPDarknet53. CSPDarknet53 can copy the feature map and send it to the next stage through the dense block, thus separating the feature map of the base layer. This allows the gradient changes to be integrated into the feature map, effectively solving the problem of gradient disappearance. In Yolov4, the Spatial Pyramid Pooling (SPP) structure is a new component added to the neck. It first divides the input feature map into segments. Then, it applies pooling operations with different sizes of pooling kernels in each segment to obtain pooled results for various sizes and receptive fields. Figure 4 demonstrates the pooling of three dimensions as an example. The maximum pooling is performed on the feature map to obtain $1 \times d$, $4 \times d$, and $16 \times d$ features, respectively, representing the feature map's dimension. These pooled results are concatenated into a fixed-length vector as the input of the next layer. As SPP processes the input feature map at multiple scales, it can capture more comprehensive scene information and enhance the adaptability of the object detection network to objects of different scales. Regarding feature fusion, YOLOv4 adopts a Path Aggregation Network (PAN), which complements Feature Pyramid Networks (FPN). The deep layer network responds efficiently to semantic features in convolutional neural networks. Still, it possesses little geometric information, which is unsuitable for target detection. In contrast, the shallow layer network responds quickly to image features but possesses few semantic features, unfit for image classification. FPN is a top-down feature

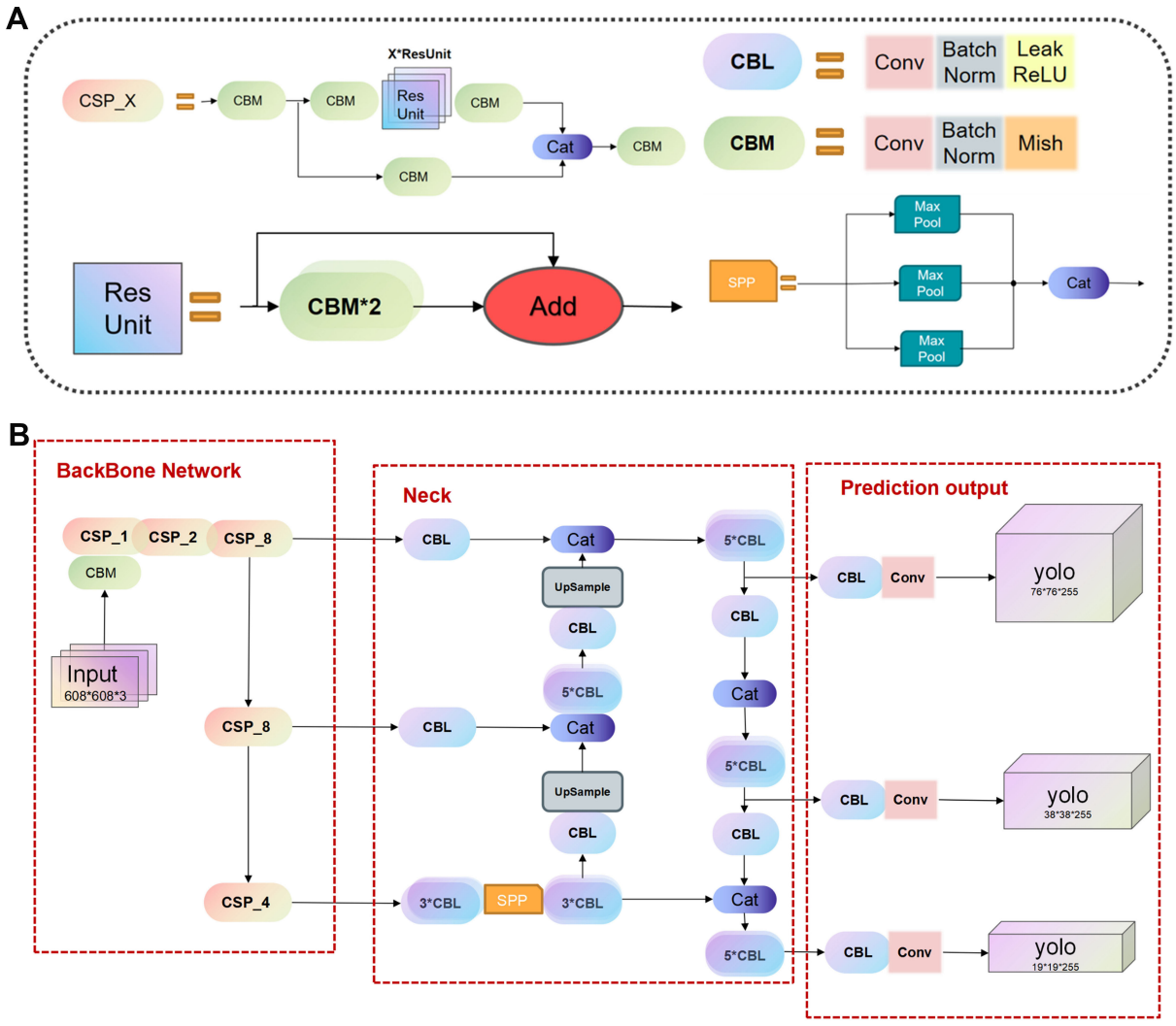

Figure 3. (A) Partial module structure in YOLOv4; (B) The structure of the YOLOv4. YOLOv4: You only look once v4.

pyramid that passes down robust semantic features at the top level through upsampling and then fuses them with lower-level features to obtain a feature map for prediction. Although FPN effectively enhances semantic information, it does not satisfactorily deliver location information. Therefore, YOLOv4 adds a bottom-up feature pyramid to the back of the FPN structure, passing location features from the lower layers to the upper layers through down-sampling and lateral linking. Such an improved feature pyramid has both semantic and location information, which solves the problem as mentioned above.

In addition, YOLOv4 uses Mosaic and Self-Adversarial Training (SAT) for data enhancement. The principle of Mosaic is to combine four training images into one for training, which can enrich the background and enhance target detection in complex backgrounds. SAT is a novel data enhancement technique that is divided into two phases. In the first stage, the neural network performs adversarial training by changing the original image without changing the network weights. In the second stage, the neural network is trained to perform standard target detection on the modified image. After the above data enhancement process, the robustness of the model is improved.

Although YOLOv4 has a potent capability, we still want to improve its performance of detection speed and

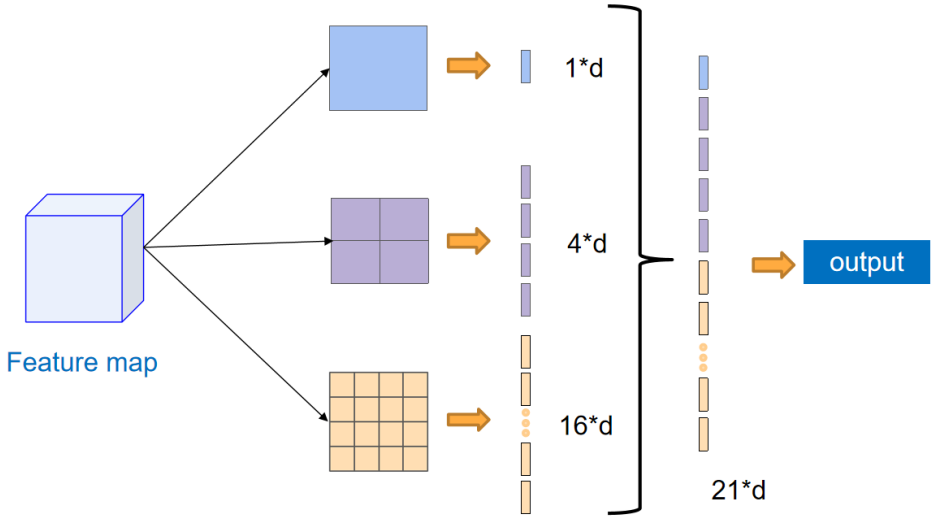

Figure 4. The principle of SPP. SPP: Spatial Pyramid Pooling.

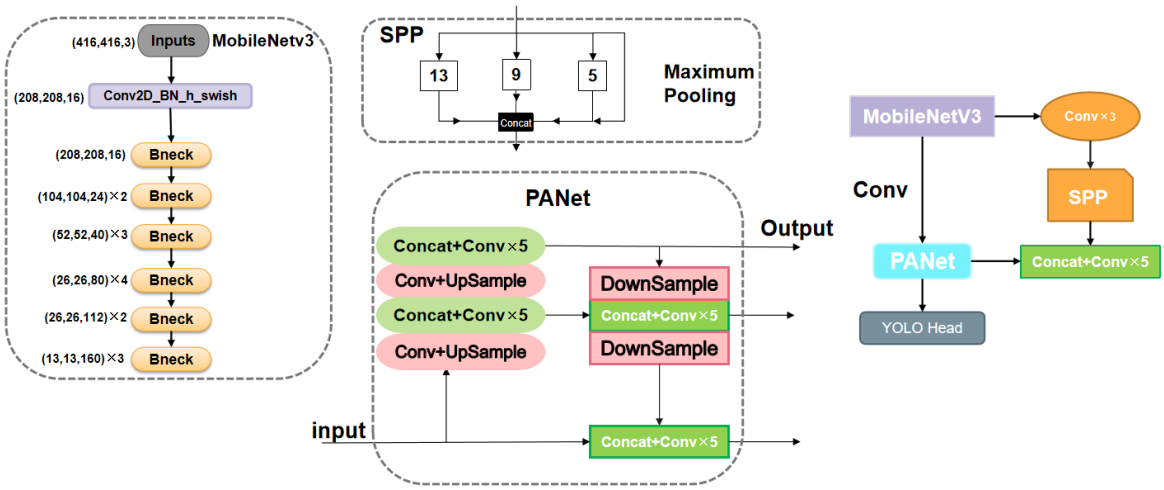

Figure 5. The structure of the MobileNetV3-YOLOv4 model.

lighten its model size to enable deployment on resource-constrained edge devices. Seeking a lightweight network to substitute ddCSPDarknet53 as the feature extraction network of the YOLOv4 will be a viable option.

The MobileNetV3 is a lightweight convolutional neural network presented by the Google team [31], widely deployed on cell phones and smart bracelets. MobileNetV3 dramatically reduces parameters and increases speed by sacrificing only a small amount of accuracy compared with traditional large-scale convolutional neural networks such as AlexNet [32] and ResNet. In the tail structure of the MobileNetV3, the average pooling is applied to cut the feature map of size 7×7 to 1×1. After that, the dimensionality of the feature map is increased by a 1×1 convolution. The whole process reduces the computational by a factor of forty-nine. Because some convolutions in the head structure with the size of 3×3 and 1×1 have little impact on the accuracy, MobileNetV3 removes them directly to improve the speed further. Additionally, MobileNetV3 cuts the convolutional core channels from 32 to 16, which is also an effective solution to make the network faster. To avoid a substantial decrease in accuracy, the Squeeze-and-Excitation Block (SE Block) is added to the core architecture of MobileNetV3. The SE Block can determine the importance of each feature channel based on their dependency relationship. The network can selectively enhance the useful features while suppressing the less useful ones

through this mechanism.

Since the MobileNetV3 meets our practical performance requirements, we choose to replace the CSPDarknet53 as the backbone to obtain the MobileNetV3-YOLOv4 model. The model structure is shown in Figure 5. MobileNetV3 reduces model size and computational requirements while maintaining high performance. MobileNetV3's transformer-based architecture provides superior feature extraction capabilities to traditional CNNs, making it particularly effective for IR imaging where challenges such as low contrast and noise interference are prevalent. Moreover, MobileNetV3's efficient integration of local, global, and input features enhances its ability to identify objects across different scales accurately. This results in improved performance for complex IR imaging tasks. Empirical results have confirmed that incorporating MobileNetV3 into YOLOv4 maintains high accuracy and reduces computational load and model size. Although MobileNetV3 may not perform as well as expected in YOLOv3, its lightweight and efficient feature learning capabilities have significantly improved IR target detection tasks in YOLOv4. This improvement is not only theoretically reasonable but its effectiveness in practical applications has also been verified through experiments. The MobileNetV3 network first extracts the features of the input image. Afterward, the SPP module performs maximum pooling on the front layer features. It connects the processed results to form a new feature layer, which increases the depth of the network and preserves the front layer features, and obtains more local feature information. The PAN block upsamples and downsamples the features extracted by the MobileNetV3 to improve the information extraction capability of the FPN block. The feature network and feature layers are fused by adaptive pooling of different layers, and the fused results are passed into YOLO Head for regression and classification. YOLO Head divides the input images into networks of corresponding sizes, and finally, the classification results and confidence levels of the objects are obtained by the predefined prior frame determination.

4. EXPERIMENTS

The experiments were conducted using a computer that had Ubuntu 18.04 pre-installed. The CPU was Intel (R) Core (TM) i59300H, 2.40GHz. The GPU was NVIDIA GeForce GTX 1650, with 64 GB of memory. To test the performance of the IR image object detection model proposed in this article, we utilized the common FLIR IR dataset and the KAIST IR pedestrian dataset. Firstly, we compared the latest detection algorithms and models proposed on various datasets regarding detection accuracy, speed, and model size. Secondly, ablation experiments were conducted on the enhanced model to assess the effectiveness of various improvement methods.

4.1. Datasets

4.1.1 The FLIR IR datasets

The dataset was an IR dataset open-sourced by FLIR in July 2018, applied for many IR image target detection training tasks. This FLIR IR dataset provided two types of images: thermal imaging images with annotations and corresponding RGB images without annotations. The FLIR dataset contained 14,452 IR images, of which 10,228 were from multiple short videos, and 4,224 were from a long video of 144 s. All of the images were taken from actual streets and highways. Figure 6 shows the FLIR dataset.

4.1.2 The KAIST datasets

The KAIST IR pedestrian dataset is a widely used benchmark for evaluating algorithms for detecting objects in IR images. The dataset comprises 95,328 pairs of images, each with a resolution of 640×512. The dataset offers meticulous manual annotations and well-matched visible and IR image pairs. It provides comprehensive coverage, spanning diverse traffic scenarios such as campuses, streets, and rural areas. Annotations differentiate between "person" for individual pedestrians and "people" for groups where individuals are more challenging to discern. We extracted 15,684 consecutive images from the raw data to streamline model training and performance evaluation. Experimental outcomes validate the dataset's efficacy in achieving high detection

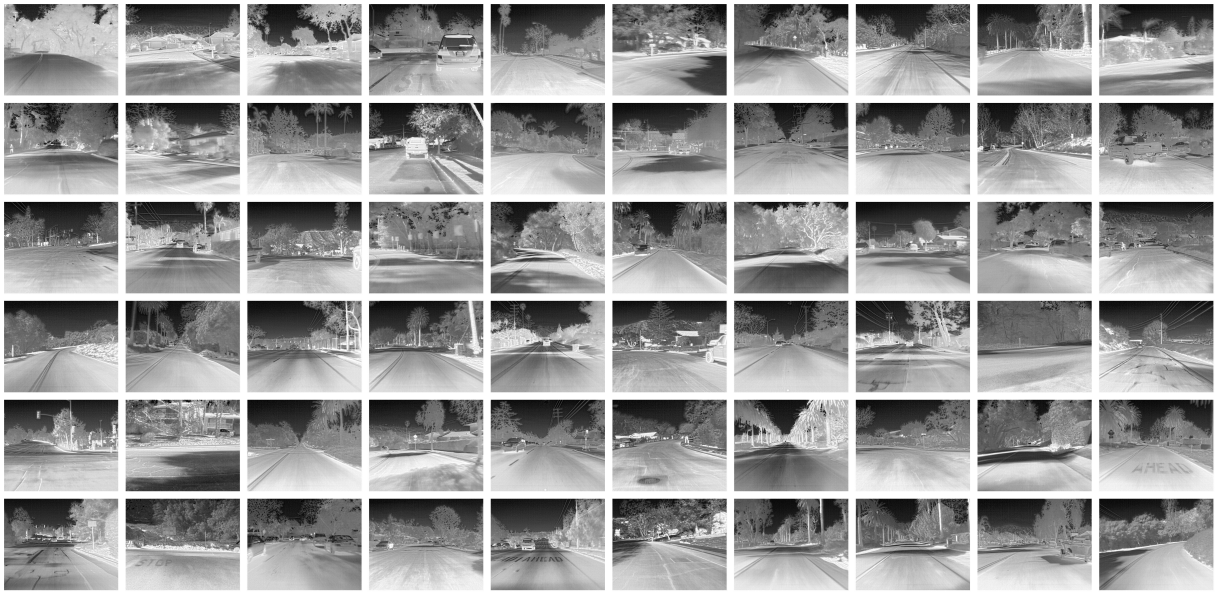

Figure 6. The FLIR dataset.

accuracy.

4.2. Parameter configuration and evaluation indicator

The experimental process drew on the idea of transfer learning. The network's initial weights at the beginning of the training were not randomly set but obtained from the YOLOv4 model after training on the ImageNet and MSCOCO datasets, reducing the time spent on training.

The most common Stochastic Gradient Descent (SGD) algorithm was used for the network optimizer's optimization algorithm, together with the Momentum algorithm that could be ported to oscillate, with the momentum taking the value of 0.9. At the start of the training, the learning rate was set to 1e-3, and the training process was set to run for 120 epochs. As the epoch number increased, the learning rate gradually decreased to 1e-5. The size of the input image was set to 416×416.

In previous research, mAP was often used to measure the target detection capability, reflecting a certain method's accuracy. Before calculating mAP, we need to get the formulas of Precision and Recall:

$$Precision = \frac{TP}{TP + FR} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

TP refers to cases where the prediction and label are both positive, while FP refers to cases where the prediction is optimistic but negative. FN refers to cases where the prediction is pessimistic while the label is positive. By utilizing a Precision-Recall (PR) curve, AP and mAP can be calculated based on the corresponding precision and recall values for each point of recall.

$$AP = \Sigma_{i=1}^{n-1} (R_{i+1} - R_i) P_{inter} (R_i + 1) \tag{4}$$

$$mAP = \frac{1}{k} \Sigma_{i=1}^{k} AP_i \tag{5}$$

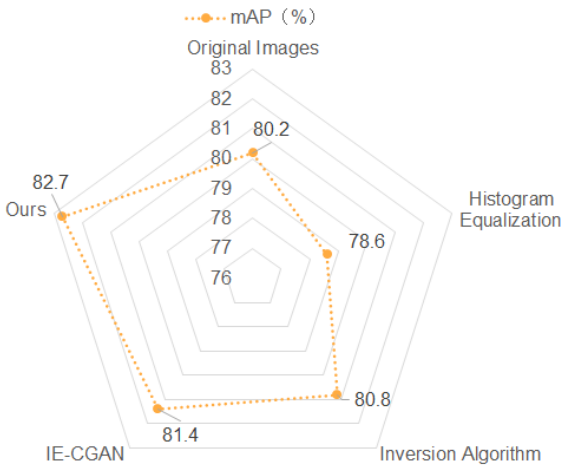

Figure 7. Detection results of different pre-processing methods.

where R_1 to R_n are the Recall values corresponding to the first interpolation of precision interpolation segment. In simple terms, AP is the area under the curve in the PR plot, and mAP is the average of all categories of AP. The experiment's Intersection over Union (IoU) threshold was set to 0.5. Besides, we used the frame per second (FPS) to evaluate the real-time detection ability of the system. However, though the state-of-the-art vehicle and pedestrian detection methods performed well in mAP or FPS, their huge model sizes made them unsuitable for deployment on edge devices. Considering that our ultimate goal is to deploy the model on resource-constrained edge devices, we also need to focus on the model size as a part of the evaluation. Therefore, we conducted a comprehensive comparison of mAP, FPS, and model sizes to highlight the advantages of our study.

4.3. Experimental results and analysis

We applied the centralized IR image pre-processing method mentioned in the third section to our improved model [Figure 7]. Compared with the original image experimental results, our improved IE-CGAN pre-processing method, mAP, has increased by 2.5%, the best detection effect among the experimental techniques. By incorporating this pre-processing method, our improved model enhances the detection accuracy and demonstrates a more robust performance across different IR target scenarios.

We selected a more lightweight model for improvement due to the necessity of hardware deployment. While other advanced models offer higher accuracy, they often feature larger architectures less compatible with our hardware constraints. Consequently, we chose a YOLO series model, which balances moderate accuracy and manageable model size, making it one of our optimal solutions. We opted for YOLOv4 as the baseline to evaluate the effectiveness of our proposed model, given its superior accuracy and enhanced detection performance, as indicated in Table 1. Furthermore, we comprehensively compared baseline models and other leading models on the FLIR IR vehicle detection dataset to ensure a diverse range of experimental evaluations. Detailed comparative experimental data can be found in Table 1. In contrast to the baseline detector YOLOv4, we introduced a new object detector in the experiment and evaluated several advanced two-stage detectors, including the classical Faster R-CNN detection model. The experimental findings reveal that, in practical scenarios, these secondary detectors perform suboptimally compared to our model. Our model has fewer parameters, lower computational costs, and superior performance. It can be deployed on hardware devices to achieve real-time object detection. Our model significantly improved by nearly 25% over YOLOv3 compared to other primary detectors. Moreover, we achieved a 2% improvement on the widely-used YOLOv4 model. The detec-

Figure 8. Some examples of the detection result on the FLIR dataset. The first column is the original image, and the second column is the result of MobileNetV3-YOLOv4.

tor proposed in this experiment outperforms the previously mentioned detection models regarding detection accuracy and computational resource efficiency. We also compared our model with YOLOv5, YOLOv8s, and YOLOv3 MobileNetv3. In the IR target detection task, our model significantly outperforms previous models and aligns more closely with the requirements of our real-time monitoring task. YOLO-IR has demonstrated outstanding performance on the FILR dataset. Our model achieved higher accuracy in this task with fewer parameters, improving by 5%, despite some performance degradation. Additionally, we compared Source Model Guidance based on YOLOv3 (SMG-Y) and PMBW (a Paced MultiStage BlockWise approach to Object Detection in Thermal Images), both based on visual converters. It can be seen that our method has an absolute advantage in detection speed and high accuracy. Meanwhile, our model size is only 110MB, which performed better than many methods. This balanced improvement in the three evaluations makes the proposed method suitable for deployment on resource-constrained edge devices. Examples of the detection results are displayed in Figure 8.

To demonstrate the excellent performance of this model, it was compared not only with many other models on the FLIR dataset but also on the KAIST dataset, and competitive results were achieved. Table 2 presents the comparison results of our model with other models on the KAIST dataset. We compared our model with recent excellent single-stage detectors and some lightweight detectors. The results indicate that our model is the smallest and superior to other detection models. Regarding accuracy, our mAP outperforms other detection models. Our model demonstrates significant performance advantages compared to other models. Compared to YOLOv3, YOLOv4, and other benchmark models, our model outperforms them in mAP. Compared to YOLOv4, our model shows a slight improvement in mAP, ranging from 81.0% to 86.8%, along with enhanced processing speed, increasing from 42 to 64.2 frames per second. Compared with pixel-wise contextual attention network (PiCA-Net), Multimodal Feature Embedding (MuFEm) + Spatio-Contextual Feature Aggregation (ScoFA), and multispectral fusion and double-stream detectors with Yolo-based information (MFDs-YOLO), our model demonstrates notable enhancements in detection accuracy. Additionally, although our model experiences a slight decrease in mAP compared to YOLO-ACN, there are significant improvements in processing speed and model size. Overall, our model achieves substantial accuracy, speed, and size advancements, making it more practical and competitive.

Table 1. Performance comparison (%) with the state-of-the-art methods on the FLIR dataset

Methods	mAP (%)	FPS (frame/s)	Model size (MB)
Faster R-CNN	84.6	6.1	577.0
VGG16	84.2	5.5	526.7
ResNet50	83.8	7.6	446.2
YOLOv3	58.2	38.5	246.4
YOLOv4	81.2	27.0	256.3
YOLOv5	73.6	39.6	191.2
TOLOv8s	74.2	158.3	/
RefineDet [33]	72.9	/	/
ThermalDet [34]	74.6	/	/
SMG-C [35]	75.6	107.0	/
SMG-Y [35]	77.0	40.0	/
YOLO-IR [36]	78.6	151.1	/
PMBW [37]	77.3	/	36.0
YOLOv3-MobileNetV3 [38]	60.59	14.40	139.60
DS-Net [15]	71.9	32.8	25.6
ours	82.7	55.9	110.0

mAP: Mean Average Precision; FPS: frame per second; R-CNN: the Region with CNN features; YOLO: you only look once; SMG-Y: Source Model Guidance based on YOLOv3; PMBW: Paced MultiStage BlockWise.

Table 2. Performance comparison (%) with the state-of-the-art methods on the KAIST dataset

Methods	mAP (%)	FPS (frame/s)	Model size (MB)
YOLOv3	79.6	36	246.4
YOLOv4	81.0	42	256.3
PiCA-Net [39]	65.8	/	/
MuFEm + ScoFA [40]	78.0	/	/
MFDs-YOLO [41]	80.3	/	/
YOLO-ACN [42]	82.3	/	177.6
ours	86.8	64.2	110.0

mAP: Mean Average Precision; FPS: frame per second; YOLO: you only look once.

Table 3. Ablation study of detection precision on the FLIR dataset

Methods	IE-CGAN	CSPDarknet53	MobileNetV3	mAP (%)
1		✓		80.2
2			✓	80.7
3	✓	✓		81.2
4	✓		✓	82.7

IE-CGAN: Image Enhancement Conditional Generative Adversarial Network; mAP: Mean Average Precision.

To intuitively demonstrate the influence of different methods on network performance, we conducted ablation experiments on the FLIR dataset using the YOLOv4 network. Specifically, we maintained the structure of YOLOv4 unchanged. Initially, we replaced the original backbone with MobileNetv3 and made further enhancements. Then, we implemented new data processing methods. We trained and tested the network on various datasets to assess the influence of these methods on network performance. As shown in Table 3, applying our data processing method IE-CGAN to the baseline model can increase the detection results mAP by 1.0%. We replaced the backbone network CSPDarknet53 of YOLOv4 with MobileNetV3, which can increase the detection results mAP by 1.5% and significantly reduce the model size. We have selected several commonly used algorithms as references to test the performance of our method, and the experimental results are shown in the Table. Our model excels in terms of detection speed and model size, achieving a frame rate of 55.9 per second and a compact model size of only 110.0 MB. The detection accuracy is good and can meet the recognition requirements. These results indicate that the model can detect onboard equipment in real time and perform lightweight tasks.

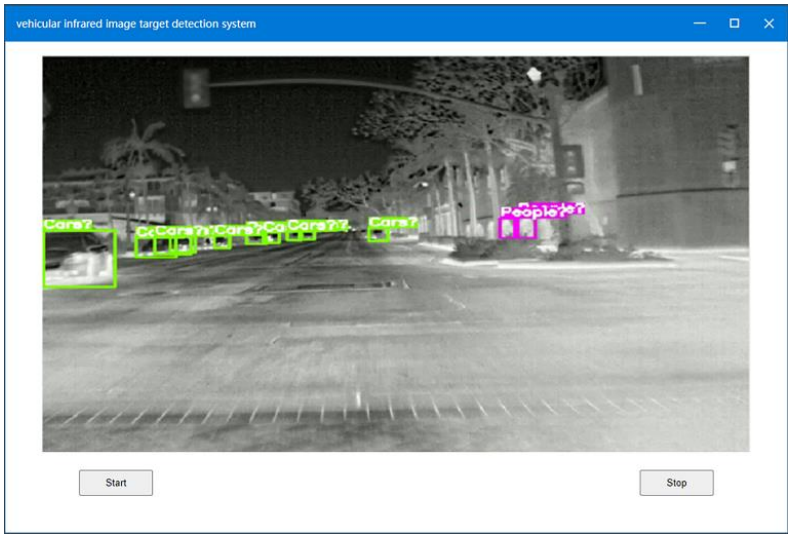

Figure 9. System operation interface.

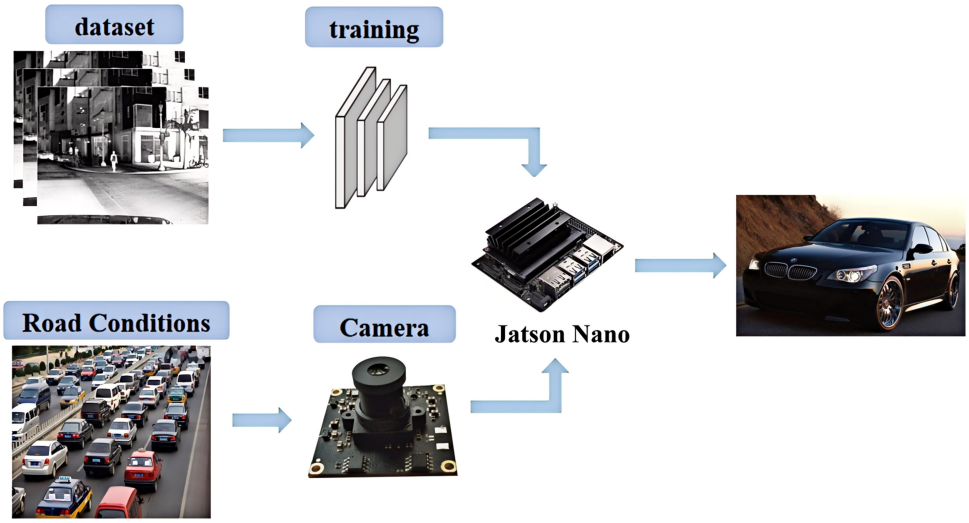

Figure 10. The workflow of the system.

5. DEPLOYMENT

The development of a deep learning application includes four steps: task modeling, data acquisition, model training, and model deployment. As the last step of implementing an application, model deployment is essential. With the development of artificial intelligence (AI), a series of embedded development boards for the AI field has been launched. Compared with Jetson Nano, Raspberry Pi 4 B's hardware condition is insufficient to support it in achieving the desired detection effect, while Jetson-TX2 and Jetson-AGXXavier are too expensive. Considering the balance between actual demand and the cost, we finally deployed the model on Jetson Nano.

The Jetson Nano is pre-installed with the Ubuntu 18.04 LTS system and has a 128-core Maxwell GPU. It can provide 472 GFLOP computing performance and 4GB of LPDDR4 memory. The outstanding hardware conditions give it a significant advantage in AI technology implementation. As an edge device, the Jetson Nano has the benefits of compact size, GPU-accelerated inference, and relatively low price, making it market competitive. In addition, the TensorRT toolkit was applied in the model inference phase, which provided high

throughput data for the model. Thus, underutilized GPU resources were solved, and the inference of the network framework was optimized to improve the inference efficiency. We build a visual system interface using the PyQT5 library [Figure 9].

Figure 10 demonstrates the workflow of the whole system. Firstly, the model is trained on the computer. Afterward, the trained model is solidified and deployed on Jetson Nano. The camera converts the acquired road conditions into images and inputs them into Jetson Nano. After the vehicle and pedestrian detection, Jetson Nano gives instructions such as braking and turning to the car.

6.CONCLUSION

This paper introduces an in-vehicle IR target detection method built on an improved YOLOv4 model, which integrates the IE-CGAN inversion algorithm to pre-process IR images. This integration enhances both image quality and detection performance. An IE-CGAN inversion algorithm is used instead of conventional methods to pre-process the IR images. Additionally, considering that the algorithm requires being deployed on the edge device, this study concentrates on improving the system's processing speed for IR images, so the backbone network of the YOLOv4 model is replaced from CSPDarknet53 to MobileNetV3, improving processing speed and efficiency. However, the dataset we used has limited diversity in image types and needs more generalizability in background models. Moreover, the model's ability to generalize requires further improvement. Our future work will focus on expanding the dataset to include a broader range of IR images, enhancing the system's robustness and generalizability across different scenarios. Moving forward, we remain committed to addressing the current limitations and enhancing the system's performance and generalizability in future work.

DECLARATIONS

Authors' contributions

Made substantial contributions to the conception and design of the study and performed data analysis and interpretation: Zhuang T, Liang X

Performed data acquisition and provided administrative, technical, and material support: Xue B, Tang X

Availability of data and materials

The FILR datasets used in this article can be found at https://www.flir.com/oem/adas/adas-dataset-form/.

Financial support and sponsorship

This work was supported by the National Natural Science Foundation of China (Grant No.62001173), the Project of Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation ("Climbing Program" Special Funds) (Grant Nos. pdjh2022a0131 and pdjh2023b0141).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Guo Y, Wang X, Yuan Q, Liu S, Liu S. Transition characteristics of driver's intentions triggered by emotional evolution in two-lane urban roads. *IET Intell Trans Syst* 2020;14:1788-98. DOI
2. Rokonuzzaman M, Mohajer N, Nahavandi S, Mohamed S. Review and performance evaluation of path tracking controllers of autonomous vehicles. *IET Intell Trans Syst* 2021;15:646-70. DOI
3. Sun Z, Bebis G, Miller R. Monocular precrash vehicle detection: features and classifiers. *IEEE Trans Image Process* 2006;15:2019-34. DOI
4. Halmaoui H, Joulan K, Hautière N, Cord A, Brémond R. Quantitative model of the driver's reaction time during daytime fog - application to a head up display-based advanced driver assistance system. *IET Intell Trans Syst* 2015;9:375-81. DOI
5. Wang Y, Xie W, Liu H. Low-light image enhancement based on deep learning: a survey. *Opt Eng* 2022;61:040901. DOI
6. Altaf MA, Ahn J, Khan D, Kim MY. Usage of IR sensors in the HVAC systems, vehicle and manufacturing industries: a review. *IEEE Sens J* 2022;22:9164-76. DOI
7. Chen B, Wang W, Qin Q. Robust multi-stage approach for the detection of moving target from infrared imagery. *Opt Eng* 2012;51:067006. DOI
8. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05); 2005 Jun 20-25; San Diego, CA, USA. IEEE; 2005. pp. 886-93. DOI
9. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23-28; Columbus, OH, USA. IEEE; 2014. pp. 580-87. DOI
10. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, NV, USA. IEEE; 2016. pp. 779-88. DOI
11. Law H, Deng J. Cornernet: Detecting objects as paired keypoints. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018). Cham: Springer; 2018. pp. 765-81. DOI
12. Kuang X, Sui X, Liu Y, Chen Q, Gu G. Single infrared image enhancement using a deep convolutional neural network. *Neurocomputing* 2019;332:119-28. DOI
13. Li J, Liang X, Shen S, Xu T, Feng J, Yan S. Scale-aware fast R-CNN for pedestrian detection. *IEEE Trans Multimed* 2017;20:985-96. DOI
14. Fan R, Wang H, Wang Y, Liu M, Pitas I. Graph attention layer evolves semantic segmentation for road pothole detection: a benchmark and algorithms. *IEEE Trans Image Process* 2021;30:8144-54. DOI
15. Feng H, Wang X, Feng M, Bu C. A lane segmentation and traffic object detection multi-task neural network for AR-HUD. In: 2021 China Automation Congress (CAC); 2021 Oct 22-24; Beijing, China. IEEE; 2021. pp. 3062-67. DOI
16. Li Y, Wang H, Dang LM, Nguyen TN, Han D, Moon H. A deep learning-based hybrid framework for object detection and recognition in autonomous driving. *IEEE Access* 2020;8:194228-39. DOI
17. Wang C, Luo D, Liu Y, Xu B, Zhou Y. Near-surface pedestrian detection method based on deep learning for UAVs in low illumination environments. *Opt Eng* 2022;61:023103. DOI
18. Liu R, Liu E, Yang J, Zhang T, Cao Y. Point target detection of infrared images with eigentargets. *Opt Eng* 2007;46:110502. DOI
19. Han J, Yu Y, Liang K, Zhang H. Infrared small-target detection under complex background based on subblock-level ratio-difference joint local contrast measure. *Opt Eng* 2018;57:103105. DOI
20. Park J, Chen J, Cho YK, Kang DY, Son BJ. CNN-based person detection using infrared images for night-time intrusion warning systems. *Sensors* 2020;20:34. DOI
21. Cao Y, Zhou T, Zhu X, Su Y. Every feature counts: an improved one-stage detector in thermal imagery. In: 2019 IEEE 5th International Conference on Computer and Communications (ICCC); 2019 Dec 6-9; Chengdu, China. IEEE; 2019. pp. 1965-9. DOI
22. Hao S, Gao S, Ma X, He T. Anchor-free infrared pedestrian detection based on cross-scale feature fusion and hierarchical attention mechanism. *Infrared Phys Technol* 2023;131:104660. DOI
23. Zhang H, Fromont E, Lefevre S, Avignon B. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: 2020 IEEE International Conference on Image Processing (ICIP); 2020 Oct 25-28; Abu Dhabi, United Arab Emirates. IEEE; 2020. pp. 276-80. DOI
24. Du S, Zhang P, Zhang B, Xu H. Weak and occluded vehicle detection in complex infrared environment based on improved YOLOv4. *IEEE Access* 2021;9:25671-80. DOI
25. Narayanan A, Kumar RD, RoselinKiruba R, Sharmila TS. 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); 2021 Mar 25-27; Chennai, India. IEEE; 2021. pp. 431-34. DOI
26. Liao Z, Zhao Y, Huang X, Wu J. CMF net: detecting objects in infrared traffic image with combination of multiscale features. In: 2021 IEEE Global Communications Conference (GLOBECOM); 2021 Dec 7-11; Madrid, Spain. IEEE; 2021. pp. 1-6. DOI
27. Zhang ZD, Tan ML, Lan ZC, Liu HC, Pei L, Yu WX. CDNet: a real-time and robust crosswalk detection network on Jetson nano based on YOLOv5. *Neural Comput Appl* 2022;34:10719-30. DOI
28. Jayasinghe O, Hemachandra S, Anhettigama D, et al. Towards real-time traffic sign and traffic light detection on embedded systems. In: 2022 IEEE Intelligent Vehicles Symposium (IV); 2022 Jun 4-9; Aachen, Germany. IEEE; 2022. pp. 723-28. DOI
29. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach* 2015;39:1137-49. DOI
30. Bochkovskiy A, Wang CY, Liao HYM. Yolov4: optimal speed and accuracy of object detection. arXiv. [Preprint.] Apr 23, 2020 [accessed 2024 Sep 14]. Available from: https://arxiv.org/abs/2004.10934.

31. Howard A, Sandler M, Chen B, et al. Searching for MobileNetv3. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27 - Nov 2; Seoul, Korea (South). IEEE; 2019. pp. 1314-24. DOI
32. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM* 2017;60:84-90. DOI
33. Devaguptapu C, Akolekar N, Sharma MM, Balasubramanian VN. Borrow from anywhere: pseudo multi-modal object detection in thermal imagery. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019 Jun 16-17; Long Beach, CA, USA. IEEE; 2019. pp. 1029-38. DOI
34. Cao Y, Zhou T, Zhu X, Su Y. Every feature counts: an improved one-stage detector in thermal imagery. In: 2019 IEEE 5th International Conference on Computer and Communications (ICCC); 2019 Dec 6-9; Chengdu, China. IEEE; 2019. pp. 1965-69. DOI
35. Chen R, Liu S, Mu J, Miao Z, Li F. Borrow from source models: efficient infrared object detection with limited examples. *Appl Sci* 2022;12:1896. DOI
36. Zha C, Luo S, Xu X. Infrared multi-target detection and tracking in dense urban traffic scenes. *IET Image Process* 2024;18:1613-28. DOI
37. Kera SB, Tadepalli A, Ranjani JJ. A paced multi-stage block-wise approach for object detection in thermal images. *Vis Comput* 2023;39:2347-63. DOI
38. Dong J, Ota K, Dong M. Real-time survivor detection in UAV thermal imagery based on deep learning. In: 2020 16th International Conference on Mobility, Sensing and Networking (MSN); 2020 Dec 17-19; Tokyo, Japan. IEEE; 2020. pp. 352-59. DOI
39. Ghose D, Desai SM, Bhattacharya S, Chakraborty D, Fiterau M, Rahman T. Pedestrian detection in thermal images using saliency maps. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019 Jun 16-17; Long Beach, CA, USA. IEEE; 2019. pp. 988-97. DOI
40. Dasgupta K, Das A, Das S, Bhattacharya U, Yogamani S. Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. *IEEE Trans Intell Trans Syst* 2022;23:15940-50. DOI
41. Hsia CH, Peng HC, Chan HT. All-weather pedestrian detection based on double-stream multispectral network. *Electronics* 2023;12:2312. DOI
42. Li Y, Li S, Du H, Chen L, Zhang D, Li Y. YOLO-ACN: focusing on small target and occluded object detection. *IEEE Access* 2020;8:227288-303. DOI