IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 8, 2021, accepted October 13, 2021, date of publication October 15, 2021, date of current version October 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3120870

YOLO-FIRI: Improved YOLOv5 for Infrared Image Object Detection

SHASHA LI[1], YONGJUN LI[1], YAO LI[1], MENGJUN LI[1], AND XIAORONG XU[2]
[1]School of Physics and Electronics, Henan University, Kaifeng 475004, China
[2]School of Computer and Electrical Engineering, Hunan University of Arts and Science, Changde 415000, China

Corresponding author: Yongjun Li (lyj@henu.edu.cn)

This work was supported in part by the Key Research and Development and Promotion Projects in Henan Province under Grant 212102210151, and in part by the National Natural Science Foundation of China under Grant U1704130.

ABSTRACT To solve object detection issues in infrared images, such as a low recognition rate and a high false alarm rate caused by long distances, weak energy, and low resolution, we propose a region-free object detector named YOLO-FIR for infrared (IR) images with YOLOv5 core by compressing channels, optimizing parameters, etc. An improved infrared image object detection network, YOLO-FIRI, is further developed. Specifically, while designing the feature extraction network, the cross-stage-partial-connections (CSP) module in the shallow layer is expanded and iterated to maximize the use of shallow features. In addition, an improved attention module is introduced in residual blocks to focus on objects and suppress background. Moreover, multiscale detection is added to improve small object detection accuracy. Experimental results on the KAIST and FLIR datasets show that YOLO-FIRI demonstrates a qualitative improvement compared with the state-of-the-art detectors. Compared with YOLOv4, the mean average precision (mAP50) of YOLO-FIRI is increased by 21% on the KAIST dataset, the speed is reduced by 62%, the parameters are decreased by 89%, the weight size is reduced by more than 94%, and the computational costs are reduced by 84%. Compared with YOLO-FIR, YOLO-FIRI has an approximately 5% to 20% improvement in AP, AR (average recall), mAP50, F1, and mAP50:75. Furthermore, due to the shortcomings of high noise and weak features, image fusion can be applied to image preprocessing as a data enhancement method by fusing visible and infrared images based on a convolutional neural network.

INDEX TERMS Attention mechanism, infrared image, image fusion, object detection, YOLOv5.

I. INTRODUCTION

Object detection in infrared images has received extensive attention within the field of computer vision due to its important research value for applications. It also occupies an irreplaceable position in many fields, such as in diagnosing diseased cells [1], video surveillance [2], drone cruise [3], infrared warning [4], infrared night vision [5], infrared guidance [6], and other civilian and military fields. Although object detection models have achieved promising results in various tasks, it is still a challenging task in infrared images, as most models only focus on visible images.

Obtained through thermal radiation, infrared images have outstanding characteristics, such as object detection from long distances, high concealment, and availability both in the daytime and nighttime. With the expansion of distance imaging, the ever-growing demands for intelligent object detection in infrared images have become more urgent. However, constructing models for infrared images to achieve the desired results has been restricted by the longer imaging wavelength, larger noise, poorer spatial resolution, and more sensitivity to temperature changes in the environment compared with those characteristics of visible images. In recent years, several studies have investigated the possibility of object detection in infrared images using methods, such as spatial filtering [7], frequency domain filtering [8], and sparse representation [9]. However, these traditional object detection methods in infrared images are restricted by a single application scenario. They also have a slow recognition speed and a weak generalizability, making them difficult to use for fully extracting important features when applied to multi-scene and real-time detection applications.

Convolutional neural networks (CNN) models have the ability to learn the deep features of the input. The high-level

The associate editor coordinating the review of this manuscript and approving it for publication was Zhongyi Guo.

data features from the original pixels of the training data can be learned to obtain a better feature expression capability for complex context information. Some networks have greatly improved their accuracy and generalize ability [10], and have solved the problems of image classification [11], image segmentation [12], superresolution [13], etc. Examples include region-based two-stage object detection algorithms, such as R-CNN [14], Fast R-CNN [15], and Faster R-CNN [16], as well as region-free one-stage object detection algorithms in the SSD (Single Shot Multi-box Detector) series [17], [18] and the YOLO series [19]–[22]. In comparison to visible images, the lower signal-to-noise ratio of infrared images makes the objects easier to submerge and interfere; the complex background of infrared images makes the object areas dark and uneven; and the distance between objects and infrared sensors is relatively long, making the objects occupy small areas of the whole image. This shows that the CNN models are robust when applied to visible images. Despite the effectiveness of these studies, using a convolutional neural network to detect weak and small objects in infrared images has become a difficult and hot research topic.

Based on the characteristics of infrared objects, we propose an object detection algorithm named YOLO-FIR for infrared images based on the region-free detector YOLOv5 [23]. As a state-of-the-art detector, YOLOv5 has the advantages of fast convergence, high precision, and strong customization. It also has strong real-time processing capabilities and low hardware computing requirements, meaning that it can be easily transplanted to mobile devices. These advantages are very helpful for ensuring the detection accuracy of the object in infrared images. Then, we make further improvements and propose a novel detection model, YOLO-FIRI, for small and weak objects in infrared images. The model proves to be reliable and efficient for object detection in infrared images. In the design of the YOLO-FIRI network, the CSP [24] module in the backbone network is extended to focus on shallow information and extract features to the maximum, while the feature extraction module is iterated to extract the detailed information and the deep features more thoroughly. Meanwhile, the SK (Select Kernel) [25] attention module is introduced and improved in residual blocks, and the features are re-weighted and fused from the channel dimension. In the detection stage, to better detect small and weak objects, multiscale feature detection is improved. Four-scale feature maps are used to detect multiscale objects, especially to enhance the detection of small objects. Additionally, to evaluate the contribution of each part of the network model designed in this paper, we conduct an ablation experiment on the KAIST [26] infrared pedestrian dataset. Finally, in the experimental analysis, the image fusion method is adopted to realize data enhancement by using the Densefuse [27] network to fuse visible and infrared images. We also use another dataset, FLIR [28], to further evaluate the performance. Experimental results demonstrate that the presented model improves the detection accuracy and ensures real-time detection speed in infrared images.

The main contributions of the research can be summarized as follows:

1. Based on studying the unique features of infrared images, this paper proposes the YOLO-FIR method for infrared objects detection with the core of YOLOv5 by analyzing the network structure, compressing channels, optimizing parameters, etc. Further improvements are made to design a novel network, YOLO-FIRI, where the feature extraction network is designed for complete use of the shallow features, and the detection head network has four layers to focus on small and weak objects. The speed and accuracy of the proposed models are qualitatively improved compared with the state-of-the-art infrared image object detection algorithms.

2. Our designed feature extraction network extends and iterates the shallow CSP module, which uses an improved attention module, forcing the network to pay more attention to the shallow and detailed features in infrared images, as well as make the model more robust by learning more distinguishable features.

3. Aiming at small objects caused by the problem of long-distance, the network detection head structure is improved, a multiscale object detection layer is added, and four-level spatial pyramid pooling is used to increase the receptive field and improve the detection accuracy of small objects.

4. In image preprocessing, we use convolutional neural network, Densefuse network, to fuse visible and infrared images, which can realize data enhancement. Thus, image fusion can be used as a data enhancement method to enhance the features of infrared images.

II. RELATED WORK

Infrared image object detection has the advantage of not being disturbed by the environment, and it is a research hotspot in the field of object detection [29]. Currently, infrared image object detection approaches have been divided into two types: traditional algorithms and CNN models.

A. TRADITIONAL INFRARED OBJECT DETECTION

Traditional infrared object detection methods consider infrared images as three parts: object, background, and noise in the images. The idea is to suppress background and noise, thus, strengthening the object to achieve object detection by using various methods. Zhao and Kong [7] first used the detection method based on spatial filtering for infrared object detection. In terms of different gray values of object and background, the background is selected and suppressed, and thus the object is detected. However, this method allowed all isolated noise points of small objects to pass, leading to a low detection rate. To address this problem, Anju and Raj [8] used the frequency difference between the object and background to separate the high-frequency part and the low-frequency part to achieve the detection task. Compared with that of spatial filtering methods, the detection effect of frequency domain filtering is a substantial promotion, but incurs high computational complexity. Jiao [9] adopted

sparse representation to cast the principle of infrared object detection in the form of a low-rank matrix and sparse matrix recovery, thus, achieving object segmentation and detection. The performance on the signal-clutter ratio (BCR) and background suppression factor (BSF) is much better than that of the filtering methods, but the nonlocal autocorrelation of the infrared image background cannot be used well, which leads to the lack of a background suppression effect. As the BCR decreases, the image background becomes increasingly complex. Enormous computational costs do not justify applicating the above models in real-time detection.

Traditional infrared pedestrian detection methods use artificially designed feature extractors, such as Haar [30], histogram of oriented gradients (HOG) [31], or aggregate channel features (ACF) [32] to extract the features of objects. Next, it takes advantage of the sliding window to extract local features, and then use support vector machine (SVM) [33] or AdaBoost [34] to determine whether there is an object in the region. Unfortunately, these infrared object detection algorithms have strong pertinence, a high time complexity, and window redundancy. They are also not robust to the changes in object diversity.

B. NEURAL NETWORK OBJECT DETECTION ALGORITHMS

Due to the improvement in computability and the widespread use of infrared imaging system equipment, many datasets are released to the public, such as KAIST [26], FLIR [28], and OTCBVS [35], which prompts deep learning to be gradually applied in the field of infrared image object detection. Thanks to the strong capability of feature expression, the CNN models open new horizons and create a large amount of excitement in object detection. They preserve the neighborhood relations and spatial locality of the input in their latent higher-level feature representations. Additionally, the number of free parameters describing their shared weights does not depend on the input dimensionality, meaning that the CNN can scale well to realistic-sized high-dimensional images in terms of computational complexity. The object detection networks mainly include two-stage detectors and one-stage detectors [36].

In 2014, R. Girshick *et al.* presented a pioneering two-stage object detector, R-CNN [14], and the main idea was to divide object detection into two steps: to generate proposals and predict objects. However, the computation was not shared and was extremely time-consuming. To accelerate inference speed and achieve better detection accuracy, Fast R-CNN [15] and Faster R-CNN [16] were developed by using ROI (Region of Interest) pooling and a novel proposal generator, RPN (Region Proposal Network), respectively. Currently, there are many network model variants based on Faster R-CNN for solving different problems, such as R-FCN [37], Mask R-CNN [12], and Sparse R-CNN [38]. In the field of infrared image object detection, Ghose *et al.* [39] proposed a method with a few modifications based on classical two-stage detector Faster R-CNN, which used the corresponding saliency maps to enhance the infrared images. However, since the process of training the saliency network was not added to the Faster R-CNN, the non end-to-end multitask training was very time-consuming. Devaguptapu *et al.* [40] developed a multimodel Faster R-CNN to obtain high-level infrared features through RGB channels, but the multimodel undoubtedly increased the inference speed of training. Park *et al.* [4] developed a CNN-based human detection method for infrared images. The performance was improved by performing pixelwise segmentation and making fine-grained predictions, whereas the proposed method lacked generality for the different datasets. Practically speaking, two-stage detectors have difficulty achieving real-time inference.

To address the issue of two-stage detectors, Redmon *et al.* [19] proposed a region-free one-stage object detection algorithm, YOLO, in 2016 that divided the image into grid cells and considered each cell as a proposal to detect the object. Compared with Faster R-CNN, YOLO omitted proposal generation and achieved end-to-end detection, which could realize real-time detection. Subsequently, the greatly improved detection speed gives rise to extensive research, such as SSD [17], YOLOv3 [21], and YOLOv4 [22]. To detect infrared objects, Kristo *et al.* [41] used the one-stage detector YOLOv3 to detect persons at night in different weather conditions. YOLOv3 is faster than the two-stage detector Faster R-CNN. However, YOLOv3 easily misses small objects, so its detection accuracy is very low. M. Li *et al.* proposed SE-YOLO [42], a real-time pedestrian object detection algorithm for small objects in infrared images, which improved the feature expression ability of the network combined with the SE block [43].To further improve the speed and accuracy of object detection, especially when objects are small and occluded, Li *et al.* [44] developed a detector, YOLO-ACN, by introducing an attention module, CIoU (Complete Intersection over Union) [45], [46] loss, improved Soft-NMS (Non-Maximum Suppression) [47], and depthwise separable convolution. The detector, YOLO-ACN, can focus on small objects and avoid the deletion of occluded objects. However, there were still a large number of parameters to save that the weight file was too large, which made it difficult to apply on mobile devices. In addition to the above methods based on the classic YOLOv3, there were some other one-stage network models for infrared image object detection. Cao *et al.* [48] presented a DNN-based one-stage detector, ThermalDet, which included a dual-pass fusion block (DFB) and a channelwise enhancement module (CEM). The mAP of ThermalDet is 74.6% in the FLIR dataset, and thus, cannot achieve the desired results. Dai *et al.* [49] presented an SSD-like object detection method TIRNet. In this method, VGG was adopted to extract features, and the residual branch was introduced to robust features. Although TIRNet only cost a little additional time, its detection performance still could not meet the actual application requirements. Song *et al.* [50] harnessed the features of infrared images and visible images to achieve fused features. Then, a multispectral feature fusion network (MSFFN) was

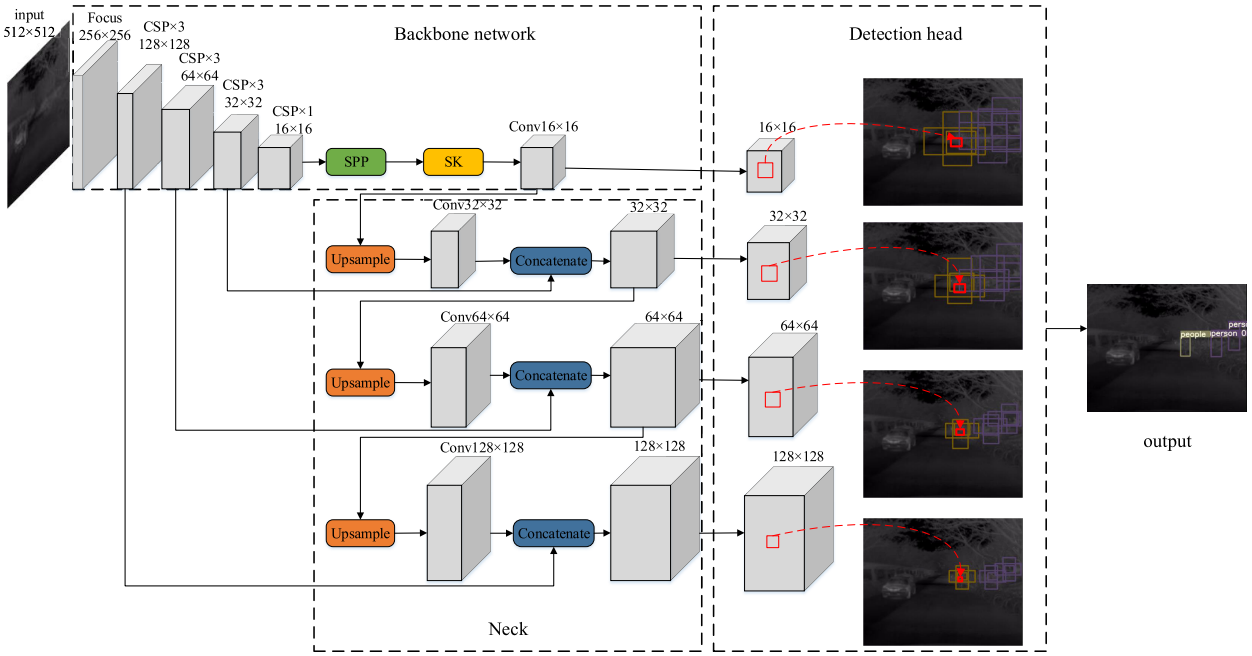


FIGURE 1. YOLO-FIRI overall network architecture. It mainly includes three parts: the Backbone network, the Neck, and the Detection head. The input single-channel infrared image is extracted through the backbone network with the extended CSP module. Multiscale features are further fused in the neck. Finally, the four-scale feature maps obtained are used to achieve multiscale object detection in the detection head.

proposed based on YOLOv3 to detect pedestrian objects, but the excellence of the MSFFN was only obvious when the input images were of a small size.

These methods achieved a better performance for nighttime object detection in different fields, such as pedestrian detection [4], [8], [41], [44] and autonomous driving [48], [49]. Despite the recent progress, it was difficult to transplant these models to mobile devices after training, especially for drone equipment, satellite equipment, infrared cameras, etc. To solve the problems in the existing models, this paper studies the state-of-the-art detector YOLOv5, which was first released on June 25, 2020. Based on studying the unique features of infrared images, this paper proposes the YOLO-FIR method for infrared objects with the core of YOLOv5 by analyzing the network structure, compressing channels, optimizing parameters, etc. Furthermore, by extending the thickness of the shallow CSP module that contains rich feature information in the backbone network, incorporating an improved SK attention module in the residual blocks to boost the feature extract ability and adding the detection layer to detect smaller objects, YOLO-FIRI, an improved infrared image object detection framework, is further proposed. In addition, the Densefuse network is used to fuse infrared images and visible images to generate more informative fused infrared images. Experimental results show that, compared with the latest infrared image object models, whether on the KAIST dataset or the FLIR dataset, the proposed object detection model YOLOv5-FIRI for infrared images brings about notable improvements in detection accuracy, detection speed, and model size. In addition, the detection accuracy of the proposed models on the fused dataset is improved to a certain degree.

III. PROPOSED METHOD

A. NETWORK ARCHITECTURE

One-stage deep convolutional neural networks YOLOv3 and YOLOv4 have achieved a good performance in object detection. YOLOv5 uses a variety of network structures and two types of CSP modules to improve YOLOv4, so that YOLOv5 is very conducive to object detection and recognition in terms of detection accuracy and computational complexity. Therefore, this paper proposes a method, YOLO-FIR, for infrared objects with the core of YOLOv5 by analyzing network structure, dividing data, compressing channels, optimizing parameters, training, and testing the model, etc. As a result, a novel network model, YOLO-FIRI, is designed and implemented to detect small and weak objects quickly in infrared images. The structure of YOLO-FIRI is shown in Figure 1, which mainly includes three parts: a backbone network with a lightweight feature extraction network, a neck network to realize cross-stage feature fusion, and a multiscale detection head.

In Figure 1, the input infrared image changes from $512 \times 512 \times 1$ (infrared images have a single channel) to $256 \times 256 \times 4$ after the focus operation. Then, in the backbone network, the extended CSP module is used to extract rich information from shallow and deep feature maps after the focus operation, and the attention mechanism introduced in the CSP module guides the assignment of different weights to realize and notice the extraction of weak and small features. In addition, the SPP (Spatial Pyramid Pooling) [51] layer can concatenate the results obtained in the channel dimension through four pooling windows to solve the alignment problem of anchors and feature maps. We use the SK attention module to enhance the extracted features. Second, in the neck

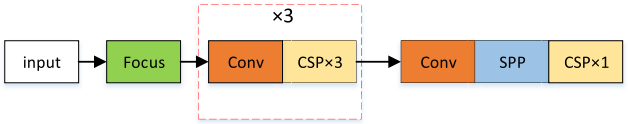


FIGURE 2. The Backbone structure with the extended shallow CSP module. We extend the shallow CSP to achieve the same number of iterations as the deep CSP (which is iterated 3 times). Then, the network can realize the focus on shallow features.

network, PANet (Path Aggregation Network) [52] is used to generate feature pyramids, and the top-down and bottom-up fusion structures are both used to effectively fuse the multiscale features extracted from the backbone network and enhance the detection of objects with different scales. Finally, in the detection head, the four sets of output feature maps are detected, and the anchor boxes are applied on the output feature maps to generate the final output vectors with a class probability score, a confidence score, and a bounding box. Then, according to the NMS [53] postprocessing, the results detected by the four detection layers are screened to obtain the final detection results. The additional set of feature maps can solve the problem of missed and false detections caused by long-distance shooting. The proposed network models demonstrate a qualitative improvement compared with the latest infrared image object detection in terms of detection accuracy, reasoning speed, and network parameters.

B. EXTENDED CSP MODULE

With network layers deepening gradually, the convolutional neural network can better extract the semantic information of high-level features, but the resolution of the high-level feature maps is low. In contrast, the resolution of the feature map in the shallow layer is high, whereas the feature semantic information extracted by the shallow network is weak. For small and weak objects with a few features in infrared images, deep convolution may cause object features to be difficult to extract or even lost. To maximize the extraction of features that are conducive to the detection of weak and small objects in infrared images, it is necessary to make full use of the high-level resolution features of the convolutional neural network in the shallow layer. Thus, in the feature extraction stage, we extend the thickness of the CSP module in the shallow feature extraction process. Through feedback and iteration step-by-step, the object features in the feature maps can be fully extracted to achieve multifeature extraction from shallow layers to deep layers. Moreover, when deepening the CSP modules in the overall feature extraction network by controlling the width and depth factors, we only extend the thickness of the CSP module to extract shallow features. The backbone structure of the entire feature extraction network is shown in Figure 2. In this way, without substantially increasing the size of the network model and the complexity of the algorithm, the ability to extract shallow feature information is enhanced, which is conducive to the detection of weak and small objects in infrared images. In addition, the CSP structure divides the feature maps into two branches to extract

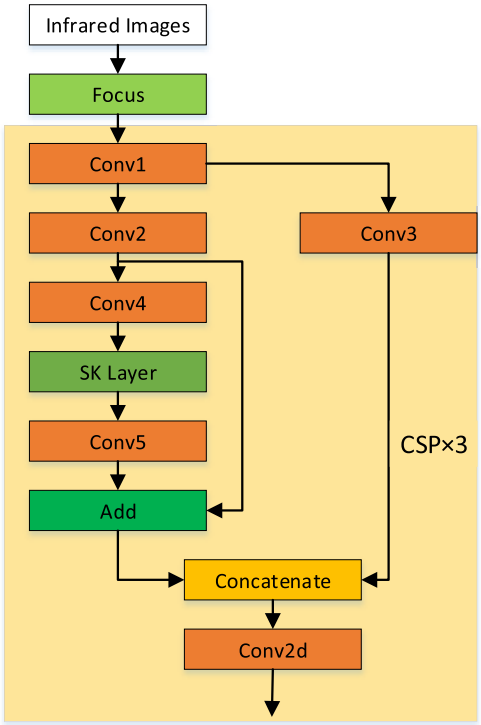


FIGURE 3. The extended CSP module with the improved SK attention module. The attention module is added between the two convolutional layers as an SK layer.

features and then merge them, which can achieve a richer gradient combination while reducing the amount of calculations.

In Figure 2, after the focus slicing operation, the Conv and CSP modules are stacked three times, the shallow layer is extended to the same number of CSP module feedback iterations as the deep layer, and feature maps of different sizes are obtained step-by-step. Then, the full extraction of fine-grained features of shallow information and deep high-level semantic information are obtained, and the specific CSP module structure is shown in Figure 3. Conv represents the three operations of standard convolution, normalization, and the activation function, while Conv2d represents standard convolution. Through the concatenate operation, the feature maps containing the two branches of the Conv and the attention module (SK Layer) are merged, and the 128×128 features are fully extracted in the shallow layer. Compared with the YOLOv5m model that adds 108 layers, we only add 18 layers to the network by extending the shallow CSP module. We ensure that the detection speed of the model is not reduced when improving the detection accuracy of the model.

C. IMPROVED SK ATTENTION MODULE

The visual system tends to pay attention to a part of the information that assists with judging the image and ignores the unimportant information [54]. In object detection, an attention mechanism can be introduced in the residual blocks of the shallow feature extraction stage to effectively select object information, and more weights can be assigned to

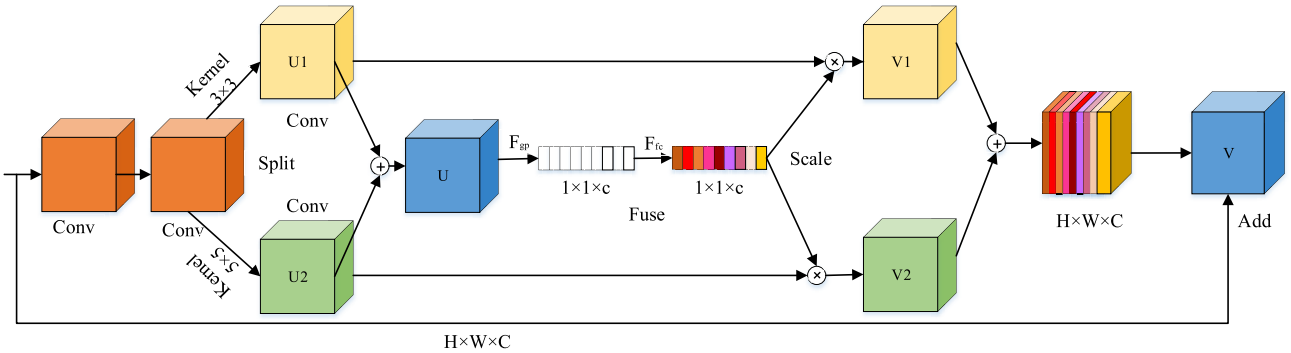


FIGURE 4. The improved SK attention module. We implement the improved SK convolution via three operators: split, fuse, and scale. In the split operation, two kernels with different sizes are split; in the fuse operation, global pooling and a full connection are used; in the scale operation, the results generated by the above two-stage are weighted.

small and weak objects to improve the feature expression ability of small objects for accurate detection. The SKNet [25] network can adaptively adjust the size of the receptive field according to multiple scales of the input information, and better extract objects with different sizes and distances. Therefore, we introduce an improved SK attention module in each CSP module and use two convolutional operations with different convolution kernel sizes to learn the channel weights. The output vector continues to perform 1×1 convolution operations. The corresponding introduction position in the CSP module is the SK layer in Figure 3, and the specific improved SK attention mechanism structure is shown in Figure 4.

As shown in Figure 4, after two Conv modules, which include standard convolution, normalization, and activation functions, the improved SK attention module is directly embedded into the residual blocks, and it is mainly divided into three parts: split, fuse, and scale. The split operation separates the input vector by performing the Conv operation with two different sizes of kernels, 3×3 and 5×5, to obtain the output vectors $U1$ and $U2$, and to obtain the vector U after the addition operation. The fuse stage uses global average pooling (F_{gp}) to compress the matrix to $1 \times 1 \times$C and uses a channel descriptor to represent the information of each channel. Therefore, the dependency between the channels is established, which can be expressed as (1), and the fully connected layer (F_{fc}) makes the relationship between the channels flexible and nonlinear. Here, two fully connected layers are also used to add more nonlinearity, fit the complex correlation between the channels, reduce the number of parameters and calculations as much as possible, and obtain the weight value, which is given by the Eq.:

$$F_{gp}(U) = \frac{1}{W \times H} \sum_{i=1}^{H} \sum_{j=1}^{W} U(i, j) \tag{1}$$

$$F_{fc}\left(F_{gp}, \omega\right) = \sigma\left(B\left(F_{gp}, \omega\right)\right) \tag{2}$$

In (1), W and H are the width and height, respectively, and i and j are the i-th row and j-th column of the image, respectively. In (2), ω is the weight, σ is the ReLU activation function [55], and B represents the batch normalization operation.

Scaling is a simple weighting operation. The weight values calculated in the fuse stage are multiplied back to the original matrix to obtain the final output of the SK blocks, which can strengthen the useful information of weak and small objects for different channelwise scenarios. The matrix vectors are added again and merged to make full use of the shallow and deep layer information. By using a simple and effective fully connected layer, the output obtained after the sigmoid activation weight value is directly multiplied by the vector U to obtain the vector V instead of generating vectors a and b (two weight matrices to multiply). Thus, the computational complexity is reduced, and the reduction of the inference speed caused by the increased network layer is avoided.

$$F_{scale}\left(U, F_{fc}\right) = V1 + V2 = U1 \cdot F_{fc} + U2 \cdot F_{fc} \tag{3}$$

In (3), $F_{scale}\left(U, F_{fc}\right)$ is channelwise multiplication, multiplying the feature maps U with the weight value obtained in the F_{fc} stage, and outputting the weighted feature maps.

SK is a lightweight module that can be directly embedded in the network. It has a strong generalize ability by acquiring different receptive field information and an adaptive adjustment structure which is beneficial to the detection of pedestrians in infrared images. Moreover, systemic improvements can be achieved with a minimal computational burden.

D. MULTISCALE FEATURE DETECTION

The YOLOv5 network uses three types of output feature maps to detect objects with different sizes and uses 8 downsampling output feature maps to detect small objects. The objects in the KAIST [26] dataset are small and weak; therefore, we add a feature scale to focus on smaller objects. When feature maps are upsampled to the size of 64×64, we continue to upsample the feature maps to obtain 4 downsampling feature maps. At the same time, the expanded 128×128 feature maps are fused with the same size feature map of the second layer in the backbone network to make full use of the shallow and deep features. After multiscale fusion, the four feature scales are 128×128, 64×64, 32×32, and 18×18, as shown in Figure 5. The 32×32 marked in the grid division represents the size of each grid. Nine anchors with three detection scales are increased to twelve anchors with four detection scales.

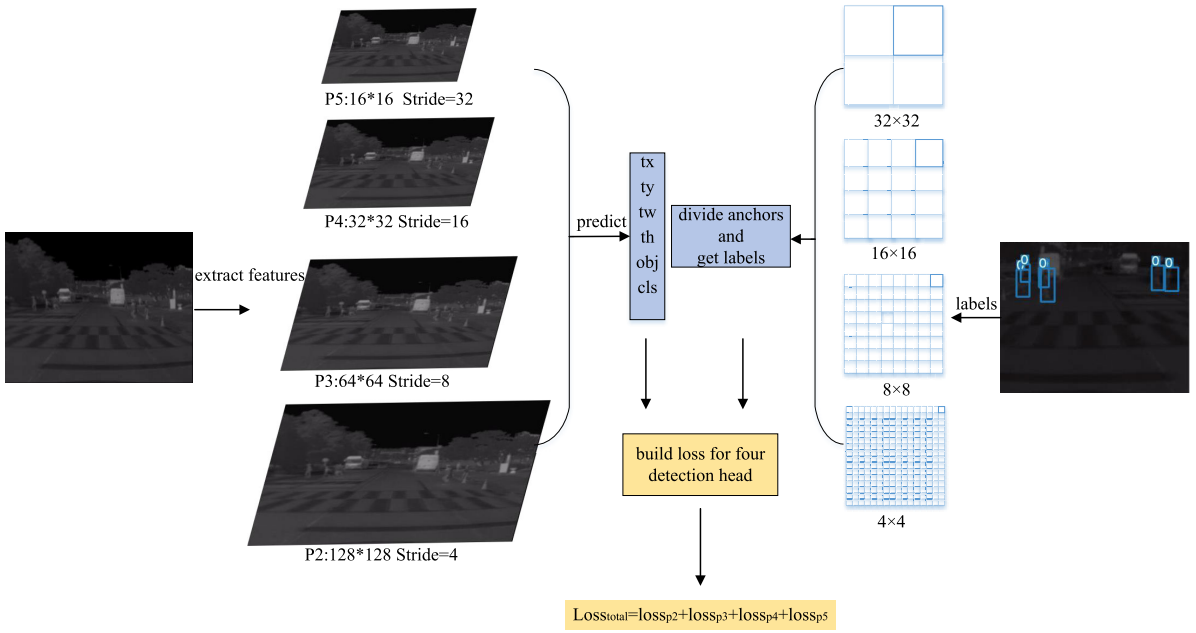


FIGURE 5. Multiscale detection structure. The input image through the network extracts features, and then four sets of feature maps are finally detected to predict the six values of each object. According to the label information, the loss between the prediction and ground truth is calculated, and then the training is completed.

YOLOv5 can adaptively calculate suitable anchors according to different datasets, making it easier for the model to learn the converge and predict objects with different scales.

In Figure 5, the left part is the prediction of the model, and the four detection layers (P2-P5) predict the values, i.e., the central point *tx* and *ty*, the width *tw* and height *th*, and the confidence score. The right part is the ground truth of the objects, and the network obtains the label information of the input images. Then, the loss between the prediction value and the ground truth is established to calculate the loss of each detection layer. Through the feedback of the loss, the model gradually optimizes the performance and completes the training. The loss calculation method for each detection layer is the same, which is obtained by calculating the sum of the bounding box regression loss, class loss, and confidence loss, e.g., the calculation of the P2 detection layer is shown as Eq.:

$$loss_{p2} = loss_{ciou} + loss_{cls} + loss_{obj} \tag{4}$$

Here, the bounding box loss ($loss_{ciou}$) uses CIoU, the class loss is calculated through BCE (Binary Cross Entropy) loss, and the confidence loss is realized by BCE with logits loss to get numerical stability.

IV. EXPERIMENT ANALYSIS

To test the performance of the infrared image object detection models YOLO-FIR and YOLO-FIRI which are proposed in this paper, we use the public KAIST and FLIR infrared pedestrian dataset. First, the latest detection algorithms and the detection models proposed in this paper are compared in terms of detection accuracy, speed, computational complexity, parameters, etc. Second, an ablation experiment is carried out on the improved YOLO-FIRI model to test the performance brought by the different improved methods. Third, the KAIST dataset provides well-aligned visible and infrared image pairs, so the Densefuse deep neural network is used to fuse the visible and infrared images, enhance the characteristics of the objects, and generate a fusion infrared image dataset. Then, YOLO-FIR and YOLO-FIRI experiment on the fused dataset. Finally, the FLIR dataset is also used to further test the performance of the proposed models.

A. COMPARISON OF THE DETECTION PERFORMANCE

The KAIST infrared pedestrian dataset is a classic public object detection dataset evaluated by most infrared image object detection algorithms. It contains large-scale, accurate manual annotations, well-aligned visible and infrared image pairs, and has a total of 95,328 pairs of images (640 × 512 resolution), including various conventional traffic scenes on campus, street, and countryside. The dataset labels include two pedestrian object categories: person and people. Those who are better to distinguish are labeled as person, and multiple individuals who are not easy to distinguish are labeled as people. Li *et al.* [56] cleaned the training set and the test set to generate a training and testing set (7,601 examples). Liu *et al.* [57] cleaned the test set to 2,252 examples and realized the data focusing on the filtering of images that contain the object. However, to ensure the diversity of positive and negative samples, we choose to use the original KAIST dataset. For the large-scale original data, we intercept 17498 consecutive images for training and testing. The experimental results show that these data is sufficient to achieve a successful model training and performance evaluation, while also achieving a high detection accuracy.

TABLE 1. Quantitative comparison of YOLO-FIRI and the other state-of-the-art object detectors. The results are reported in terms of AP, AR, mAP50, and F1 percentage to evaluate the accuracy, and detection times, parameters, weight size, and FLOPs to assess the speed. All the experiments use the KAIST test set. The speed and accuracy of the proposed models are qualitatively improved compared with the state-of-the-art infrared images object detection algorithms.

Method	AP/%	AR/%	mAP/%	F1/%	mAP50:75/%	Time/ms	Params/M	Weight/MB	FLOPs/B
YOLOv3	73.5	76.9	79.6	74.8	57.0	25	61.5	246.4	155.1
YOLOv4	76.9	75.8	81.0	76.3	58.9	37	63.9	256.3	128.4
YOLO-ACN	76.2	87.9	82.3	81.6	59.3	20	47.4	177.6	111.8
YOLO-FIR	92.1	88.1	93.1	90.0	82.4	12	7.1	16.2	16.4
YOLO-FIRI	**96.0**	**96.2**	**98.3**	**96.1**	**95.6**	**14**	**7.2**	**15.0**	**20.4**

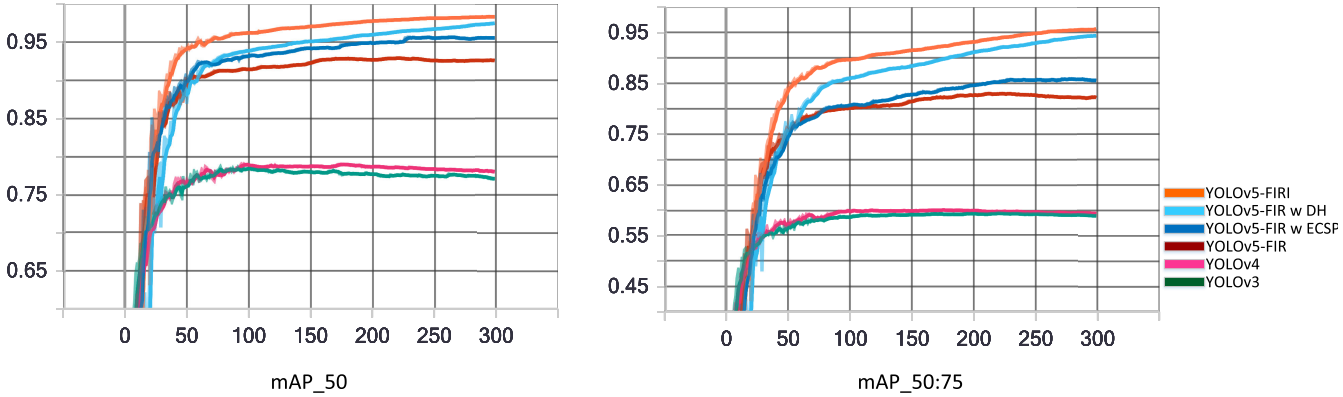


FIGURE 6. Performance comparison of the mean average precision using the training set of the KAIST. The left curves are mAP50 (IoU=0.5). The right curves are mAP50:75 (the AP by 6 IoU thresholds of 0.5:0.05:0.75 is averaged). These curves separately represent YOLOv3, YOLOv4, and the proposed models YOLO-FIR and YOLO-FIRI. YOLO-FIR w ECSP and YOLO-FIR w DH are the proposed YOLO-FIR with the extended CSP module and the detection head.

Table 1 is the comparison of the state-of-the-art infrared image object detection algorithms and the proposed YOLO-FIR and YOLO-FIRI in this paper in terms of the various evaluation indicators. Here, the size of the input image is 512 × 512, the training epochs are set to 300, the batch size is 16, the initial learning rate is 0.001 and learning rate decay of 0.01 every 5 epochs, the IoU threshold is set to 0.20, and the momentum and weight decay are 0.937 and 0.0005, respectively. We keep mosaic as 1 to use the mosaic data enhancement algorithm to expand the diversity of object samples. The training of all experiments is based on the PyTorch and carried out on the GeForce GTX 1660 GPU.

In Table 1, YOLO-FIR and YOLO-FIRI are the proposed infrared image object detection models based on YOLOv5. Compared with YOLOv3, YOLOv4, and YOLO-ACN, YOLO-FIR and YOLO-FIRI have greatly improved in various indicators. In particular, YOLO-FIRI is 24.8%, 26.9%, and 26.0% higher than the latest classic one-stage object detection algorithm YOLOv4 in detection accuracy indicators AP, AR, and F1, and the mAP50 has also been improved by approximately 21.4%. Compared with YOLO-FIR, the AP, AR, and mAP50 are improved by 3.9%, 8.1%, and 13.2%, respectively. The object features in the infrared images are weak and small. If the IoU threshold is set too large, it will be more unfavorable for detecting objects in infrared images. In contrast, setting the highest IoU value to 0.75 is more common and suitable for actual applications. We use mAP50:75 to calculate the average detection accuracy under six different IoU thresholds. Table 1 shows that the mAP50:75 of YOLO-FIRI is approximately 62.3% higher than that of YOLOv4, and the detection time and calculation amount decrease by approximately 62.2% and 84.1%, respectively, which greatly improves the real-time processing capability and reduces the hardware calculation requirements. In terms of the number of parameters, YOLO-FIRI directly reduces from tens of millions to millions. The corresponding weight file size also reduces from more than 200 MB of YOLOv3 and YOLOv4 to 15 MB of YOLO-FIRI, which is more conducive to model transplantation for mobile devices. This is mainly because the YOLO-FIR and YOLO-FIRI models use CSPDarknet as the backbone, and perform channel compression, as well as parameter optimization. As a result, the models are able to solve problems, such as repeating gradient information in network optimization in the backbone of other large-scale convolutional neural network frameworks. The gradient changes are integrated into the feature map from beginning to end, thus reducing the number of parameters and flop value of the model, which not only ensures the inference speed and accuracy, but also reduces the model size.

YOLO-FIRI is an improved infrared image object detection framework for weak and small objects in infrared images. Compared with YOLO-FIR, under the condition that the detection time, computational complexity, parameters, etc. are basically unchanged, the detection accuracy indicators AP, AR, mAP and F1 have been improved by approximately 4.2%, 9.1%, 5.6%, and 6.8%, respectively, especially mAP50:75 by approximately 13.2%. This indicates that the detection accuracy has significantly progressed after

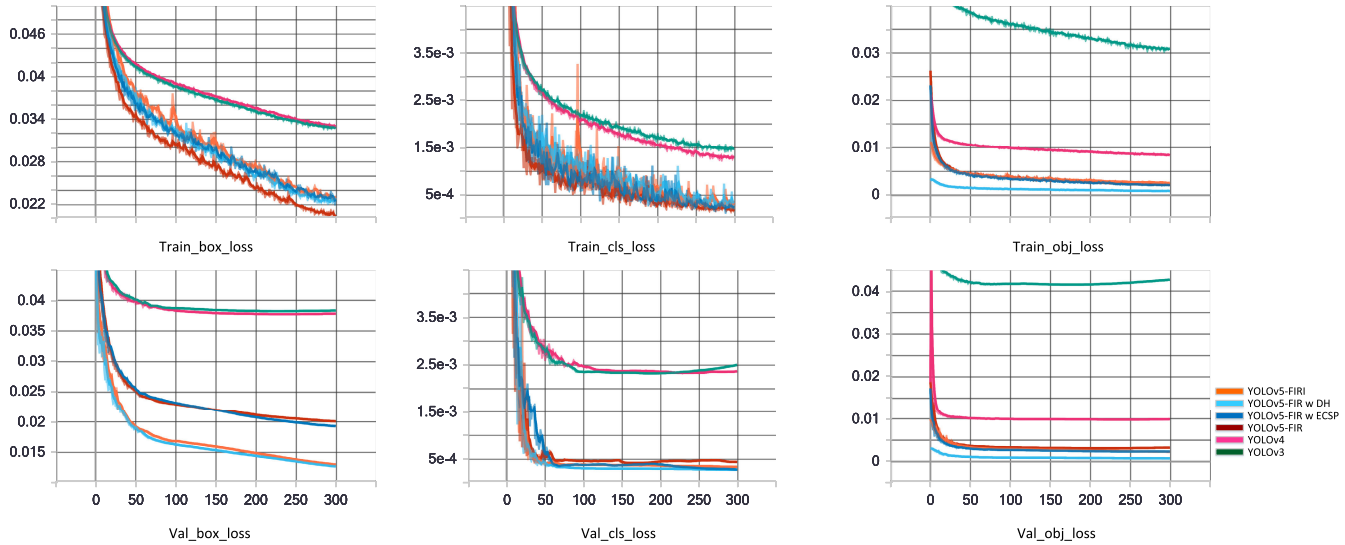


FIGURE 7. Performance comparison of three kinds of loss using the KAIST dataset. The first row is the loss of the training set of KAIST, and the three figures in the first row from left to right are the box regression loss (box_loss), the class loss (cls_loss), and the object loss (obj_loss). The second row is the loss of the test set of KAIST, and the three figures in the first row from the left are the same as those in first row.

improving the detection of weak and small objects in infrared images. To further compare the accuracy differences between the two classes, we test each accuracy evaluation index on the KAIST dataset with YOLO-FIR and YOLO-FIRI. The results are shown in Table 2. In the detection of the two classes, AP and AR increased by 4.6%-8.1%. Moreover, the mAP50 for person and people reached 98.8% and 99.0%, respectively, and the mAP50:75 for the class of people increased by 13.6%.

To study the changes within the different detection models in the training process, Figure 6 shows the training accuracy of YOLOv3, YOLOv4, YOLO-FIR, and the different improved methods based on YOLO-FIR. The left image is the mean average precision of the two classes when the IoU is 0.5 (mAP50), and the right image is the mean average precision of the six different thresholds when the IoU is 0.5 to 0.75 (mAP50: 75). Because YOLOv4 improves the feature extraction network and data enhancement technology of YOLOv3, its detection accuracy is slightly better than that of YOLOv3. However, the YOLO-FIR and YOLO-FIRI detection methods proposed in this paper have a much higher performance than YOLOv4 whether in the mAP50 on the left or the mAP50:75 on the right, and the number of trainings required to enter the stable state is also relatively low. Compared with YOLOv3 and YOLOv4, the accuracy of mAP50 approaches 80%, whereas the improved YOLO-FIRI approaches 98%; the mAP50:75 approaches 60%, whereas the improved YOLO-FIRI approaches 95%, which greatly improves the average detection accuracy. In addition, as the epochs increase, the detection performance of YOLOv3 and YOLOv4 does not always exhibit an upward trend. If the training is continued, overfitting will occur [58], which will cause a slight decrease in the detection performance. However, YOLO-FIR and YOLO-FIRI adapt a variety of CSP structure extraction and fusion features. As the number of trainings increases, their detection performance maintains a steady trend.

TABLE 2. Performance of precisions about two classes on the test set of the KAIST. This is mainly due to YOLO-FIRI's improvements in extending and iterating CSP modules, introducing attention mechanisms, and improving multiscale detection heads.

Method	classes	AP/%	AR/%	mAP50/%	mAP50:75/%
YOLO-FIR	person	92.0	92.4	95.9	87.8
YOLO-FIRI	person	**97.4**	**97.0**	**98.8**	**96.9**
YOLO-FIR	people	93.2	88.5	91.6	83.9
YOLO-FIRI	people	**97.2**	**96.6**	**99.0**	**97.5**

Loss plays an important role in the training process, which reflects the relationship between the true value and the predicted value. The smaller the loss is, the closer the prediction value is to the true value, and the better the performance of the model. Figure 7 shows the loss convergence rates of YOLOv3, YOLOv4, YOLO-FIR, and different improved methods based on YOLO-FIR. From Figure 7, we can see that the bounding box loss, class loss, and object loss in the training set and the validation set present a falling trend and eventually stabilize. Whether in the training or the validation, the improved YOLO-FIRI has been considerably reduced compared with YOLOv4. Compared with the proposed YOLO-FIR, although the curve is relatively close, there is still a slight decrease for the bounding box regression loss in the validation set. YOLOv4 is 0.037 at 300 epochs, whereas the bounding box regression loss of YOLO-FIRI is 0.012, which means that the proposed method can significantly accelerate the network's training process and converge to a lower loss when optimizing the neural network.

Examples of the test results on the YOLOv4 and the YOLO-FIRI models are shown in Figure 8. In terms of the number of detected objects, YOLO-FIRI detects 1 or 2 more pedestrian objects than the YOLOv4 in each image. As shown in (a2) and (b2), YOLO-FIRI can solve the

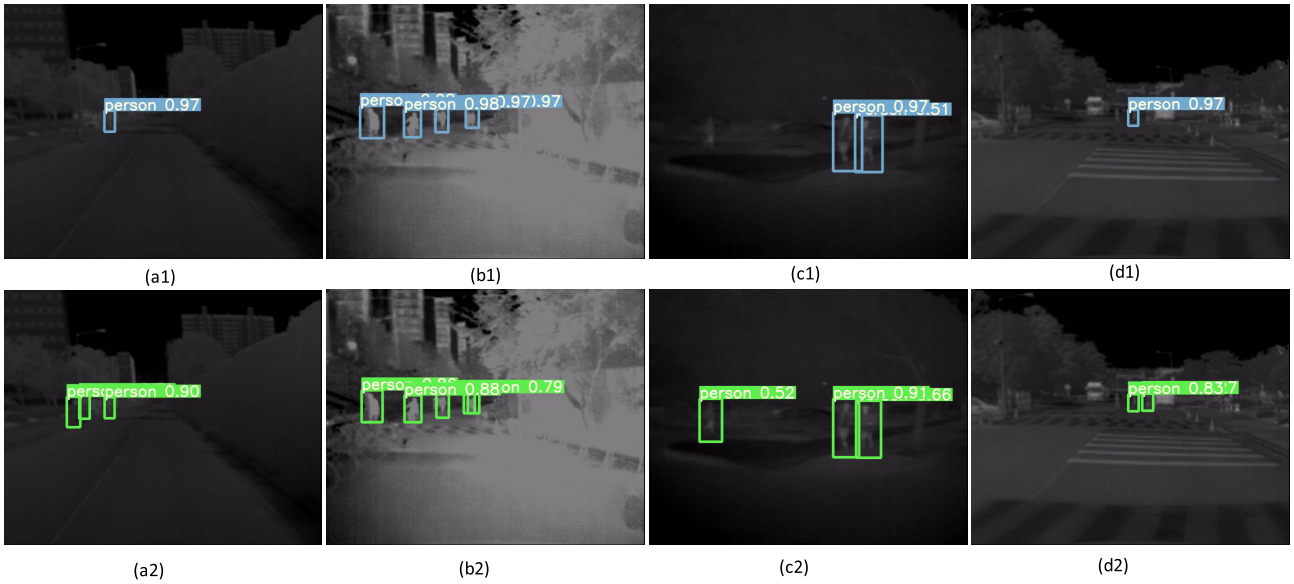


FIGURE 8. Some examples of the detection result on the test set of the KAIST dataset. The first row is the result of YOLOv4, and the second row is the result of YOLO-FIRI. We used the same four images to compare the performance of the detection models.

TABLE 3. Ablation study of detection precision on the test set of KAIST.

CSP	SK	HEAD	AP/%	AR/%	mAP50/%	mAP50:75/%
√			92.0	93.9	97.5	94.3
	√		92.4	87.4	95.5	85.5
		√	93.2	93.1	97.0	94.1
√	√	√	96.0	96.3	98.3	95.6

missed detection problem of YOLOv4 by effectively extracting features for the possible occlusion of parallel pedestrians; for the long-distance pedestrians detected in (c2) and (d2), YOLO-FIRI realizes the detection of small objects through multiscale detection. Although the object in the infrared image is difficult to distinguish from the background, the improved model can still achieve the detection of the pedestrian in the infrared images with different distances by enhancing shallow features, fusing multiple features, and improving multiscale detection.

B. ABLATION STUDY

To see the effect of the different improved technologies more intuitively on the performance of the model, we conduct an ablation experiment. Specifically, keeping the structure of YOLOv5s unchanged and only improving the extended CSP module, we can observe the impact of the performance. Then, we improve the SK attention module, and multiscale feature detection to observe the experimental results and analyze the influence.

Our ablation experiment also keeps training for 300 epochs. When the training result is stabilized, the training is completed and then tested on the KAIST test set. The indicators are shown in Table 3. By introducing the extended CSP, the improved SK attention module, and the added detection head, the accuracy indicators of object detection have been improved accordingly. When the integration of these four improvements is tested as the final network model, the tested indicators show better detection accuracy than the three methods introduced separately. The corresponding mAP50 increased by 2.8%, and the mAP50:75 achieved a maximum increase of 10.2%.

C. INFRARED OBJECT DETECTION ON THE FUSED KAIST

In the field of image analysis, the quality of the image directly affects the design of the algorithm and the accuracy of the detection. Compared with visible images, infrared images have lower resolution and blurred visual effects. At present, the performance of the object detection model based on deep learning proves to be high quality and the image performance is better as well. However, when it is applied to infrared images to detect weak and small objects, the performance of the model is greatly reduced. Therefore, to improve the detection performance of weak and small objects, data enhancement as an effective method has been proven to solve the challenges brought by object detection tasks, such as Cutout [59], CutMix [60], Keep Augment [61], and other data enhancement methods by improving the image resolution [17], [62]. Image fusion can be a data enhancement method that performs fusion processing for a variety of different types of source images. Compared with a single image, image fusion has a substantial enhancement in image quality and clarity. The KAIST dataset provides infrared and visible image pairs of the same scene. Given the lack of a publicly available infrared dataset, the generated fusion dataset also achieves data diversity.

Visible images can reflect the spectral information properties of objects, containing more detailed information, and are more in line with visual characteristics of the human eyes. And the thermal radiation characteristics of infrared images are more sensitive to objects and areas, which

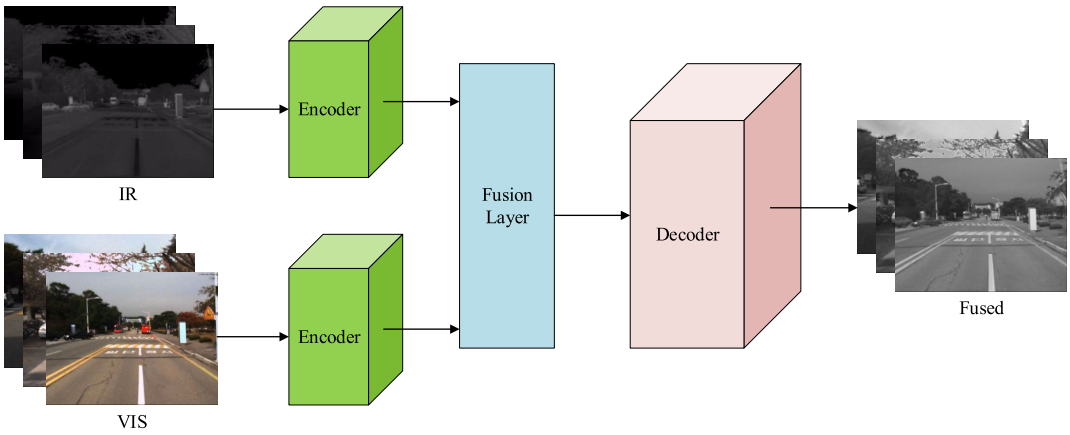


FIGURE 9. The architecture of the Densefuse network. The input images are the well-aligned visible and infrared images. The encoder network uses a convolutional and dense block to extract more useful features from the source images. The fusion layer is used to fuse the extracted features. The decoder network can rebuild the fused infrared images to realize the fused KAIST dataset.

can avoid interference caused by scene changes. Therefore, infrared image and visible image are complementary, and the image fusion method can be used to fuse infrared and visible images into a more informative infrared image by using the fusion method to achieve the purpose of data enhancement.

Figure 9 shows the Densefuse [27] framework for fusing infrared and visible images, which mainly includes three parts: encoder, fusion layer, and decoder. The encoder mainly includes two kinds of layers, C1 and a dense layer. C1 contains a 3×3 convolution kernel to extract rough features, and the dense layer contains three convolutional layers to extract deep high-level features. The encoder convolution kernel size and convolution stride are 3×3 and 1, respectively, which can receive images of any size, while the dense layer can retain depth features as much as possible in the encoding network. The fusion layer uses an additive strategy to fuse the infrared and visible image features extracted by the encoder, which can be written as Eq.:

$$F^m(x, y) = \lambda Vis^m(x, y) + (1 - \lambda) Ir^m(x, y) \tag{5}$$

In (5), $Vis^m(x, y)$ represents the visible image of the m-th channel, $Ir^m(x, y)$ represents the infrared image of the m-th channel, $F^m(x, y)$ indicates the fusion result of the m-th channel, and λ is the weighting coefficient. The output of the fusion layer enters the decoder, which contains four 3×3 convolutional layers to reconstruct the fused image. The loss function of the network $H(x, y)$ is weighted by the structural similarity loss function $H_{SSIM}(x, y)$ and the pixel loss function $H_P(x, y)$.

$$\begin{aligned} H(x, y) &= \gamma H_{SSIM}(x, y) + H_P(x, y) \\ &= \gamma \left(1 - SSIM\left(Out(x, y), In(x, y)\right)\right) \\ &\quad + \|Out(x, y) - In(x, y)\|_2 \end{aligned} \tag{6}$$

In (6), $Out(x, y)$ and $In(x, y)$ represent the output image and the input image, respectively. $H_P(x, y)$ is the Euclidean distance between $Out(x, y)$ and $In(x, y)$. $SSIM(\cdot)$ is the structural similarity. Considering that there are three orders of magnitude differences between the pixel loss and SSIM loss, we take 100 here.

In this experiment, we use the Densefuse network, which is shown in Figure 9, to obtain the fused KAIST dataset. The image comparison of the infrared images, visible images, and fused images of the KAIST are shown in Figure 10. Among them, the detailed information of some objects is marked with red boxes for highlighting. Compared with the infrared image (a1), the fused image (c1) has a clearer body posture and contour, which expands the data features, so that the object features are more obvious and easier to extract. Compared with the visible image (b2), the fused image (c2) avoids the influence of the red bus occlusion in the background, making the object easier to recognize instead of being filtered out as the background. Compared with the infrared image (a3) and the visible image (b3), the number of objects and human characteristics in the fused image (c3) are clearer.

For the visible, infrared, and fused infrared datasets, the proposed network models are trained separately. Table 4 compares the test results of visible, infrared and fused infrared images on YOLO-FIR and the improved network model YOLO-FIRI. On the detection model of YOLO-FIR, when fused dataset is used as input, both AP and mAP50 are improved, and the detection accuracy of mAP50 is improved by 0.7% compared to a single infrared image. On the YOLO-FIRI, the fused infrared dataset achieved the best results, and the mAP increased to 98.5%. The KAIST dataset has serious occlusion problems in visible images, but it can effectively improve this problem on the infrared dataset. Therefore, on the YOLO-FIRI models, the performance of the fused infrared dataset is better than that of the visible dataset and the infrared dataset.

After training, the models are further tested on two different types of datasets. Figure 11 shows some randomly selected images of infrared and fused datasets. In Column (a), the fused image can extract pedestrian objects that are occluded due to the shooting angle; in Column (b), pedestrian

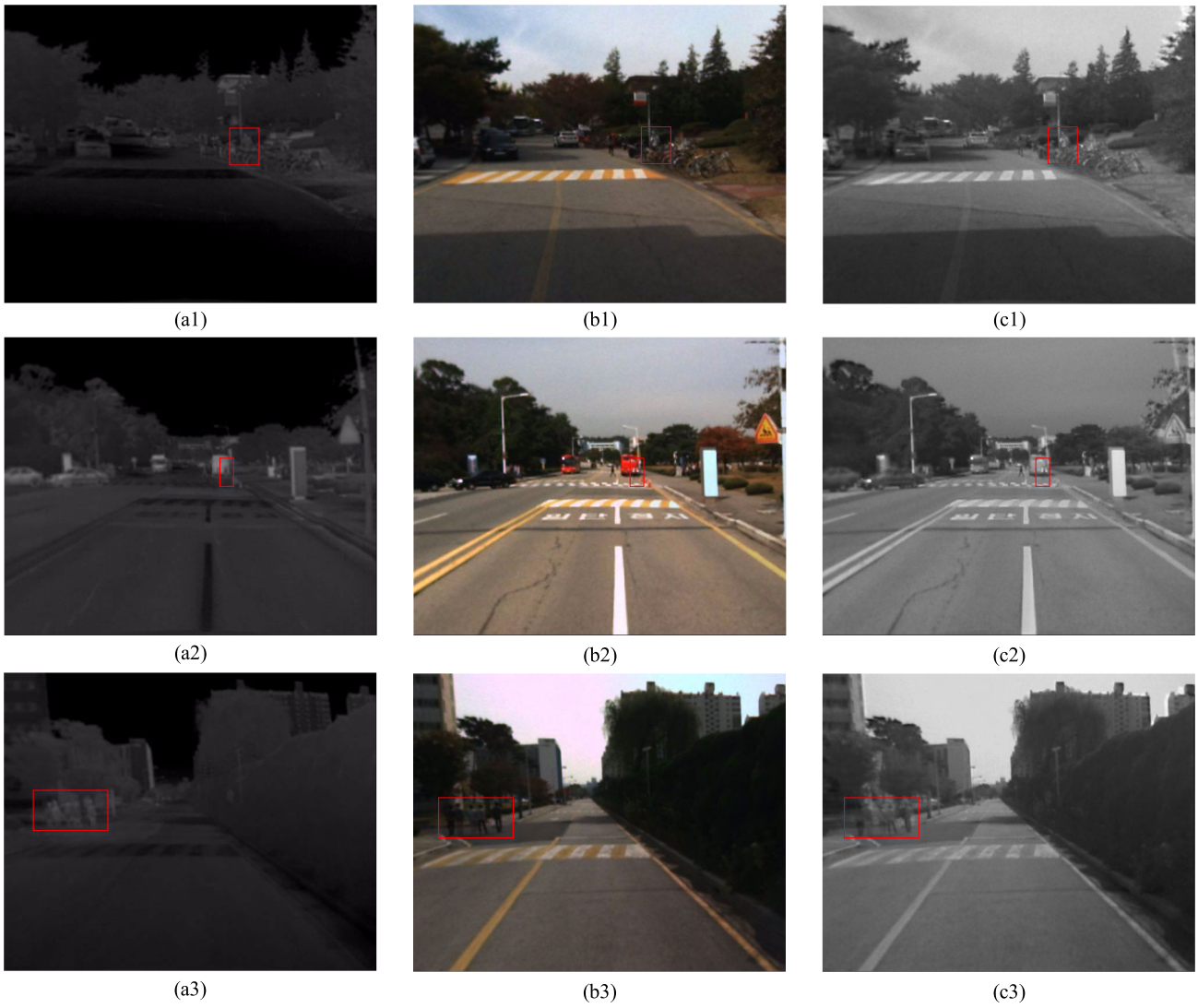


FIGURE 10. Some image examples of infrared images, visible images, and fusion images on the KAIST dataset. The first column is the infrared images, the second column is the visible images, and the third column is the fused images.

TABLE 4. Performance comparisons on the visible (VIS) dataset, infrared dataset, and fused dataset of the KAIST.

Type	AP/%	AR/%	mAP/%	F1/%
YOLO-FIR: Accuracy of Object Detection				
VIS	92.4	87.3	95.5	89.7
IR	92.1	88.1	93.1	90.0
Fused	92.4	84.9	93.8	88.1
YOLO-FIRI: Accuracy of Object Detection				
VIS	92.0	93.9	97.4	92.9
IR	96.0	96.2	98.3	96.1
Fused	**97.4**	**96.4**	**98.5**	**96.8**

objects at the edge can also be detected more accurately after fusion. In the densely crowded images of the two Columns (c) and (d), the number of pedestrian objects detected has increased from two and six to three and seven, respectively. We can see that the fused test set can detect more pedestrian objects than the infrared test set, and the overall accuracy rate has been improved.

D. INFRARED OBJECT DETECTION ON THE FLIR DATASET

On the KAIST dataset, we compare different models of the YOLO series. To better test the detection performance of the proposed models on infrared images, we test YOLO-FIRI and compare the different detection approaches on the FLIR dataset [28]. In the FLIR dataset, it provides both visible and infrared images, whereas the visible and infrared image pairs are not well aligned. Some infrared images do not have corresponding visible images, and only the infrared images are labeled. Therefore, we simply train on the infrared images and do not need to adapt a pretrained detector with RGB images. The training set and the test set contain 8862 and 1366 images with 640 $\times$ 512 resolution, and a total of three classes are included. In the experiments, we use the train set and test set as provided in the dataset benchmark. The experimental setting is still the same as that of the KAIST dataset.

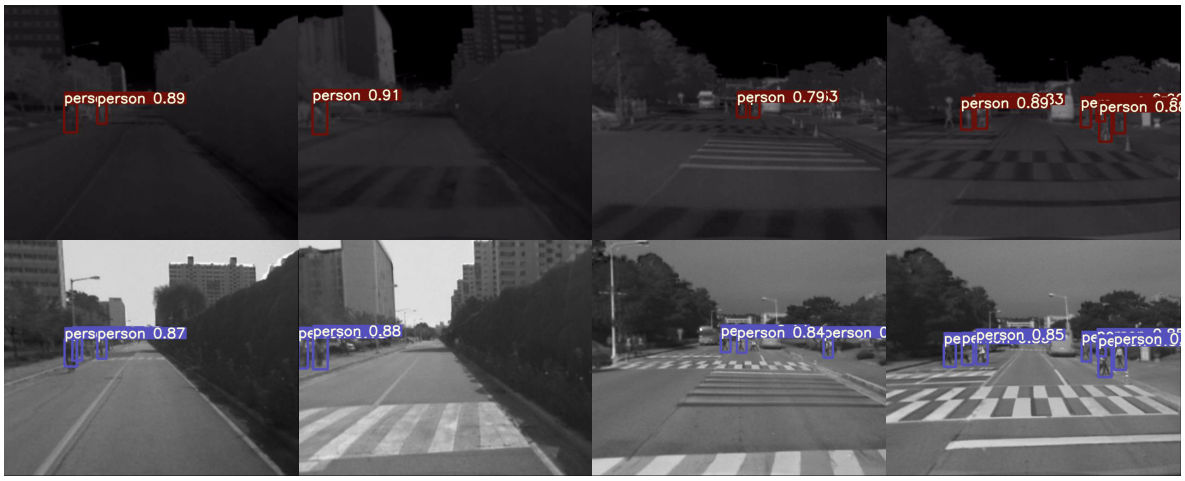


FIGURE 11. Some detection results on the infrared and fused KAIST dataset. We use the proposed YOLO-FIRI to test the two kinds of datasets. The first row is the images in the infrared KAIST dataset and the second row is the images in the fused KAIST dataset. Although the YOLO-FIRI achieves better detection results on the infrared KAIST dataset, the fused dataset obtained through data enhancement can further improve the performance with the proposed models.

TABLE 5. Quantitative comparison of FLIR measured by the common metric AP and mAP in percentage with IoU = 0.5. We evaluate the three classes labeled by FLIR and take reported numbers from [48] for most compared methods. Our YOLO-FLIR outperforms the previous works. Bolded numbers mark the best results.

Method	Person/%	Bicycle/%	Car/%	mAP/%
Faster R-CNN [40]	39.6	54.7	67.6	53.9
MMTOD-CG [40]	50.3	63.3	70.6	61.4
RefineDet [63]	77.2	57.2	84.5	72.9
TermalDet [48]	78.2	60.0	85.5	74.6
YOLO-FIR	85.2	70.7	84.3	80.1
YOLO-FIRI	**85.8**	**85.3**	**90.6**	**83.5**

As shown in Table 5, we compare the value of AP for each class and the mAP of the different detectors, which are presented with percentage. In [40], when only Faster R-CNN was used as the detector, the mAP was 53.9%. However, the MMTOD-CG used the Faster R-CNN as a baseline and combined a pretrained detector, which increased the mAP by 7%. TermalDet [48] was proposed based on RefineDet [63], which took into account the features of each layer as the final detection, and the accuracy increased to 74.6%. The proposed methods YOLO-FIR and YOLO-FIRI are based on the state-of-the-art detector YOLOv5 in this paper. Under the premise of ensuring speed, the accuracy is further improved with the full use of features in the shallow layers, the attention of the important feature, and the improved detection head. We can observe that the values of AP and mAP all outperformed those in previous work, and our YOLO-FIRI further reaches 83.5%. In other words, the region-free framework YOLO-FIRI can learn more features from infrared images and the situation of false detection and long-distance missed detection has improved.

V. CONCLUSION

To overcome the drawbacks of infrared image object detection, in this paper, we propose a one-stage region-free object detector YOLO-FIR for infrared images, which is based on the YOLOv5 and the application for infrared images. Combining the features of infrared images, we further propose an improved YOLO-FIRI based on YOLO-FIR. Precisely, we extend and iterate the shallow CSP module of the feature extraction network and combine an improved attention module to the residual blocks to maximize the use of shallow features, forcing the network to learn the robust and distinguishable features. Additionally, the network detection head is improved, multiscale object detection layers are added, and the detection accuracy of infrared small objects is improved. Compared with YOLOv4, YOLO-FIRI has made a qualitative leap in various indicators. The mAP of YOLO-FIRI is increased by approximately 37% on the infrared images of KAIST, the detection time is reduced by approximately 62%, the network parameters are reduced by more than 89%, and the weight size is reduced by more than 93%. Compared with YOLO-FIR, the mAP of YOLO-FIRI reaches 98.3% and increases approximately 13% on KAIST. The AP for the bicycle class of YOLO-FIRI on FLIR also reaches 85%, which is an increase of 15%. Our proposed model's state-of-the-art performance can be attributed to the combination of learned shallower features and attention features, which allows our model to detect infrared objects based on their low resolution and unclear features. Because the KAIST dataset provides well-aligned visible and infrared images, we prove that data enhancement can be realized to further improve the detection accuracy of infrared images through the use of the convolutional neural network in image preprocessing to fuse visible and infrared images.

In this paper, we mainly focus on the single infrared images which are still and chaotic. It would be interesting to use the infrared video to realize the object detection because the video sequences have a strongly correlation between the front and rear frames. Therefore, the detection performance in infrared video object detection will be better.

REFERENCES

[1] S. G. Kandlikar, I. Perez-Raya, P. A. Raghupathi, J.-L. Gonzalez-Hernandez, D. Dabydeen, L. Medeiros, and P. Phatak, "Infrared imaging technology for breast cancer detection—Current status, protocols and new directions," *Int. J. Heat Mass Transf.*, vol. 108, pp. 2303–2320, May 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0017931016336031

[2] H. Zhang, C. Luo, Q. Wang, M. Kitchin, A. Parmley, J. Monge-Alvarez, and P. C. Higuera, "A novel infrared video surveillance system using deep learning based techniques," *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 26657–26676, Oct. 2018, doi: 10.1007/s11042-018-5883-y.

[3] D. Cazzato, C. Cimarelli, J. L. Sanchez-Lopez, H. Voos, and M. Leo, "A survey of computer vision methods for 2D object detection from unmanned aerial vehicles," *J. Imag.*, vol. 6, no. 8, p. 78, Aug. 2020. [Online]. Available: https://www.mdpi.com/2313-433X/6/8/78

[4] J. Park, J. Chen, Y. K. Cho, D. Y. Kang, and B. J. Son, "CNN-based person detection using infrared images for night-time intrusion warning systems," *Sensors*, vol. 20, no. 1, p. 34, Dec. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/20/1/34

[5] K. Piniarski and P. Pawlowski, "Efficient pedestrian detection with enhanced object segmentation in far IR night vision," in *Proc. Signal Process., Algorithms, Archit., Arrangements, Appl. (SPA)*, Sep. 2017, pp. 160–165.

[6] S. Li, C. Wang, and H. Huang, "Infrared imaging guidance missile's target recognition simulation based on air-to-air combat," in *Optical Sensing and Imaging Technologies and Applications*, vol. 10846, M. Guina, H. Gong, J. Lu, and D. Liu, Eds. Bellingham, WA, USA: SPIE, 2018, pp. 768–780, doi: 10.1117/12.2505607.

[7] K. Zhao and X. Kong, "Background noise suppression in small targets infrared images and its method discussion," *Opt. Optoelectron. Technol.*, vol. 2, pp. 9–12, Oct. 2004.

[8] T. S. Anju and N. R. N. Raj, "Shearlet transform based image denoising using histogram thresholding," in *Proc. Int. Conf. Commun. Syst. Netw. (ComNet)*, Jul. 2016, pp. 162–166.

[9] P. Jiao, "Research on image classification and retrieval method based on deep learning and sparse representation," M.S. thesis, Xi'an Univ. Technol., Xi'an, China, 2019.

[10] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, Jul. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220301430

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[12] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[15] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.

[18] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: http://arxiv.org/abs/1701.06659

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[20] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[21] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[22] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*. [Online]. Available: http://arxiv.org/abs/2004.10934

[23] G. Jocher, A. Stoken, and J. Borovec, "Ultralytics/yolov5: V4. 0-Nn. SiLU () activations weights & biases logging PyTorch hub integration," Zenodo, Tech. Rep., Jan. 2021. [Online]. Available: https://zenodo.org/record/4418161, doi: 10.5281/zenodo.4418161.

[24] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2020, pp. 390–391.

[25] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conferecnce Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[26] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "KAIST multi-spectral day/night data set for autonomous and assisted driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 934–948, Mar. 2018.

[27] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2018.

[28] (2018). *FREE FLIR Thermal Dataset for Algorithm Training*. [Online]. Available: https://www.flir.in/oem/adas/adas-dataset-form

[29] L. Fang, X. Wang, and Y. Wan, "Adaptable active contour model with applications to infrared ship target segmentation," *J. Electron. Imag.*, vol. 25, no. 4, pp. 1–10, 2016, doi: 10.1117/1.JEI.25.4.041010.

[30] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004, doi: 10.1023/B:VISI.0000013087.49260.fb.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[32] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[33] P.-H. Chen, C.-J. Lin, and B. Schölkopf, "A tutorial on υ-support vector machines," *Appl. Stochastic Models Bus. Ind.*, vol. 21, no. 2, pp. 111–136, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.537

[34] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. ICML*, 1996, pp. 148–156.

[35] S. M. Z. S. Z. Ariffin, N. Jamil, and P. N. M. A. Rahman, "DIAST variability illuminated thermal and visible ear images datasets," in *Proc. Signal Process., Algorithms, Archit., Arrangements, Appl. (SPA)*, Sep. 2016, pp. 191–195.

[36] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks and Machine Learning*, T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds. Berlin, Germany: Springer, 2011, pp. 52–59.

[37] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," in *Advances in Neural Information Processing Systems*, vol. 29, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/325995af77a0e8b06d1204a171010b3a-Paper.pdf

[38] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14454–14463.

[39] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[40] C. Devaguptapu, N. Akolekar, M. M. Sharma, and V. N. Balasubramanian, "Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[41] M. Kristo, M. Ivasic-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using Yolo," *IEEE Access*, vol. 8, pp. 125459–125476, 2020.

[42] M. Li, T. Zhang, and W. Cui, "Research of infrared small pedestrian target detection based on YOLOv3," *Infr. Technoiogy*, vol. 42, no. 2, pp. 176–181, Feb. 2020.

[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[44] Y. Li, S. Li, H. Du, L. Chen, D. Zhang, and Y. Li, "YOLO-ACN: Focusing on small target and occluded object detection," *IEEE Access*, vol. 8, pp. 227288–227303, 2020.

[45] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, vol. 34, no. 7, pp. 12993–13000. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6999

[46] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," 2020, *arXiv:2005.03572*. [Online]. Available: http://arxiv.org/abs/2005.03572

[47] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5561–5569.

[48] Y. Cao, T. Zhou, X. Zhu, and Y. Su, "Every feature counts: An improved one-stage detector in thermal imagery," in *Proc. IEEE 5th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2019, pp. 1965–1969.

[49] X. Dai, X. Yuan, and X. Wei, "TIRNet: Object detection in thermal infrared images for autonomous driving," *Int. J. Speech Technol.*, vol. 51, no. 3, pp. 1244–1261, Mar. 2021, doi: 10.1007/s10489-020-01882-2.

[50] X. Song, S. Gao, and C. Chen, "A multispectral feature fusion network for robust pedestrian detection," *Alexandria Eng. J.*, vol. 60, no. 1, pp. 73–85, Feb. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016820302507

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[52] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5449–5457.

[53] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. Int. Conf. Pattern Recognit.*, vol. 3. Aug. 2006, pp. 850–855.

[54] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2015, *arXiv:1502.03044*. [Online]. Available: http://arxiv.org/abs/1502.03044

[55] A. L. Maas, A. Hannun, and A. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30, no. 1, p. 3.

[56] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," 2018, *arXiv:1808.04818*. [Online]. Available: http://arxiv.org/abs/1808.04818

[57] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," 2016, *arXiv:1611.02644*. [Online]. Available: http://arxiv.org/abs/1611.02644

[58] D. M. Hawkins, "The problem of overfitting," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 1, pp. 1–12, 2004, doi: 10.1021/ci0342472.

[59] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*. [Online]. Available: http://arxiv.org/abs/1708.04552

[60] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.

[61] C. Gong, D. Wang, M. Li, V. Chandra, and Q. Liu, "KeepAugment: A simple information-preserving data augmentation approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1055–1064.

[62] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals for accurate object class detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 424–432.

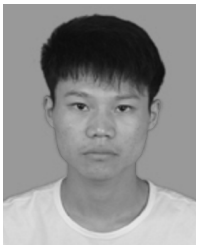
[63] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.

SHASHA LI was born in Luoyang, Henan, China, in 1996. She received the B.S. degree in electronic science and technology from the Henan Institute of Engineering, Zhengzhou, China, in 2019. She is currently pursuing the M.S. degree in optical engineering with Henan University, Kaifeng, China. Her research interests include computer vision, image processing, and object detection.

YONGJUN LI was born in Kaifeng, Henan, China, in 1977. He received the M.S. degree in circuit and systems from Guangxi Normal University, in 2005, and the Ph.D. degree in communication and information systems from Xidian University, in 2017.

From 2005 to 2008, he was a Lecturer with the School of Physics and Electronics, Henan University, where he has been an Assistant Professor, since 2012. He presided over and participated in a number of national natural science and technology funds, and took part in the Shaanxi Province Key Science and Technology Innovation Team Project. He has applied for six national invention patents, and published more than 30 papers in *Multimedia Tools and Applications*, *Optical Engineering*, *Journal of Electronic Imaging*, *Journal of Zhengzhou University*, and *Journal of Henan University*. His research interests include image processing and artificial intelligence.

YAO LI was born in Henan, China, in 1998. He received the B.S. degree from Hainan University, China, in 2020. He is currently pursuing the M.S. degree with Henan University. His current research interests include computer vision, image processing, and deep learning.



MENGJUN LI was born in Shangqiu, Henan, China, in 1995. He is currently pursuing the M.S. degree in optical engineering with Henan University, Kaifeng, China. His research interests include computer vision and image super-resolution.

XIAORONG XU was born in Weinan, Shanxi, China, in 1979. She received the M.S. degree in computer software and theory from Guangxi Normal University, Guilin, China, in 2005. Since 2005, she has been a Teacher with the Hunan University of Arts and Science. She has published many academic papers. Her research interests include image processing and pattern recognition.