IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 18, 2022, accepted May 1, 2022, date of publication May 3, 2022, date of current version May 27, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3172504

Lightweight Spatial Sliced-Concatenate-Multireceptive-Field Enhance and Joint Channel Attention Mechanism for Infrared Object Detection

ZHIHENG PAN, (Graduate Student Member, IEEE), LIUCHAO XU, CHUANDONG LIANG, KUI PAN, MI ZHAO, AND MIN LU

College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832000, China

Corresponding author: Min Lu (lm_shz@163.com)

This work was supported in part by the Launch Project of High-level Talent Scientific Research of Shihezi University, and in part by the Shihezi University International Science and Technology Cooperation Promotion Plan under Grant RCZK202005 and Grant GJHZ202108.

ABSTRACT Infrared object detection has high application value in the field of remote sensing due to its anti-interference ability and long detection distance. However, infrared images suffer from many disadvantages such as poor fine-grained information, low resolution and contrast, which makes infrared object detection methods have rather poor performance while utilizing conventional object detection methods. Two novel lightweight attention mechanisms were proposed in this study to solve the problem. Sliced concatenate and multi receptive-field spatial group-wise enhance (SCMR-SGE) module, utilizing grouping feature operation, enhances the sub-features by generating attention factors at each location in each semantic group and suppresses irrelevant information. Joint attention module is used to selectively enhance or inhibit channel information through attention factors generated by three different pooling layers. Unlike the previous work, each module was used only once, and was embed into two modules into feature pyramid network (FPN) instead of backbone network. The mAP50 of our method based on YOLOv5m alone reached 82.7%, which was the best result on the original FLIR dataset which didn't process the imbalanced sample problem. At the same time, the detection speed can still be maintained at around 60 FPS on single GPU. Our experiments demonstrated that our lightweight attention mechanisms have better performance than mainstream ones, and the method of embedding our attention mechanisms into the CNN is effective and universal.

INDEX TERMS Infrared object detection, lightweight attention mechanism, convolutional neural network.

I. INTRODUCTION

Infrared imaging technology can convert the invisible object surface temperature into a visible thermal image representing the object surface temperature distribution. Infrared imaging technology uses the infrared spectrum of the object, which has better anti-interference performance and larger operating range than the visible spectrum in bad weather [1]. These characteristics make it have high application value in the fields of security system, fire alarm, automatic driving and so on [2]–[6].

For many years, object detection for infrared images has always been a challenging task. Compared with the optical image, the infrared image is obtained by ''measuring'' the heat radiated by the object, which makes infrared image suffer from low resolution, low contrast and low signal-to-noise ratio (SNR) [2]. Therefore, the object detection task for the infrared image is more difficult than that for the visible image. In daily life, the infrared characteristics of most objects are not obvious, and the boundary between them is also fuzzy, which can cause a lot of trouble to the task when the target area has complex background.

In the past few years, benefiting from the rapid development of Convolutional Neural Network (CNN) and its strong performance, infrared target detection has ushered in new development opportunities. With a large amounts of data training, CNNs can well fit a model for specific tasks. Influential methods such as SSD [7], YOLO [8], Retina-Net [9],

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Faster R-CNN [10], ResNet [11] and so on, have proved their superior performance. However, these methods, with little consideration for the infrared object detection, are aimed at the conventional object detection task. Optical images, benefiting from their relatively high resolution, contain fine-grained detail which is essential to object detection. On the other side, the lack of information makes it tough to extract inherent features in infrared images for object detection task.

Many researchers have investigated approaches to improve the feature extraction capability for infrared targets. In [12], Mcintosh *et al.* proposed Target to Clutter Ratio (TCR) Metric to derive the optimum eigenvectors, which can simultaneously represent targets and discriminate them from clutter. In [13], Nasrabadi *et al.* introduced a powerful method using a combination of a two-dimensional wavelet transform, which decomposes the images into uniform subbands, and an efficient algorithm called the Reed Xiaoli (RX), which can detect multi-variate anomaly. In [14], Liu *et al.* utilize eigenvectors obtained by computing every location in the image through a target map function, where large values indicate the target position. These methods have improved the detection performance on infrared targets. However, they involve a large number of mathematical operations with high complexity, which means they are easy to encounter operations not integrated in the deep learning framework. This makes them have difficulty to give full play to the acceleration function of the GPU, resulting in reduced training and detection speed and poor generality. CNN based methods such as [15]–[17] trying to use paired visible images as complimentary. To alleviate the issue that infrared images lack of fine-grained features, a series of networks with fusion architectures are built to enhance original infrared images with detailed visual characteristics. But these methods still have some limitations, such as the need for tons of precisely paired color-thermal images increases the operational cost and are hard to collect.

On the other hand, inspired by human visual attention mechanism, researchers introduce the attention mechanisms into CNNs to improve the performance of method with low cost. By roughly scanning the global area of image, humans can easily determine the target areas needed to be focused on. Similarly, the attention mechanism assigns weights to the information processed by the method, and invests more attention resources in high weight areas to get details from them while suppressing other irrelevant areas. A local cross-channel interaction strategy was introduced in ECAnet [18] without reducing the dimension to learn channel attention. SEnet [19] uses global average pooling to extract the global features of each feature map, and then obtains the weight of each channel through MLP and sigmoid function, which makes the key channels obtain higher weight and suppresses the irrelevant channels. Based on SEnet and Inception [20], SKnet [21] can adaptively select the size of receptive field according to the input of the network, so as to distinguish the importance between different channels. CBAM [22] and BAM [23] utilize both global average pooling and maximum pooling on the basis of SEnet, and introduce spatial attention mechanism to make the method obtain the weight from channel and space. [24] introduces transformer [25] block to attention mechanism for infrared object detection task. However, most of existing work directly embed the attention mechanism into all residual structures in backbone of CNNs. In fact, how to embed the attention mechanism into the CNN and which layers of the CNNs are embedded also have a certain impact on the performance of the method.

Generally speaking, researchers nowadays focus on two technology routes: either improving the feature extraction ability of their methods or enriching the fine-grained features in infrared images. In this article, we propose a novel spatial attention module and a novel channel attention module to improve feature extraction capabilities. Both two modules are lightweight and can be easily embedded into existing methods. In particular, we find suitable position for our modules and minimize the number of them we use.

To sum up, the contributions of this article are as follows:

1) A novel spatial attention mechanism called SCMR-SGE module is proposed to improve detection performance of our method. SCMR-SGE can effectively highlight multi interested targets in multi areas with diverse high order semantics and suppress background noise. Meanwhile, it is still a lightweight attention mechanism attaching almost no extra parameters and calculations.
2) We propose a novel channel attention mechanism called joint attention (JA) module utilizing three different pooling layers to selectively enhance or suppress specific channels. Specifically, we directly compare the important factors obtained from channel attention mechanism which utilize three different pooling layers to determine the final weight, which makes the allocation of attention resources more reasonable.
3) Unlike most of other methods that use multiple attention modules, we only use each module once and embed them in FPN structure rather than the backbone network to achieve better performance. The experiment results show that our arrangement is effective.

The rest of this article is arranged as follows: In Section 2, we introduce the method and structure in detail. Experiment settings are presented in Section 3. We make full explanation of our experiments setting and results, along with discussions based on our comparative experiments in Section 4. Finally, we summarize our work in Section 5.

II. METHODS

A. SLICED CONCATENATE AND MULTI-RECEPTIVE FIELD SGE

The first-generation source using the similarity between local and global as an attention mask was proposed in [26]. Authors believe that features generated by CNNs is composed of many sub-features, which can usually be distributed in grouped form in the feature of each layer. By scaling the feature vectors over all the locations in each group, SGE module can

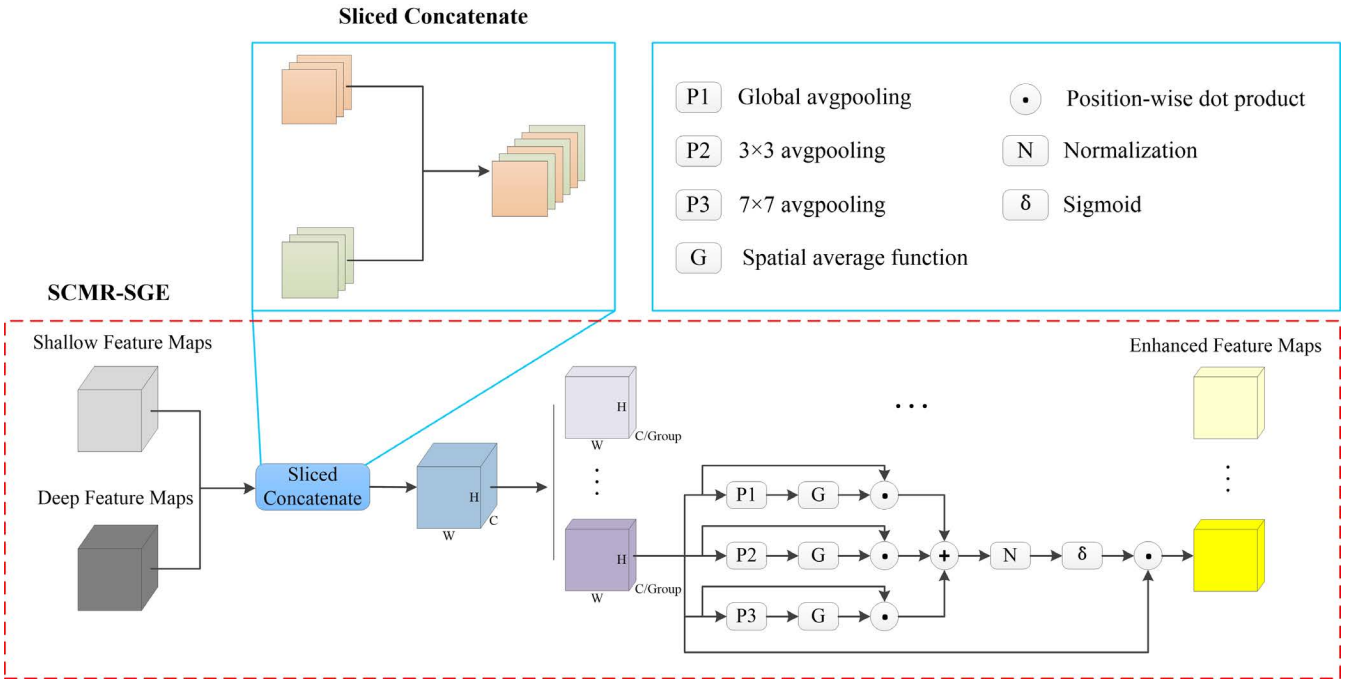

FIGURE 1. Structure of proposed SCMR-SGE Module. This attention module can obtain semantically enhanced feature maps.

spatially suppress the possible noise and highlight the correct semantic feature regions with minimal additional computational burden. These feature vectors are determined only by the similarity between global and local features within each group. However, each group in SGE module is only composed of several adjacent feature maps, thereby reducing the information flow and information representation capabilities between groups. To solve this problem, in [27], author proposed the SSE module. The SSE module introduced the idea of channel shuffle in Shufflenet [28], which shuffled feature maps originally from different groups into new groups to strengthen the information flow between groups. However, such a way of information exchange can only make information flow between adjacent groups, the input and output channels from distant groups are still uncorrelated. Furthermore, SGE module employs global average pooling to approximate the semantic vector that one group learns to represent. Obviously, this operation can rapidly gather global information in feature maps, which is efficient in image classification tasks because they only need to focus on the most interested area in feature maps. However, this may cause problems in the field of object detection tasks, because they may have multiple interested areas in an image. In this case, the global average pooling operation cannot well separate all interested areas, resulting in the loss of information in some areas causing missed detection.

Mainly inspired by the above research, we proposed SCMR-SGE module. The FPN structure in YOLOv5 combines the deep feature maps with the shallow ones for feature fusion. Feature maps generated by deep layers in CNNs usually contains rich semantic information but their resolution is relatively low. On the contrary, those high-resolution feature maps generated by shallow layers have less semantic information but rich in spatial information [29]. In order to make better use of the feature fusion function of FPN, the idea of sliced concatenate is introduced in our module. The overall structure of proposed SCMR-SGE module is shown in **Figure 1**. First, feature maps generated by relative shallow layers in CNNs concatenate those ones generated by deep layers through sliced concatenate operation. Then, SCMR-SGE divides feature maps into G groups along the channel dimension. After that, as mentioned above, to approximate a more appropriate semantic vector that this group learns to represent, SCMR-SGE adopts multi average pooling modes to acquire global and local statistical features respectively. Finally, more relevant features between channels are aggregated, and an attention factor is generated at each spatial location within each group to learn more advanced semantic information through spatial enhance operation [27].

In [30], sliced concatenate function is achieved through loop structure, which does not increase floating point operations (FLOPs). However, [31] points out that it is insufficient to use FLOPs as the only metric for computation complexity. The original design in CASEnet resulted in a large increase in network memory accesses cost (MAC) and reducing degree of parallelism. Compared with the original method, this operation almost doubled the time cost in training stage and the latency in detection stage in our experiment. Therefore, we modified the original operation to improve

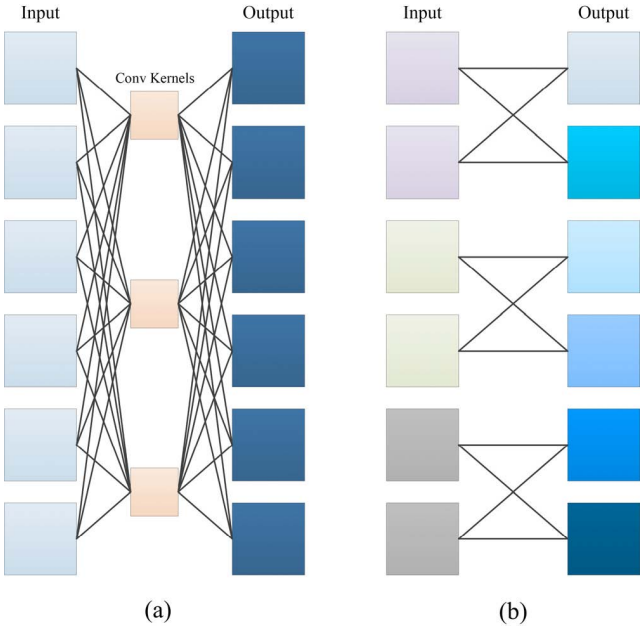

FIGURE 2. Illustration of standard convolution and group feature. (a) Standard convolution; (b) Convolutional layers with three groups.

operational efficiency. Modified sliced concatenate operation can be completed by the following steps: first, we generate an all-zero tensor with the same shape as the tensors needed to be sliced concatenated. Then, the first tensor to be operated concatenate this all-zero tensor, which is also used for concatenating the second tensor to be operated. Next, reshape these two tensors channel dimension into (C, 2), where each tensor has 2C channels after concatenating, and transpose into (2, C). Finally, reshape them into 2C channels and add these two tensors. Sliced concatenate operation ensures each group contains feature maps generated by both shallow layers and deep layers, which carry strong position information and semantic information, respectively.

Grouping feature can significantly reduce the amounts of calculations [28]. It can be seen from **Figure 2** that each group learns different features by fewer parameters compared with standard convolution. Suppose the size of input feature map is $C \times H \times W$ and the number of convolution kernels is N, the total parameter amount of grouping features is $N \times (C/G) \times H \times W$ under the circumstance of channels are divided into G groups. However, the total parameter amount of standard convolution is $N \times C \times H \times W$. It means that the total parameter amount is reduced to the original (1/G).

It is difficult for a CNN to obtain favorable feature distributions in deep networks. To solve this problem, semantic features in critical regions are enhanced through learning the overall information of each entire group. In particular, local averaging function employs two different sizes of pooling to refine receptive field. The procedure for spatial enhancement is as follows: global and local statistical features g are obtained through global averaging function and local averaging function to approximate the semantic vectors that this group learns to represent

$$g_{ij} = \frac{1}{k^2} \sum_{i=i-s}^{i+s} \sum_{j=j-s}^{j+s} x_{ij} \tag{1}$$

where k is pooling size, $s = (k-1)/2$, $x_{ij} \in R^{C/G}$. In this article, we choose 3, 7 and global as pooling size, respectively. In particular, when the pooling size is set to global, the operation will be simplified to global average pooling.

Then, a corresponding importance coefficient c_i is generated by a simple dot product for each feature, which measures the similarity between the semantic feature g and the local feature x.

$$c_{ij} = g_{ij} \cdot x_{ij} \tag{2}$$

Next, c is normalized over the space in order to prevent coefficients deviation between various samples.

$$\mu_c = \frac{1}{m} \sum_{i=1,j=1}^{m} c_{ij} \tag{3}$$

$$\lambda_c = \frac{1}{m} \sum_{i=1,j=1}^{m} (c_{ij} - \mu_c)^2 \tag{4}$$

$$\hat{c}_i = \frac{c_{ij} - \mu_c}{\lambda_c + \varepsilon} \tag{5}$$

where $m = H \times W$, ε is a constant added for numerical stability. To make sure that the identity transform can be represented through normalization inserted in the network, a pair of parameters γ, β are introduced for each coefficient $\hat{c}_i$ to scale and shift the normalized result:

$$a_i = \gamma \hat{c}_i + \beta \tag{6}$$

Finally, importance coefficient generated through sigmoid function is employed to scale the original x_{ij} spatially to obtain the enhanced feature $\hat{x}_{ij}$.

$$\hat{x}_{ij} = x_{ij} \cdot \sigma(a_i) \tag{7}$$

where σ represents sigmoid function. All the enhanced feature maps constitute the final output, which significantly eliminate the interference of background noise in infrared images and highlight the interested areas. This operation can greatly improve the performance of initial YOLOv5s, yet the same, just like SE or CBAM module, SCMR-SGE module can be embedded in existing mainstream network structures and the additional computational burden is negligible.

In short, SCMR-SGE module utilizes attention factors guided by the similarities between the global and multi receptive field local information, which can simultaneously integrate attention in channels and spaces, to improve feature extraction capabilities. In particular, sliced concatenate operation promote the information exchange between each group, which can make the result more robust. SCMR-SGE module enables each feature group to autonomously enhance high order semantic features and suppress possible noise with negligible additional parameters.

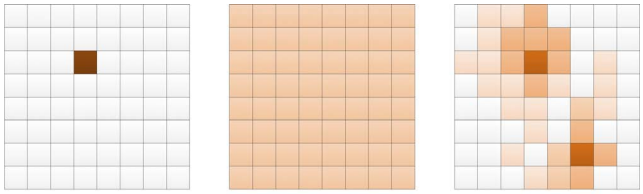

FIGURE 3. Simplified illustration of different pooling layers. (a) Maximum Pooling; (b) Average pooling; (c) Log-sum-exp pooling. The shade of color represents the weight of this pixel.

B. JOINT ATTENTION MODULE

Squeeze-and-Excitation (SE) module, which is the first attention module take into consideration the relationship between channels, was proposed to improve the quality of representations by explicitly modelling the inter dependencies between the channels [19]. "Feature recalibration" strategy is adopted in SE module. Specifically, squeeze operation embeds global information and excitation operation implements adaptive recalibration. The importance of each feature channel is automatically acquired through this procedure, then the useful features are promoted according to this important factor and the features that are not useful for the current task are suppressed. However, Woo *et al.* pointed out that features generated by global average pooling are suboptimal features in order to infer fine channel attention, and they suggest to use max-pooled features as well [25].

As mentioned above, the pooling layers play an important role in determining what information should be passed down by network. **Figure 3** is a toy example of different pooling layers. Suppose that two photos are taken at the same scene, but someone set a bonfire in one of them. In this case, if our target is not this bonfire, maximum pooling will tend to express the feature of this bonfire and the result of average pooling will also be shifted. Representing the temperature distribution of the object is a central feature of infrared images and distinguishes it from visible light images, which means objects of the same kind tend to share similar activation value. Thus, we need a smooth pooling approach that allows similar pixels to share similar weights. We thereby utilize the Log-Sum-Exp (LSE) pooling proposed in [32]. The LSE pooling function is defined as

$$x = \log[\frac{1}{S} \cdot \sum_{(i,j)\in S} exp(r \cdot x_{ij})] \tag{8}$$

where x_{ij} is the activation value of each pixel in the pooling region S, and $S = s \times s$ is the total number of pixels in S. r is a hyperparameter which controls how smooth one wants the approximation to be. Infinite value implies the effect will similar to the maximum pooling, zero value will have an effect similar to the average pooling. The advantage of this aggregation is by controlling the value of r, pixels have similar scores will share a similar weight in the training procedure, which controls the level of similarity. On account of suffering from overflow and underflow problems, the LSE function is modified as

$$x_p = x_{\max} + \frac{1}{r} \cdot log[\frac{1}{S} \cdot \sum_{(i,j)\in S} exp(r(x_{ij} - x_{max}))] \tag{9}$$

where $x_{max} = max\{|x_{ij}|, (i, j) \in S\}$.

[22] argues that different pooling layers can gather other important clues about distinctive object features to refine channel-wise attention. Thus, simultaneously using different pooling layers can greatly improve representation capacity of networks rather than using each of them independently. Based on this conclusion, A novel Joint Attention (JA) Module, which employs three different pooling layers in **Figure 3**. simultaneously, is proposed mainly inspired by the researches mentioned above. The overall structure of JAM is shown in **Figure 4**. Our idea is simple: since different pooling approaches extract different feature factors, then it might be better to compare the extracted feature factors directly. Channels will be enhanced if anyone of the three approaches gives a high importance factor, otherwise, suppress this channel instead. In particular, we double our operation with the aid of the up-sampling function in FPN network to make the results more robust.

The operation of proposed JA Module can be described as follow: First, three different descriptors are generated by three different pooling layers, respectively. Then, each of these three descriptors is forwarded to a shared network, which is composed of multi-layer perceptron (MLP) with one hidden layer, to generate three important factors. Finally, the three vectors are compared to selectively enhance or suppress them. Specifically, we take the maximum value if any value of the three vectors in the same position is greater than 0.5, conversely, if all three values in the same position are less than 0.5, their minimum value is taken. In short, the proposed channel attention is computed as

$$A(F) = ES\{\sigma(MLP(avgpool(F), \sigma(MLP(maxpool(F), \sigma(MLP(LSEpool(F), \}\} \tag{10}$$

where σ denotes the sigmoid function, MLP weights are shared for all the inputs, ES denotes the enhance and suppress operation. Then, we employ up-sampling function and repeat the ES operation to eliminate the offset caused by the linear interpolation calculation in the up-sampling function as much as possible:

$$\hat{A}(F) = ES\{A(F), A[upsample(F)]\} \tag{11}$$

important factors can be further refined after up-sample stage, which makes the JA module more robust.

To sum up, JA module utilizes three different pooling layers to determine enhance or suppress channel attention factors. This will allow JA module make full use of the channel factors representing fine, prominent and smoothed features respectively to improve representation capability. In addition, JA module also take the advantage of the up-sample operation to recalibrate the final attention factors, which contributes to its compelling effectiveness in practice.

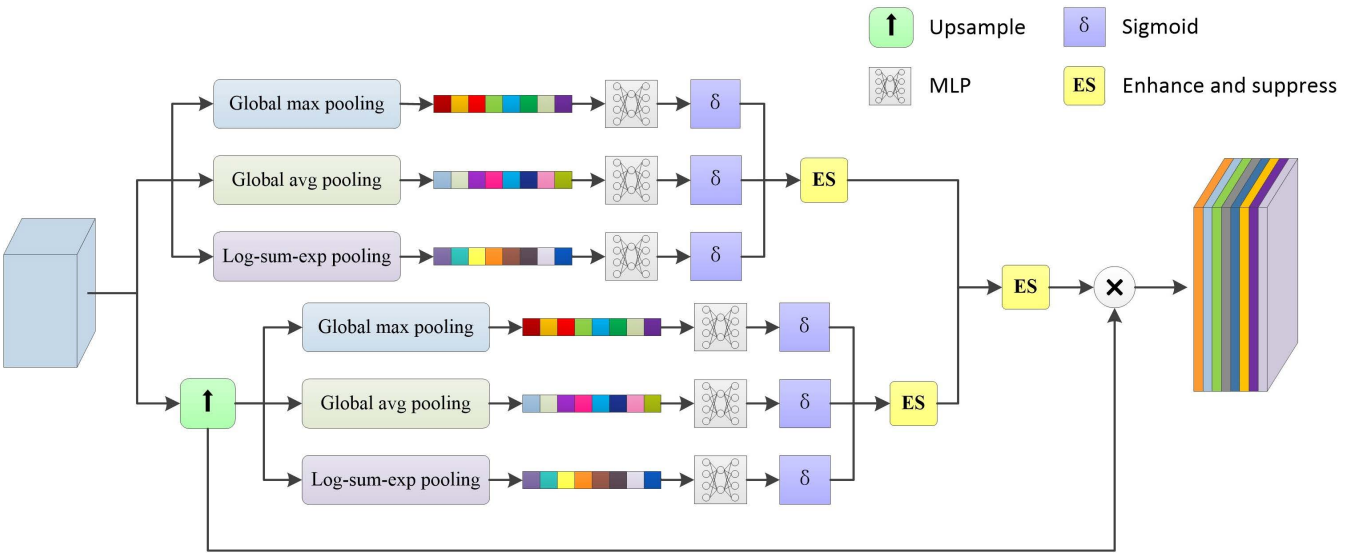

FIGURE 4. Structure of proposed Joint Attention (JA) Module. This module can enhance/suppress and obtain up-sampled feature maps.

C. POSITION OF EMBEDDING

YOLOv5 [33] adopts the modular design idea and has four versions: YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. The only difference between the four versions is the number of CSP modules stacked. Stacking more CSP modules can obtain better feature extraction capability, which is conducive to our comparative experiments. By embedding our modules at the same position of FPN network in different version of YOLOv5, we can determine whether our method is universal according to the experimental results. The overall structure of YOLOv5s is illustrated in **Figure 5.** YOLOv5m, YOLOv5l and YOLOv5x reuse 2, 3 and 4 groups of CSP modules at the position of CSP module in YOLOv5s respectively.

Since we want to make the method as efficient as possible, although the modules proposed above are lightweight enough yet we still want to employ these models as less as of possible. Thus, we decide to focus on the FPN network, which is designed for feature fusion, to refine the features extracted by backbone network. As mentioned above, these two modules, SCMR-SGE and JAM, focus on 'where' and 'what' need to be paid attention respectively, should be embedded in different layers at FPN network. Considering that SCMR-SGE module needs feature maps containing relatively more semantic information, concatenate module marked by blue rectangle accepts feature maps generated by the last CSP module in backbone network and the final outputs of backbone network, which makes it suitable for SCMR-SGE module. On the other side, replacing up-sampling function with JAM marked by red rectangle can refine the feature maps processed by SCMR-SGE module without interfering the coarse features extracted by shallower layers. Benefiting from YOLOv5s adopts FPN [26] and PANet [34] for the neck network at the same time, feature maps processed in such way can carry stronger semantic feature and flow to each prediction head. This arrangement, focusing on feature fusion stage instead of feature extraction stage, is different from the common practice of embedding the attention mechanisms into the residual structure, yet still can benefit from the improvement of feature extraction capability. We will further discuss this point in Section 4.

III. EXPERIMENTS SETTINGS

We conducted a series of comparative experiments to verify the advancement and robustness of the proposed methods. This section will also introduce our dataset, experimental settings and evaluation metrics.

A. DATASET INTRODUCTION

FLIR-ADAS dataset [35] was collected from a vehicle camera view on the streets and roads of Santa Barbara, California. The dataset, taken at daytime (60%) and night (40%) under clearly cloudy weather conditions, contains 8862 images for training and 1366 images for validation. Labeled objects in the dataset include cars, persons, bicycles and dogs. It needs to be pointed out that the dataset lack of dog labels, which may cause the trained model have poor generalization ability and suffer from overfitting problem. Thus, labels of dogs are ignored in our experiment. In particular, even after ignoring the dog label, the number of people, cars and bicycles in the sample is still uneven. Data enhancement strategy for bicycle class used in some researches is not employed in our method. Some images in FLIR dataset are shown in **Figure 6**.

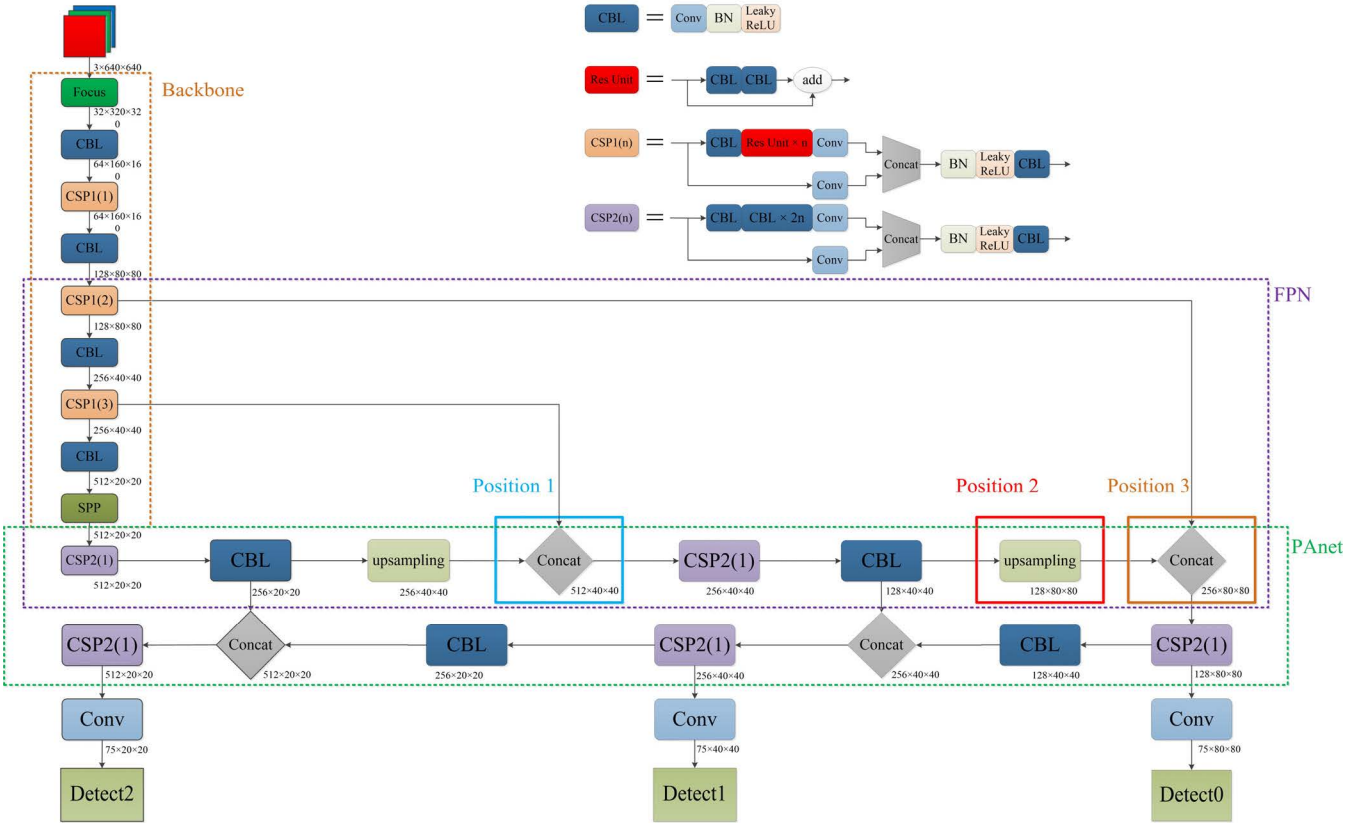

FIGURE 5. Structure of YOLOv5s. Modules marked by blue and red rectangles at position 1 and position 2 are replaced by SCMR-SGE Module and JA Module, respectively.

FIGURE 6. Display of some images in FLIR dataset.

B. DATASET INTRODUCTION

All experiments were implemented on a PC with Intel(R) Core(R) i7-8700K CPU, NVIDIA RTX 3060(12GB LHR) GPU, CUDA 11.1, CUDNN 9.1, and the operating system was Win10. The deep learning framework was Pytorch 1.9.0. The optimizer used stochastic gradient descent (SGD) and the learning rate decay strategy was cosine. The initial learning rate was 0.01, finally decay at rate of 0.001, batch size was 48, warmup 3 epochs, and epochs for training were optimized to 100. The input image resolution was 640×512.

C. DATASET INTRODUCTION

To evaluate the performance of the proposed methods, mAP50, mAP0.5:0.95 and latency were used as metrics. Average precision (AP) is the area surrounded by the precision-recall curve and coordinate axis. Mean average precision (mAP), which is the average of multiple categories of AP, is the key metric in object detection algorithm. There are three categories in our experiment, so mAP50 can be regarded as the average of AP values of these categories. Generally speaking, the better detection capability, the higher mAP value. The precision and recall are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

where TP (True Positive) is the positive sample of correct detection, and FP (False Positive) is positive sample of error detection. On the contrary, FP and FN are negative samples of correct and error detection, respectively. Whether the label is a positive sample or a negative sample is determined by the

TABLE 1. Comparison with other mainstream attention mechanisms.

Module	Embedding	mAP50	mAP50:95	Para.	Latency
SE	FPN	80.8	42.8	7.02M	8.9ms
SE	backbone	80.9	43.1	7.02M	9.4ms
CBAM	FPN	80.8	43.0	7.05M	9.2ms
CBAM	backbone	80.8	42.9	7.06M	10.1ms
SGE	FPN	80.8	42.7	7.02M	8.4ms
SGE	backbone	80.4	42.6	7.02M	8.9ms
MR-SGE	FPN	81.1	42.7	7.02M	9.1ms
MR-SGE	backbone	80.6	42.4	7.02M	10.7ms
SCMR-SGE	FPN	81.3	42.9	7.02M	9.2ms

IoU. The IoU is defined as follow:

$$IoU = \frac{P \cap T}{P \cup T} \tag{14}$$

where P and T refer to predicted bounding box and ground truth. The threshold of IoU is set to 0.5 in most cases, which means if samples with IoU greater than 0.5 are considered as positive samples. Then, AP is defined as:

$$AP = \int_0^1 p(r)dr \tag{15}$$

where r denotes recall, p denotes precision and p(r) is a function which takes r as parameter. In most cases, mAP50 is the most important metric.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. ABLATION STUDY

1) SCMR-SGE MODULE

In this section, the effectiveness of the proposed spatial enhancement module is verified. When comparing the methods, all parameters are set to the same. After a series of experiments, we choose to divide feature maps into 64 groups in SGE module for best performance. The latency of method is evaluated by all video frames attached in datasets.

We choose SE and CBAM Module to compare with our module since these two modules have been wildly used in recent object detection methods. Original SGE Module is also employed as control study. All modules here are embedded into concatenate structure marked by blue rectangle shown in **Figure 5**, respectively. The performance of each module is shown in **Table 1**.

As seen from the table, compared with SE and CBAM Module, original SGE module has similar performance but lower latency. Benefiting from grouping features idea in SGE module, our modified module can still maintain a relatively low computational burden. Compared with original SGE module, Multi-Receptive-field SGE increases the mAP50 by 0.5% at the cost of increasing 0.7ms latency. What's more, when it comes to SCMR-SGE Module, the redesigned sliced concatenate operation only increases 0.1ms latency in return for 0.2% mAP50 increasing. To verify the increase in performance, we also embed mentioned attention mechanisms in the backbone. It can be seen that mAP of embedding attention mechanisms in FPN structure is almost the same as that of embedding in backbone. However, embedding them in backbone will cause obvious detection speed loss due to multiple use. In particular, SGE and MR-SGE module show rather poor performance when they are embedded in backbone, which proves they are more suitable for processing deep semantic features. The results above indicate SCMR-SGE Module can improve detection performance with a small computational cost, which is essential in real-time object detection task.

TABLE 2. Comparison with other mainstream attention mechanisms.

Module	mAP50	mAP50:95	Para.	Latency
SE	80.7	43.1	7.02M	8.7ms
SE(max+avg pool)	80.9	43.2	7.02M	9.1ms
SE(max+avg+lse pool)	80.5	42.3	7.02M	9.3ms
CBAM	80.2	42.7	7.02M	9.7ms
JA	81.4	43.0	7.02M	9.6ms

2) JA MODULE

In this section, we conducted a series of experiments to verify the effectiveness of our JA module. We still introduce YOLOv5s method equipped with SE and CBAM modules as comparative experiment. All modules here are added to upsample structure marked by red rectangle shown in **Figure 5**, respectively. The performance of each module is shown in **Table 2**.

Three SE modules using three different pooling layers are employed to embed into initial method. All these modules improve the performance of the method. In particular, SE module using average pooling combined with maximum pooling shows better performance, which is consistent with what we mentioned in Section 2. It is also noteworthy that, compared with SE modules, the performance of CBAM Module is rather poor. It is likely that as the neural network gets deeper, channels are getting more and more important, which makes it inappropriate to employ spatial attention modules to filter the deep feature maps. What's more, when SE module adopts more than two different pooling layers, the method of directly adding the pooling results is no longer effective. Compared with all above modules, proposed JA module increase mAP50 to 81.4%, which is higher than any of them. The result further confirms that different pooling layers can gather different salient features to refine channel-wise attention. At the same time, the loss of our method's detection speed is acceptable.

B. INFRARED OBJECT DETECTION IN FLIR-ADAS DATASET

1) EFFECT OF SCMR-SGE AND JA MODULES

We selected several representative detection results to intuitively compare our method with the original YOLOv5s. The confidence threshold for bounding box and IoU in both two

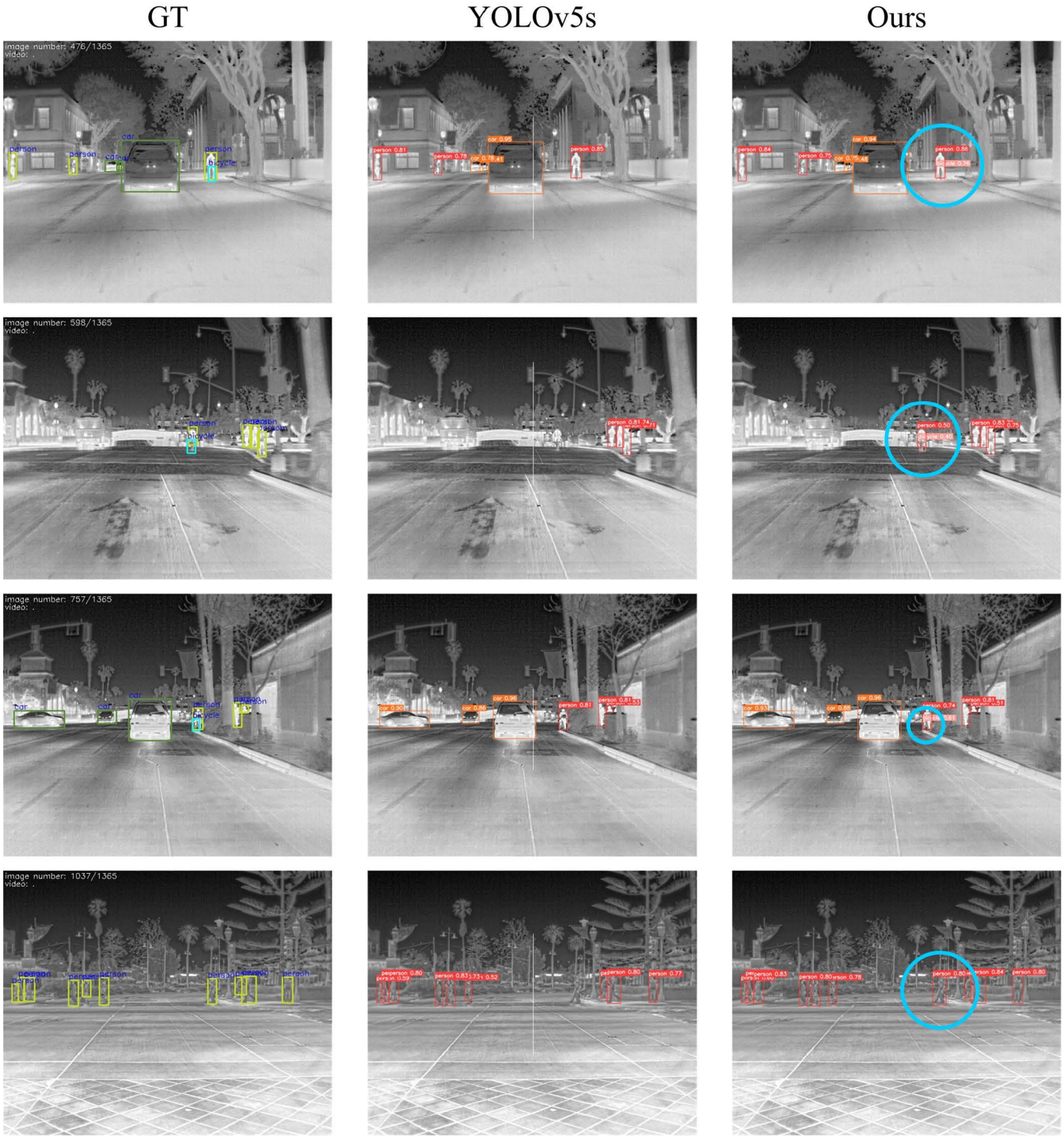

FIGURE 7. The detection results of FLIR dataset conducted by YOLOv5s and our method, respectively. The first column is ground truth, the second column is the results of YOLOv5s and the third column is the results of our method. Areas marked by blue circles indicates that YOLOv5s has missed detection here.

methods are set to the same. Output results are shown in **Figure 7.**

Targets missed by the original YOLOv5s method are usually in clutter, such as bicycles and pedestrians with trees and other interference objects in the background. These objects are partially occluded or their own thermal characteristics are not obvious, which puts forward higher requirements for the detection capability of the method. **Figure 7** shows that our method can effectively distinguish targets from complex background and reduce missed detection. Targets such

TABLE 3. The performance of comparative methods is shown in the table. The best AP is highlighted in bold.

Method	car	person	bicycle	mAP50	Para.	FPS
SSD	61.6	40.9	43.6	48.7	34.3M	47
Faster-RCNN	67.6	39.6	54.7	53.9	278M	21
Retina-net	71.5	52.3	61.3	61.7	38.0M	41
RefineDet[36]	84.5	77.1	57.2	73.0	58.2M	24
ThermalDet[37]	85.5	78.2	60.0	74.6	59.1M	22
YOLOv5m	91.2	**86.9**	65.6	81.3	20.86M	67
YOLOv5s	90.1	82.9	65.3	79.4	7.02M	109
YOLOv5s+Ours	**91.3**	85.1	**68.3**	**81.6**	7.03M	94

TABLE 4. Comparison with SCMR-SGE after embedding it at another possible position in FPN.

Module	Position	mAP50	mAP50:95
SCMR-SGE	1	81.3	42.9
SCMR-SGE	3	80.3	42.1

as bicycles which are largely obscured by the riders, and pedestrians with characteristics similar to the tree trunk are successfully detected, which intuitively shows the superior performance of our method.

2) COMPARISON WITH OTHER METHODS

We also conduct comparative experiments with different methods to verify the outstanding performance of our method. The performance of each method is shown in **Table 3**.

Compared with original YOLOv5s, our proposed method gains all-round improvement of detection performance with negligible performance loss. APs for all three targets have increased significantly. After adding our two modules, mAP50 reaches 81.6%, which is 2.2% higher than before addition. It is worth mentioning that AP for the bicycle class, which lack of samples and salient features in clutter, having a 3% increasing, reaches 68.3%. On the other side, the parameters of the modified method have only increased by 0.01M, resulting in 15 frame per second performance loss. This is quite acceptable especially when compared with YOLOv5m, the widening and deepening version of YOLOv5s. Our method not only has higher accuracy than the original YOLOv5m method, but also has great advantages in detection speed.

C. DISCUSSION

To minimize performance loss, we only use each module once and embed them into the FPN. Two reasons can be account for the effectiveness of our arrangement. First, for SCMR-SGE, in [26], author mentioned that spatial enhance attention is used to enhance feature maps containing rich semantic information, which means it should be embedded at relative deep layers. In order to prove this point, we have also done experiments that embedded our SCMR-SGE Module at another concatenate function marked with green rectangle in **Figure 5**, which are shown in **Table 4**.

TABLE 5. Comparison with different channel attention mechanism after embedding them at another possible position in FPN.

Module	Position	mAP50	mAP50:95
JA	2	81.4	43.0
JA	3	80.7	43.0
SE(avg+max pool)	2	80.9	43.2
SE(avg+max pool)	3	80.7	42.9

TABLE 6. Comparison with YOLOv5m after embed our modules in FPN.

Method	car	person	bicycle	mAP50	Para.	FPS
YOLOv5m	91.2	86.9	65.6	81.3	20.86M	67
YOLOv5m+ours	**91.6**	**87.3**	**69.3**	**82.7**	20.87M	58

It can be clearly seen that embedding SCMR-SGE at position 3 does not show good results, which proves that it is not suitable to enhance feature maps generated by shallow layers because they still contain rough information such as edges and textures. On the other hand, our JA module is proposed to filter channel information, which means it should be used when some channels are not helpful to the task. Thus, the position of JA module should behind SCMR-SGE Module, where can not only inherit upstream feature enhancement effect and filter semantic information, but also transfer the filtered feature map to each detection head as the network goes down through PANet. We also investigate different positions for JA module and employ SE module as comparative experiments, which are shown in **Table 5.**

It can be seen from the table that the performance of both two methods decreases after changing position from 2 to 3. Similarly, considering that position 3 contains feature maps processed by shallow layers, which are essential to pinpoint the location of targets, it is unreasonable to filter the feature maps when each of them is of value. What's more, important factors can be recalibrated with the help of the up-sampling function at position 2, which cannot be realized in other positions. Thus, we can draw a conclusion that the embedding positions of these two modules are reasonable.

To verify the performance of our method, we also embed our modules at the same positions in YOLOv5m method. The performance of embedding our modules before and after are shown in **Table 6.**

It can be seen from the table that mAP for all targets also have improved. After embedding our modules, the modified YOLOv5m method is state-of-art on original FLIR dataset.

At the same time, our modules don't have much impact on the detection speed of the method, the parameters only increase 0.01M and the performance loss was only 9 FPS.

Since the only difference between YOLOv5s and YOLOv5m method is that YOLOv5m double the number of CSP structure, which is shown in **Figure 5**, for better feature extraction capability. It means that effort, trying to improve the feature extraction capabilities of backbone network, will

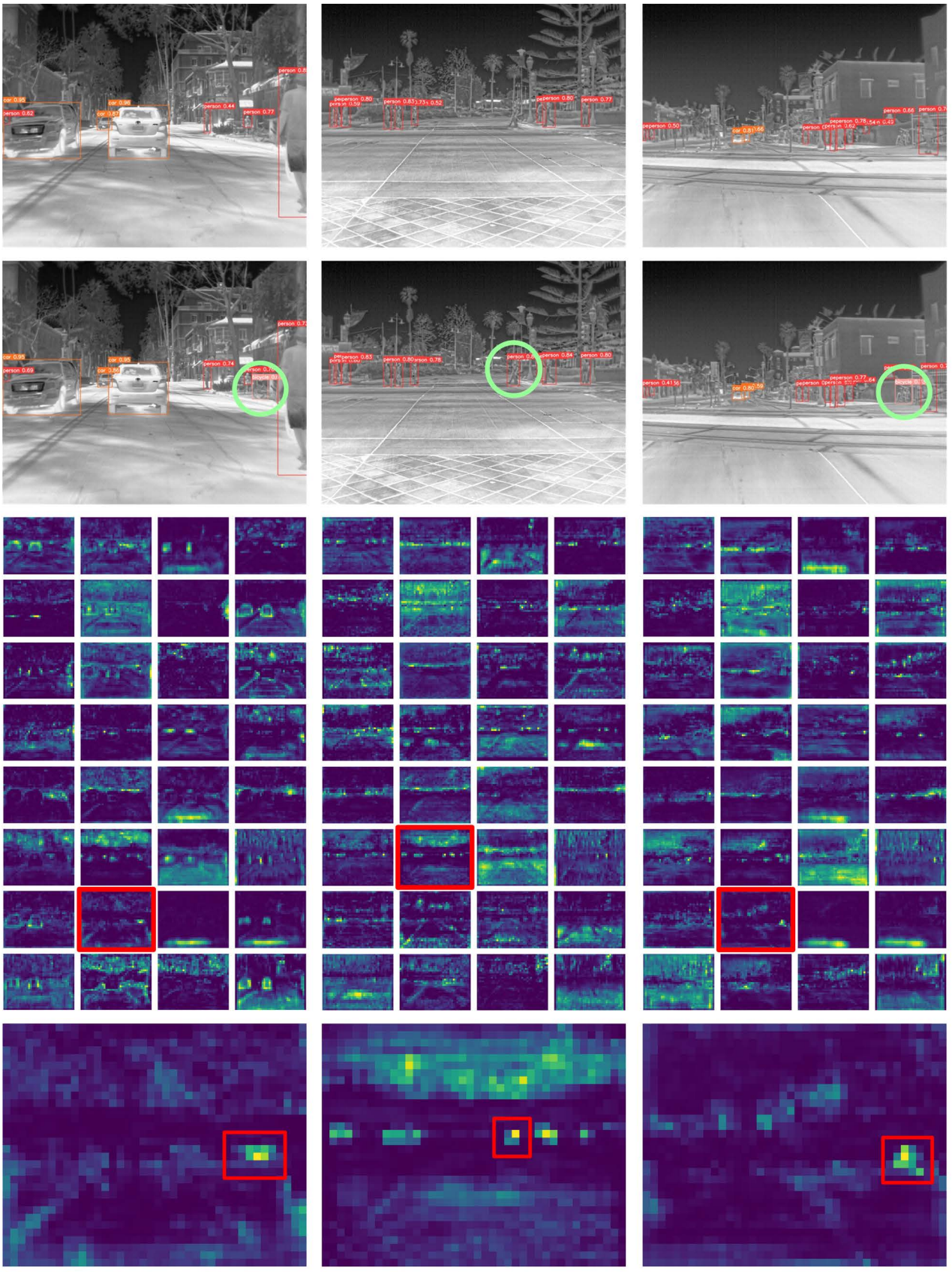

FIGURE 8. Demonstration of proposed modules' effectiveness. The first row is detection output obtained by YOLOv5s. The second row is detection output obtained by our method. The third row is the first 32 feature maps processed by SCMR-SGE and JA module. The fourth row is the enlarged significant feature maps which are marked with red rectangle in the third row.

TABLE 7. Experiment results on KAIST dataset.

Module	AP50	AP95
YOLOv5s	88.1	50.7
YOLOv5s+SE	89.6	51.1
YOLOv5s+ours	90.7	51.3
YOLOv5m	90.0	51.4
YOLOv5m+SE	90.7	51.7
YOLOv5m+ours	91.6	52.5

still pay off when they are introduced to our methods. Thus, we can draw a conclusion that our arrangement of modules is effective and universal.

We also visualize the effectiveness of our proposed modules to verify the robustness of our method, which are shown in **Figure 8.**

We choose three representative images and visualize the first 32 feature maps processed by our SCMR-SGE Module and JA Module. The areas marked with green circles in the second row indicates that there are objects miss-detected by the original YOLOv5s method. Feature maps marked with red rectangles in the third row are those carrying strong enhanced semantics information which contributes to successfully detecting the originally missed targets, and the fourth row shows the enlarged images of those significant feature maps for better viewing. It can be clearly seen that our modules successfully enhanced interested targets where are missed in original method. All the conducted experiments show that, for infrared object detection task, our proposed methods in this article have advantages in both detection accuracy and detection efficiency.

D. EXPERIMENTS ON THE KAIST DATASET

We also complete a series of comparative experiments on re-annotated KAIST dataset to demonstrate the robustness and advantage of our proposed method. KAIST pedestrian dataset includes 95328 images in total, containing 103128 dense annotations. Each scene is composed of RGB color images and corresponding long wave infrared images. The dataset captured various conventional traffic scenes including campus, street and countryside during the day and night, respectively. However, the original dataset suffers from rather poor annotation quality, Li *et al.* cleaned up and sampled images every 2 frames from training videos, excluded heavily occluded, truncated and small (< 50 pixels) pedestrian instances, and re-annotate targets to improve the dataset quality [36], which are chosen as our comparative experiment dataset. Re-annotated KAIST dataset has 7601 images and we divide them into 6413 images for training and 1188 images for validation. All labels introduced to our experiments are set as person since the resolution is too low to distinguish the difference between person and people in long wave infrared images.

It can be seen from the **Table 7** that compared with baseline YOLOv5s and YOLOv5m methods, modified methods equipped with our modules also have significant increase in AP50 and AP95. In other words, modified methods have better feature extraction capability than their baselines, which proves the effectiveness of our modules.

V. CONCLUSION

Aiming at the problem of low contrast and lack of fine-grained and texture information in infrared images, we propose two novel attention mechanism to improve the feature extraction capabilities for infrared object detection methods. SCMR-SGE module can simultaneously enhance high-order semantic information and suppress possible background noise. At the same time, JA module can selectively enhance or suppress channel information to improve detection performance. Moreover, our proposed method is different from other methods that embed attention mechanism to each module of the backbone network, which only utilizes SCMR-SGE and JA modules at two essential positions at FPN network to filter spatial and channel information, respectively.

Experimental results on the FLIR-ADAS and KAIST datasets show that proposed method can effectively improve the detection performance for infrared targets with low cost. Our proposed method can maintain the detection speed at around 60fps while the mAP of FLIR dataset reaches 82.7% basing on YOLOv5m only, which is higher than other state-of-art methods, such as RefineDet and ThermalDet. But it also needs to be noted that, although our method can gain all-round improvement on mAP of all categories, the mAP of bicycle class is still not high, and the problem of poor generalization ability caused by the imbalanced of dataset samples still exists, which deserves further research.

REFERENCES

[1] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014, doi: 10.1109/TGRS.2013.2242477.

[2] R. Gade and T. B. Moeslund, "Thermal cameras and applications: A survey," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 245–262, Jan. 2014, doi: 10.1007/s00138-013-0570-5.

[3] C. Q. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013, doi: 10.1109/TIP.2013.2281420.

[4] W. Schroeder, P. Oliva, L. Giglio, and I. A. Csiszar, "The new VIIRS 375 m active fire detection data product: Algorithm description and initial assessment," *Remote Sens. Environ.*, vol. 143, pp. 85–96, Mar. 2014, doi: 10.1016/j.rse.2013.12.008.

[5] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, and H. Wang, "Learning deep multi-level similarity for thermal infrared object tracking," *IEEE Trans. Multimedia*, vol. 23, pp. 2114–2126, 2021, doi: 10.1109/TMM.2020.3008028.

[6] C. D. Elvidge, K. Baugh, M. Zhizhin, F. C. Hsu, and T. Ghosh, "VIIRS night-time lights," *Int. J. Remote Sens.*, vol. 38, no. 21, pp. 5860–5879, Jun. 2017, doi: 10.1080/01431161.2017.1342050.

[7] W. Anguelov, D. Erhan, D. Szegedy, C. Reed, S. Fu, and C. Berg, "SSD: Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2].

[8] 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)*, Seattle, WA, USA, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[10] 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, doi: 10.1109/TPAMI.2016.2577031.

[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)*, Seattle, WA, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[12] B. McIntosh, S. Venkataramanan, and A. Mahalanobis, "Infrared target detection in cluttered environments by maximization of a target to clutter ratio (TCR) metric using a convolutional neural network," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 1, pp. 485–496, Feb. 2021, doi: 10.1109/TAES.2020.3024391.

[13] A. Mehmood and N. Nasrabadi, "Wavelet-RX anomaly detection for dual-band forward-looking infrared imagery," *Appl. Opt.*, vol. 49, no. 24, pp. 4621–4632, 2010, doi: 10.1364/AO.49.004621.

[14] R. Liu, "Point target detection of infrared images with eigentargets," *Opt. Eng.*, vol. 46, no. 11, Nov. 2007, Art. no. 110502, doi: 10.1117/1.2802301.

[15] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Inf. Fusion*, vol. 50, pp. 148–157, Oct. 2019, doi: 10.1016/j.inffus.2018.11.017.

[16] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," 2019, *arXiv:1901.02645*.

[17] 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 2019, doi: 10.1109/TIP.2018.2887342.

[18] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 11534–11542, doi: 10.1109/CVPR42600.2020.01155.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[21] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 510–519, doi: 10.1109/CVPR.2019.00060.

[22] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[23] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," in *Proc. 29th Brit. Mach. Vis. Conf. (BMVC)*, Newcastle, U.K., Sep. 2018, pp. 147–161.

[24] C. Zhao, J. Wang, N. Su, Y. Yan, and X. Xing, "Low contrast infrared target detection method based on residual thermal backbone network and weighting loss function," *Remote Sens.*, vol. 14, no. 1, p. 177, Jan. 2022, doi: 10.3390/rs14010177.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 1–11.

[26] X. Li, X. Hu, and J. Yang, "Spatial group-wise enhance: Improving semantic feature learning in convolutional networks," 2019, *arXiv:1905.09646*.

[27] Z. Cui, X. Wang, N. Liu, Z. Cao, and J. Yang, "Ship detection in large-scale SAR images via spatial shuffle-group enhance attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 379–391, Jan. 2021, doi: 10.1109/TGRS.2020.2997200.

[28] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6848–6856, doi: 10.1109/CVPR.2018.00716.

[29] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[30] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam, "CASENet: Deep category-aware semantic edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1761–1770, doi: 10.1109/CVPR.2017.191.

[31] N. Ma, X. Zhang, H. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 122–138, doi: 10.1007/978-3-030-01264-9_8.

[32] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1713–1721, doi: 10.1109/CVPR.2015.7298780.

[33] J. Glenn, S. Alex, B. Jirka, S. Christopher, C. Liu, R. Prashant. (2021). *Yolov5*. Accessed: Jul. 10, 2021. [Online]. Available: https://github.com/ultralytics/yolov5

[34] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 8759–8768, doi: 10.1109/CVPR.2018.00913.

[35] FA Group. (2018). *Flir Thermal Dataset for Algorithm Training*. (Accessed: Aug. 17, 2021). [Online]. Available: https://www.flir.in/oem/adas/adas-dataset-form/

[36] 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. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 4203–4212, doi: 10.1109/CVPR.2018.00442.

[37] 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)*, Chengdu, China, Dec. 2019, pp. 1965–1969, doi: 10.1109/ICCC47050.2019.9064036.

[38] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proc. 29th Brit. Mach. Vis. Conf. (BMVC)*, Newcastle, U.K., Sep. 2018, pp. 1–12.

ZHIHENG PAN (Graduate Student Member, IEEE) received the B.S. degree in optoelectronic information science and engineering from the Nanjing University of Posts and Telecommunications, Jiangsu, China, in 2019. He is currently pursuing the M.S. degree in electronic information with Shihezi University, Xinjiang, China. His research interests include image processing, object detection, and fine-grained visual categorization.

LIUCHAO XU received the B.S. degree in electrical engineering from Shihezi University, Shihezi, Xinjiang, China, in 2020, where he is currently pursuing the M.S. degree in electrical engineering. His research interests include photovoltaic array fault diagnosis and location.

CHUANDONG LIANG received the B.S. degree in electrical engineering from Shihezi University, Shihezi, Xinjiang, in 2021, where he is currently pursuing the M.S. degree in electrical engineering. His research interests include integrated energy buildings, energy transportation systems, and solar panel automatic cleaning devices. He was a recipient of the Outstanding Graduate of Shihezi University, the First Prize of National College Students and Mathematics Competition (CMC), and the First Prize of National College Students and Mathematical Modeling Competition (CUMCM).

KUI PAN received the bachelor's degree in energy and power engineering from Chongqing Jiaotong University, in 2021. He is currently pursuing the master's degree in mechanical engineering with Shihezi University. His research interests include distributed photovoltaic power generation and electric vehicles.

MI ZHAO was born in Shihezi, Xinjiang, China, in 1980. She received the B.S., M.S., and Ph.D. degrees from Xidian University, Xi'an, China, in 2002, 2006, and 2009, respectively.

She has authored or coauthored over 20 publications. She has also coauthored the book *Optimal Supervisory Control of Automated Manufacturing Systems*. She is currently an Associate Professor with the College of Machinery and Electricity. Her main research interests include modeling, analysis, and control of smart grids and supervisory control of discrete event systems.

MIN LU received the B.S. degree in control engineering from Xinjiang University, Xinjiang, China, in 2008, and the Ph.D. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2020. She is currently a Professor with Shihezi University. Her research interests include reliability research of power electronics devices and wind power generation technology.