UAV-Based Human Detection With Visible-Thermal Fused YOLOv5 Network

Xiongxin Zou, Tangle Peng, and Yimin Zhou, *Member, IEEE*

***Abstract*—Timely and effective search and rescue (SAR) is highly desired in the disaster rescues. Unmanned aerial vehicles (UAVs) can quickly conduct aerial searches to assist SAR with the equipped sensors. A visible-thermal human detection model based on an improved you only look once version 5 (YOLOv5) network is proposed to compensate the deficiencies of visible data with thermal images. The complementary information between the visible images and thermal images are considered with a partially shared two-stream backbone network, so as to better preserve the information of each branch while reducing the domain distinction and extracting the modality-invariant features. Features of the two modalities are fused via a fusion module with a multidimensional attention mechanism. By taking the pixels outside the region of interest as negative samples, the extra loss function can suppress the uncorrelated feature extraction of the backbone to enhance the effective feature representation. The proposed visible-thermal human detection model has been deployed on the UAV with satisfied human detection performance. Comparative experiments on the multispectral pedestrian dataset KAIST have also been performed to demonstrate that the proposed model outperforms other visible-thermal object detection models with log-average miss rate.**

***Index Terms*—Object detection, two-stream network, unmanned aerial vehicle (UAV), visible-thermal fusion, YOLOv5.**

I. Introduction

THE disasters have increased dramatically as a result of destruction of the natural environments. People are always killed in these natural disasters, and 432 global disasters were recorded in 2021, resulting in 10492 deaths [1]. Survival rates for the victims would decrease as the trapped time increases, especially in complex disaster sites (such as mountains, earthquakes, floods), making it more difficult to carry out search and rescue (SAR) effectively and timely. Unmanned aerial vehicles (UAVs) have been widely used in the SAR missions in recent decades due to their flexibility and maneuverability to quickly locate the victims via various target detection technologies.

UAVs commonly use the airborne camera to capture the aerial images for the object detection. However, in the dim environments, it is difficult to extract the effective features due to the lost contrasts and details of the visible images, which is unreliable for the missions involving the safety of human life. Besides visible camera, multitype sensors can be installed on the UAVs for the perception, i.e., thermal cameras [2], radar [3], and LiDAR [4]. The thermal camera can capture the images via the thermal radiation, which is highly effective for the detection without interference from other ambient illumination conditions. Hence, a lot of research focus on how to combine the visible and thermal images for the human detection. The two-stream network architecture is normally adopted in the visible-thermal object detection networks, where the feature-level fusion is more popular than the decision-level fusion due to the reduced computation burden and more complete representations.

Considering different imaging principles, there are large domain distinctions between the visible and thermal images [5]. Specifically, the former mainly describes the reflected light information with abundant texture gradient variations, while the latter represents the thermal radiation with strong contrast pixel intensities. How to extract fused features retaining richer effective information from the two modes is worthy further investigating. Certain methods have been developed to reduce the domain distinctions to obtain more conducive cross-modal features. Xu et al. [6] have trained the backbone's cross-modal representation through a regional reconstruction network to refactor the thermal images from the associated visible images. A cross-modal feature learning module [9] is designed to achieve the information interaction between the two modalities, so that different levels of the cross-modal representation can be learned at each layer of the backbone.

The visible and thermal images have different contributions to the detection under diversified illumination conditions, hence, the weights of the different modalities can be tuned adaptively to improve the fused feature performance. Zhang et al. [10] have proposed a confidence perception fusion method to select features with more useful information and suppress useless information via reweighting the importance of the features. An

Manuscript received 10 April 2023; revised 27 July 2023; accepted 27 August 2023. Date of publication 26 September 2023; date of current version 23 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61973296 and Grant U1913201, in part by the Shenzhen Science and Technology Innovation Commission Project under Grant JSGG20210802154535003 and Grant JCYJ20220818101206015, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021B1515120038. Paper no. TII-23-1255. *(Corresponding author: Yimin Zhou.)*

Xiongxin Zou and Tangle Peng are with the School of information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan 411201, China, and also with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: xx.zou@siat.ac.cn; tl.peng@siat.ac.cn).

Yimin Zhou is with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: ym.zhou@siat.ac.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2023.3310792.

Digital Object Identifier 10.1109/TII.2023.3310792

extra network based on only visible images is used to weight the results of the visible and thermal branches, which are then fused via the gated fusion layer [11].

The abovementioned methods have addressed the issue of poor fusion performance between the visible and thermal images, possessing better detection accuracy. However, most methods are based on the faster region-based convolutional neural network (R-CNN), which may not meet the real-time requirements in certain application scenarios. Several studies have made improvements based on one-stage models, such as single shot multibox detector [12], [13] and RetinaNet [14]. The you only look once (YOLO) series of the object detection algorithms have been widely applied in practical deployments. A multimodal object detection approach based on YOLO version 5 (YOLOv5) has been proposed to achieve feature fusion via the channel attention (CA) and weighted summation [15]. Furthermore, the coordattention module is introduced to capture the effective features through parallel decoupled spatial attention (SA) and CA so as to improve the detection accuracy [16]. Then, a cross-feature enhancement module is designed to enhance the feature representation and an improved multihead attention module is used for the feature combination [17]. Similarly, a feature interaction module is introduced to enhance cross-modal information interaction during the feature extraction, while a self-attention feature fusion module is established to describe the long-range dependencies among the features [18]. Although these one-stage-based models can achieve a better balance between the detection accuracy and speed, they did not directly consider the model deployment on the portable devices during the improvement process.

Hence, this article investigates the challenges arising from significant domain difference between the visible and infrared image modalities, which would lead to poor fusion and issues related to deploying multimodal models at the edge. A visible-thermal infrared object detection network based on YOLOv5 is, thus, proposed to assist UAVs in capturing images for the SAR operations.

Compared to the aforementioned YOLOv5-based methods, the proposed approach utilizes the shared convolutional operators to reduce the modality discrepancies. In the design of the fusion module, the method can avoid the computationally expensive self-attention mechanisms but adopt a more lightweight modified spatial-CA instead, achieving a better balance between the detection accuracy and speed. Moreover, the fusion module can automatically select more effective features through the element-wise maximum operations. The use of the background weakening loss in the backbone can further enhance the detection performance by suppressing the nontarget feature learning. The simplification of the neck and head components enables the model to maintain lightweight and low loss rates.

The main contributions of this article are summarized as follows.

1) A multimodal object detection model is designed based on YOLOv5 with a partial operator sharing two-stream network architecture considering the semantic similarity of various modal features and limited power of the UAV platform.
2) A dual attention fusion (DAF) module is proposed to adaptively fuse different modal features with the same semantics.
3) A background weakening constraint is further proposed to make the backbone network to focus on the object regions so as to obtain high-quality human features.
4) A series of experiments are performed to verify the effectiveness of the proposed visible-thermal human detection network on the KAIST dataset and feasibility for the edge deployment on UAVs.

The rest of this article is organized as follows. Section II introduces the model and improvement schemes proposed in this article. The experimental settings and results are explained in Section III. Finally, Section IV concludes this article.

II. Proposed Method

A. Overall Architecture

YOLOv5-l is taken as the baseline model in the proposed visible-thermal object detection model, as demonstrated in Fig. 1, where the overall structure can be divided into the following three parts: backbone, neck, and head. The backbone is a modified Darknet [19] based on cross-stage partial (CSP) network [20], which can be divided into four parts, each of which is labeled with a depth d ($d \in \{0, 1, 2, 3\}$), and the output feature maps with different depths are labeled as F_d. As shown in the top-left corner of Fig. 1, the CSP module is denoted as CSP_n^x, where $x \in \{0, 1\}$ indicates the use of residual structures in the internal blocks, and n is their number. In the CSP module, the features are split into two groups and transmitted to different branches in order to eliminate the redundant gradient information in the residual structure. This strategy can effectively reduce the computation in the bottleneck layer without loss of accuracy. The neck can aggregate the features with different scales by feature pyramid network [21] and path aggregation network [22] to achieve the interaction of detail and semantic information. Considering that the human body always maintains a similar size under the perspective of the drone, unlike the baseline model with three detectors to predict the targets with significant size variations, the proposed model only uses a single detector at the head.

Using the same backbone network to encode two kinds of images with large domain distinction will lead to a huge information loss. Therefore, we adopt a two-stream network structure in the backbone so that the two branches with the same structure can extract features from visible and thermal images independently to retain as much as useful information. The feature maps extracted with the visible branch and the thermal branch are marked as F_d^v and F_d^t, while the finally fused feature maps are marked as F_d^f. Although the independent backbone networks can retain more information, it would cause large domain distinction between different modal features, which is negative for the feature fusion. Inspired by the method in [23], a strategy of sharing deeper networks is adopted to use the encoders with the same parameters to obtain invariant modality features.

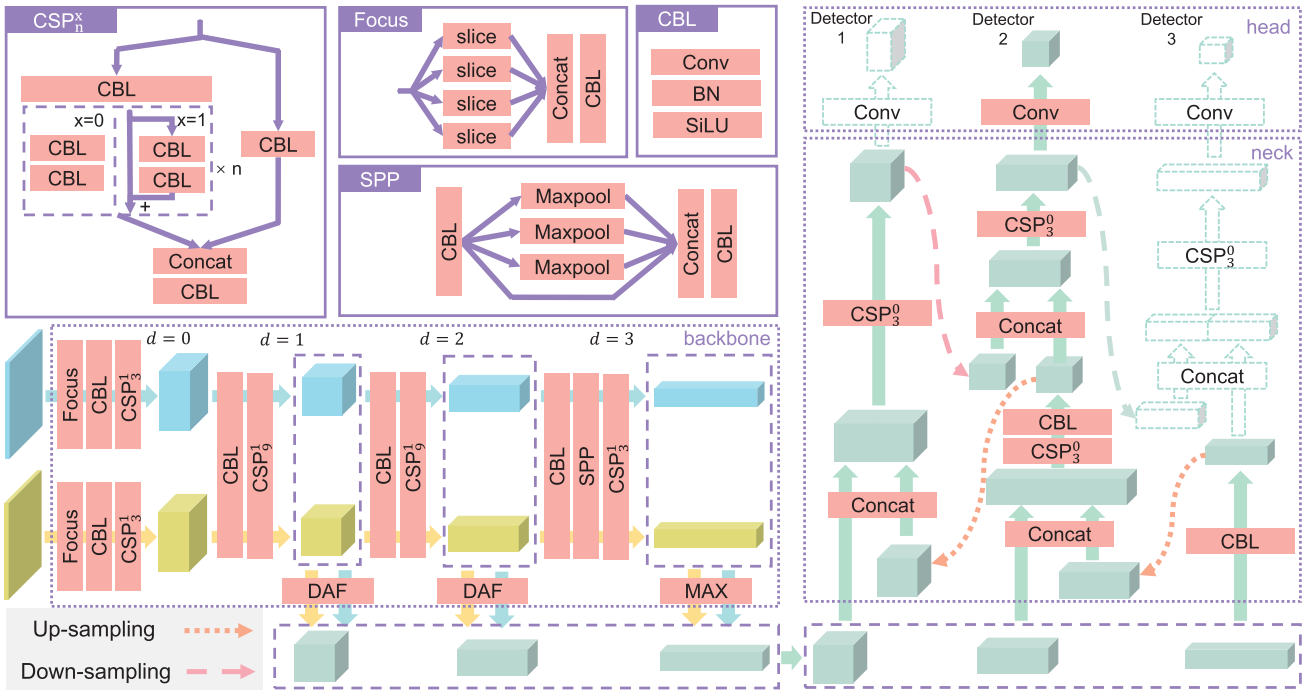

Fig. 1. Overall architecture of the proposed model, where the blue blocks are the visible feature maps, yellow blocks are the thermal feature maps, and green blocks are the fused feature maps. The up-sampling operation is performed using the nearest-neighbor interpolation, while the down-sampling is achieved through convolution. The transparent blocks within the dashed border are discarded, so only one detector is used.

With the deepening of the network layers, semantic information is more abundant and the feature difference caused by the sensor imaging principle in the original images will be much lessened. In this case, using the same convolution operator to encode feature maps will not or only result in minimal information loss. Hence, as shown in Fig. 1, the proposed model has shared layers $(d = 1, 2, 3)$ in which the two branches perform convolution operations using the same convolution operator, and the layers of $d = 0$ remain independent. In contrast to the fully independent backbone structure, the partially shared two-stream backbone structure can effectively reduce the model parameters while maintaining the detection performance. Moreover, since the same convolution operator is used in the last few layers, the final feature maps extracted from the two branches can maintain semantic consistency in the corresponding channels, which can facilitate the fusion of the two modality features in the subsequent steps.

B. DAF Module

Attention mechanism can adaptively improve the contribution of the useful information to aid feature fusion of the visible and thermal images. The DAF module, combining the spatial and CA mechanisms, is designed to adaptively adjust the importance of different modality features in varied illumination environments. To start with, as illustrated in Fig. 2(a), we concatenate (Concat) the features of the two modalities, and transfer them to the SA module. In Fig. 2(b), the SA module can first generate two feature maps via calculating the maximum and mean of the feature values at the same pixel position, then a space weight map is predicted by the two 1×1 convolution layers so as to weight each pixel position of the inputs.

The SA module can adaptively enhance the feature values of the latent object regions, thereby suppressing the interference of the irrelevant regions. Next, the global importance of each channel of the two modalities is evaluated through the CA module to further suppress the impact of invalid channels on the subsequent steps. In the classic CA module [24], a channel weight vector can be predicted via two fully connected layers from the global average pooling results of each channel. The two fully connected layers would reduce the computational complexity via decreasing the dimension of the first layer, but it is not conducive to capture the dependencies across all channels. Therefore, an efficient channel attention module [25] is introduced [see Fig. 2(c)], several 1×1 convolution layers are used to replace the fully connected layers while avoiding the channel dimension reduction during the channel weights prediction.

The most common used method to merge features is to directly Concat or add up the values of their corresponding pixel position (Add). In the DAF module, a different method, called Max, is applied to choose the maximum of the corresponding feature values of the two sets of feature maps as the merged values, written as

$$V_{i,j,k} = \max\left\{V^v_{i,j,k},\ V^t_{i,j,k}\right\} \tag{1}$$

where V is the feature values of $F \in \mathbf{R}^{w \times h \times c}$, (i, j) is the spatial coordinates, and k is the channel number. Hence, it can effectively avoid the channel dimension doubling in the

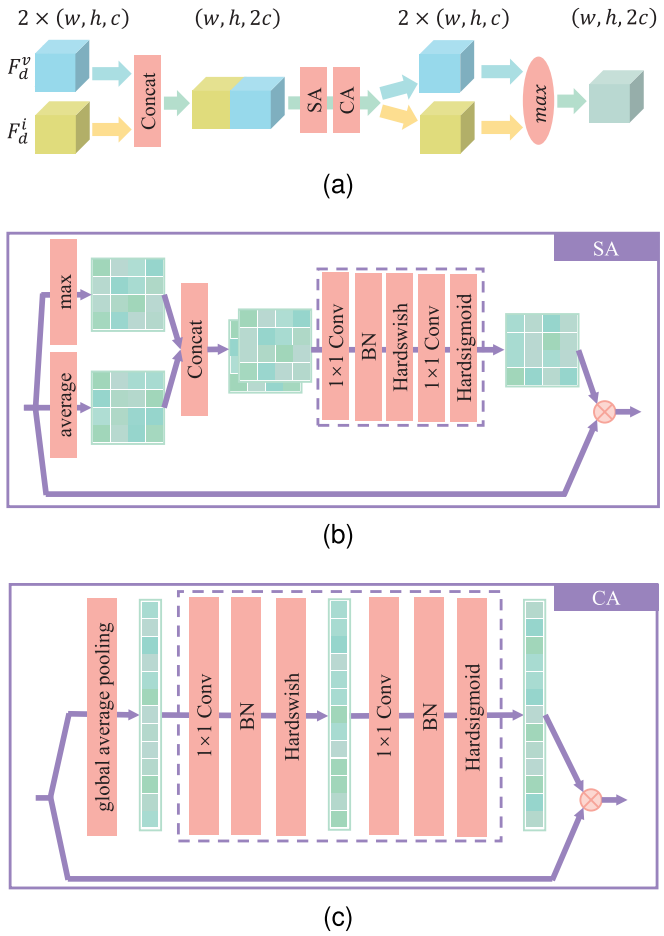

Fig. 2. (a) DAF module. (b) SA module. (c) CA module.

subsequent procedure. Besides, the Max operation can prevent serious detection degradation if one of the fused images is invalid.

C. Loss Function

The loss function of YOLOv5 consists of three parts: classification loss L_{cls}, object confidence loss L_{obj}, and bounding box loss L_{box}. L_{cls} and L_{obj} employs the binary cross entropy (BCE) function [26],

$$\text{BCE} = x - x \times z + \log\left(1 + e^{-x}\right) \tag{2}$$

and L_{box} employs the generalized intersection over union (GIoU) [27]

$$\text{GIoU} = 1 - \left(\frac{A \cap B}{A \cup B} - \frac{C - (A \cup B)}{C}\right). \tag{3}$$

Thus, the total loss function of YOLOv5 can be written as

$$L^i_{\text{detect}} = \alpha_{\text{cls}} L^i_{\text{cls}} + \alpha_{\text{obj}} L^i_{\text{obj}} + \alpha^i_k \alpha_{\text{box}} L^i_{\text{box}} \tag{4}$$

$$L_{\text{detect}} = \sum_{i=1}^{n} L^i_{\text{detect}} \tag{5}$$

where n denotes the number of the detectors, $d \in \{\alpha_{\text{cls}}, \alpha_{\text{obj}}, \alpha_{\text{box}}\}$ are the loss weights, and α_k is the weight of the detectors.

Besides, a background weakening constraint is proposed as part of the loss function, denoted by L_{bw}, where the background regions are used as negative samples to suppress the extraction of the nonobject features, thereby reducing the noise in the backbone encoded features. The pretrained parameters of YOLOv5, trained on COCO2017 dataset, are used to initialize the backbone and fine-tuned on the training set, which can improve the model accuracy and reduce the fitting times. Although the pretrained parameters contains a large number of unnecessary feature operators, the proposed L_{bw} can effectively suppress these operators. Besides, L_{bw} can assist the DAF module to better focus on the target area.

Specifically, based on the ground truth, a set of corresponding masks are first generated, denoted as $M = \{M_1, M_2, M_3\}$. Then, the feature maps encoded from the backbone are masked to generate a set of region of interest (ROI) labels

$$T^m_d(i,j) = F^m_d(i,j) \circ M_d(i,j) \tag{6}$$

where $\circ$ denotes the element-wise product and m is the modality of the feature maps. Smooth $L1$ loss is applied to train the model to suppress the nonobject features, written as

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{others.} \end{cases} \tag{7}$$

So the background weakening loss is

$$L^m_{\text{bw}}(F^m, T^m_d) = \text{smooth}_{L1}(F^m - T^m_d) \tag{8}$$

$$L_{\text{bw}} = L^v_{\text{bw}} + L^t_{\text{bw}}. \tag{9}$$

Fig. 3 depicts the detailed workflow of L_{bw}. So the proposed loss function of the model is

$$L = \alpha_{\text{detect}} L_{\text{detect}} + \alpha_{\text{bw}} L_{\text{bw}} \tag{10}$$

where α_{detect} and α_{bw} denote the weights of L_{detect} and L_{bw}.

III. Experiments and Result Analysis

A. Dataset and Settings

A series of experiments are performed to verify the proposed method with KAIST Multispectral Pedestrian Dataset (KAIST) [28]. KAIST contains 95 000 aligned multispectral images with a single frame size of 640 × 512 and a total of 103 128 annotated bounding boxes. In the experiments, the sanitized annotations from [10] and [30] are used. The training set contains 8892 images and 21 542 annotations, while the testing set contains 2252 images and 2317 annotations. Two commonly evaluation metrics used in the pedestrian detection are applied: log-average miss rate (LAMR) and recall rate. Here, the miss rate (MR) indicates the percentage of the missed objects (false negative) in all ground truths (true positive and false negative)

$$\text{MR} = \frac{\text{FN}}{\text{TP} + \text{FN}}. \tag{11}$$

TABLE I
EVALUATIONS ON THE KAIST DATASET

Method	Backbone			LAMR↓			Recall↑
	Type	Parameter	FLOPs	All	Day	Night	All
IAF R-CNN [11]	VGG-16	1.47×10^7	1.00×10^{11}	15.73	14.55	18.26	91.20
CIAN [29]	VGG-16	1.47×10^7	1.00×10^{11}	14.12	14.78	11.13	97.25
MSDS-RCNN [30]	VGG-16	1.47×10^7	1.00×10^{11}	11.63	10.60	13.73	94.30
AR-CNN [10]	VGG-16	1.47×10^7	1.00×10^{11}	9.34	9.94	8.38	97.25
MBNet [9]	ResNet50	2.35×10^7	2.70×10^{10}	8.13	8.28	7.86	98.42
MLPD [31]	VGG-16	1.47×10^7	1.00×10^{11}	7.58	7.96	6.95	96.70
Our model	CSPDarknet-s	$\mathbf{4.21 \times 10^6}$	$\mathbf{4.41 \times 10^9}$	7.58	8.80	**5.05**	98.49
	CSPDarknet-l	2.71×10^7	3.05×10^{10}	**6.48**	**6.65**	5.82	**98.69**

The bold minimum values in the columns for Parameter, FLOPs, and LAMR, and the maximum value for Recall.

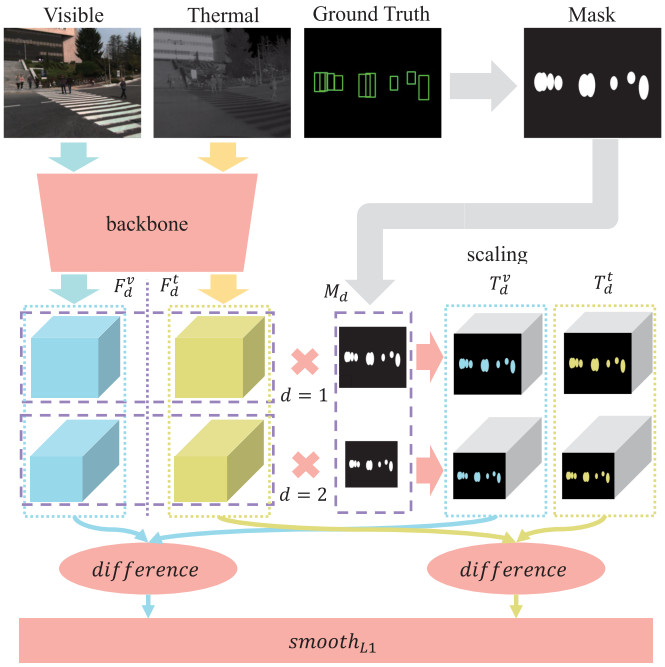

Fig. 3. Calculation of the background weakening loss.

LAMR is the mean of MR at nine false positive per image (FPPI) rates evenly distributed in the log-space from 0.01 to 1. The recall rate is the proportion of the detected objects (true positive) in all ground truths

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{12}$$

In order to evaluate the lightweight effect of the model, the amount of model parameters and the floating-point operations (FLOPs) are added as the evaluation indicators.

The experimental hardware environment includes 2 × NVIDIA GeForce RTX 3080 (10 G) GPUs and 2 × Intel Xeon Silver 4214R CPU. During the training process, the stochastic gradient descent optimizer is employed, along with the cosine annealing policy to set the learning rate (with an initial value of $1.25 \times e^{-3}$). The basic data augmentation techniques, such as random scaling, inversion, and color transformations are applied as well.

B. Comparative and Ablation Experiments

The proposed model is compared with other models in multispectral pedestrian detection on the KAIST dataset. Among them, IAF R-CNN [11], MSDS-RCNN [30], and AR-CNN [10] are based on the faster R-CNN, while CIAN [29], MBNet [9], and MLPD [31] are based on SSD. In comparison to these methods, the proposed model is based on YOLOv5, with a particular emphasis on the mitigating cross-modal domain disparities for better feature fusion and model deployment on the UAVs. Table I lists the LAMR and recall rates for these models in the daytime, nighttime, and all-day scenarios, while the model's backbone architecture, parameter count, and FLOPs are presented as well. It can be seen that the LAMR of the proposed model has the best performance in all test environments, reduced by about 1 percent. Besides, the MR-FPPI curves for each model are depicted in Fig. 4, illustrating that the proposed model can achieve lower MR at relatively low FPPI.

As listed in Table II, the ablation experiments demonstrate the variations in all the evaluation indicators, starting from the baseline model YOLOv5-l, and gradually applying the partially shared two-stream backbone strategy, the DAF module, the L_{bw} constraint and the single detector strategy. The experimental results identify that the proposed model can achieve a favorable balance between the model detection performance and lightweight design. Furthermore, to explore more lightweight solutions, various improvement strategies are applied to the smaller YOLOv5-s, resulting in a reduction of approximately 84% in the parameter amount and 86% in FLOPs. Although the detection performance is slightly reduced after switching to a smaller baseline model, the proposed model can still outperform the compared models.

C. Quantitative Analysis

1) Sharing Strategies: Different sharing strategies in the backbone are tested. Table III lists the LAMR and parameter quantities of the model with different convolution kernel sharing conditions. The strategy of sharing all the convolution operators has inferior LAMR than the other strategies. After the layer with $d = 0$ is set as an independent convolution operator, LAMR can be greatly improved and reduced to 7.51%. This suggests that the visible and thermal images have large discrepancies in the

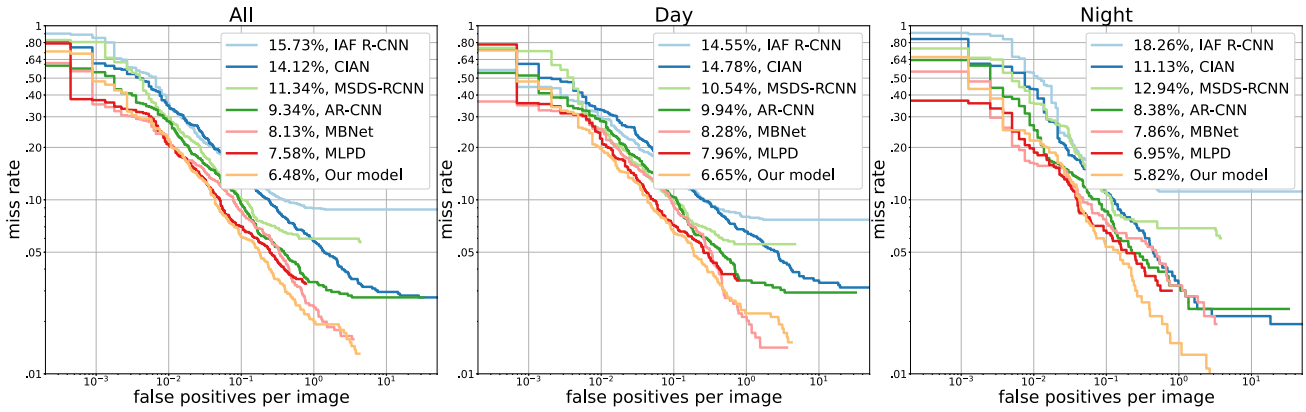

Fig. 4. Performance of the model evaluated with MR-FPPI curves in the daytime, nighttime, and all-day scenarios.

TABLE II
ABLATION STUDY OF THE PROPOSED MODEL

Baseline model	Improvement strategy	LAMR↓			Recall↑	Parameter	FLOPs
		All	Day	Night	All		
YOLOv5-l	Two-stream	7.51	8.34	5.53	98.97	7.16×10^7	1.18×10^{11}
	Two-stream+L_{bw}	7.26	7.69	6.49	99.04	7.16×10^7	1.18×10^{11}
	Two-stream+DAF	7.77	7.93	7.43	98.83	4.76×10^7	9.53×10^{10}
	Two-stream+DAF+L_{bw}	7.69	7.61	7.56	**99.18**	4.76×10^7	9.53×10^{10}
	Two-stream+DAF+L_{bw} +One-detector	**6.48**	**6.65**	5.82	98.69	3.52×10^7	9.04×10^{10}
YOLOv5-s	Two-stream+DAF+L_{bw} +One-detector	7.58	8.80	**5.05**	98.49	$\mathbf{5.50\times10^6}$	$\mathbf{1.30\times10^{10}}$

The bold minimum values in the columns for Parameter, FLOPs, and LAMR, and the maximum value for Recall.

TABLE III
EVALUATION WITH DIFFERENT TWO-STREAM BACKBONES

Share scheme				LAMR↓			Parameter
0	1	2	3	All	Day	Night	
				7.64	**8.34**	6.26	9.85×10^7
			✓	8.43	8.51	8.71	8.12×10^7
		✓	✓	8.15	8.68	7.08	7.36×10^7
	✓	✓	✓	**7.51**	**8.34**	**5.53**	7.16×10^7
✓	✓	✓	✓	8.85	8.46	10.07	$\mathbf{7.14\times10^7}$

The bold minimum values in the columns for Parameter, FLOPs, and LAMR, and the maximum value for Recall.

TABLE IV
EVALUATION OF SIMILARITY WITH DIFFERENT DEPTHS

Method	$d=0$	$d=1$	$d=2$	$d=3$
dist $(\cdot)$	0.00134	0.00026	0.00007	0.00002
cos $(\cdot)$	0.99920	0.99956	0.99972	0.99964

attributes of the modalities at the shallow layers such that the same convolution operator cannot overcome this difference to obtain the meaningful information from both modalities simultaneously. The feature maps of deeper layers would weaken these differences, then the shared convolution operator can be used with minimal information loss. Furthermore, Table IV presents the Euclidean distance dist $(\cdot)$ and cosine similarity cos $(\cdot)$ of the convolutional kernel parameters for the two backbone branches when the sharing strategy is not employed. Consistent with the intuition, as the depth increases, the Euclidean distance between the two branches decreases, and the cosine similarity increases.

TABLE V
EVALUATIONS WITH DIFFERENT MERGER STRATEGIES

Fusion strategy	LAMR↓			Parameter	FLOPs
	All	Day	Night		
Concat	**7.51**	**8.34**	**5.53**	7.16×10^7	1.18×10^{11}
Add	8.15	8.94	6.45	$\mathbf{4.69\times10^7}$	$\mathbf{9.53\times10^{10}}$
Max	7.90	8.50	6.83	$\mathbf{4.69\times10^7}$	$\mathbf{9.53\times10^{10}}$

The bold minimum values in the columns for Parameter, FLOPs, and LAMR, and the maximum value for Recall.

According to Table III, the number of the model parameters gradually decreases as the number of the shared layers increases. The scheme with fully shared convolution operators is about 2.7×10^7 less than that fully independent scheme. Besides, the parameter amount of the scheme with sharing depth of $\{1,2,3\}$ increases only 0.3% than the scheme fully shared. With the combination of the two evaluation indicators, the shared depth of $\{1,2,3\}$ in the two-stream backbone is a better option.

2) Merger Strategies: The comparative experiments are performed on the two mainstream merger strategies, Concat and Add, as well as the Max method. The dimension of the feature maps in the Concat is reduced to ensure that the three merger strategies are as close as possible in terms of spatial complexity by a convolution module. Table V enumerates the LAMR, the parameter numbers, and FLOPs for the three strategies. The parameter numbers are reduced considerably with Add and Max, by about 34% compared to Concat. At the same time, the computational complexity is reduced by 19%. Since the corresponding

TABLE VI
EVALUATION WITH THE DETECTOR AND PRIOR BOUNDING BOX SETTINGS

Detector index			Default				Clustering				Manual			
			LAMR↓			Recall↑	LAMR↓			Recall↑	LAMR↓			Recall↑
1	2	3	All	Day	Night	All	All	Day	Night	All	All	Day	Night	All
✓			86.33	84.59	90.43	35.26	48.77	43.25	59.72	76.50	19.14	16.57	24.66	94.36
	✓		**6.96**	**6.99**	**6.81**	98.96	22.23	23.53	19.27	94.01	54.91	59.91	43.54	58.64
		✓	94.01	95.24	91.01	8.68	68.98	74.40	58.08	34.89	76.79	80.33	72.27	27.81
✓	✓		7.64	7.61	7.46	**99.18**	17.03	16.25	18.72	97.46	9.33	9.28	9.61	**99.31**
	✓	✓	7.04	7.06	6.82	98.96	19.86	22.21	14.94	96.08	55.61	60.80	43.96	58.45
✓		✓	83.20	82.31	85.18	42.66	24.14	23.49	25.60	98.42	15.41	13.65	18.55	**99.31**
✓	✓	✓	7.67	7.60	7.56	**99.18**	**12.23**	**12.52**	**11.86**	**99.45**	**9.06**	**8.96**	**9.25**	**99.31**

The bold minimum values in the columns for Parameter, FLOPs, and LAMR, and the maximum value for Recall.

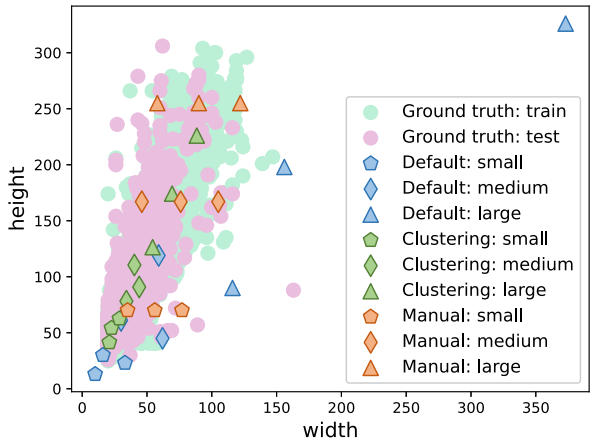

Fig. 5. Distribution of the ground truth and the prior bounding boxes.

channels of the cross-modalities feature maps encoded by the shared convolution operators have the same semantics, Add and Max can achieve similar detection performance to the Concat. The latter slightly outperforms in LAMR, hence, Max is adopted in the proposed model.

3) Detectors and Prior Bounding Box Setting: During the experiments, different detectors contributed differently to the detection results, as seen in Table VI. The detectors with index $\{1, 2, 3\}$ corresponding to three feature maps with various sizes. In the default prior bounding box of YOLOv5, the detector with index $\{2\}$ contributes the most to the model. In order to study the influence of prior frame setting on the detector, two prior bounding box settings are additionally used, one is obtained by clustering the bounding boxes of the training set and the other consists of manually setting prior boxes with the uniform distribution. The prior bounding boxes and the ground truth of the datasets are shown in Fig. 5. The contribution of the individual detector is more uniform in the two settings, but the overall result fails to achieve that of the default setting. Therefore, settings with only the detector with indexed as $\{1\}$ and the default prior bounding boxes in the head is the optimal solution. In order to eliminate the decrease of the back-propagation gradient after removing two detector heads, different α_{detect} are set for the training, and the results are listed in Table VII. When $\alpha_{\text{detect}} = 3.0$, the proposed model can achieve the optimal performance, indicating that for the tasks with a single object scale, it is not conducive to improve the detection performance via the multiscale detection heads.

TABLE VII
EVALUATION WITH DIFFERENT α_{DETECT} SETTINGS

α_{detect}	LAMR↓			Recall↑
	All	Day	Night	All
1.0	6.88	7.44	**5.76**	98.90
2.0	6.75	**6.53**	7.18	**99.11**
3.0	**6.48**	6.65	5.82	98.69
4.0	7.70	8.13	7.12	98.56

The bold minimum values in the columns for Parameter, FLOPs, and LAMR, and the maximum value for Recall.

D. Qualitative Analysis

1) Attention Mechanisms: In the proposed model, the attention mechanism is used to guide the feature fusion. To closely observe the effect of the attention mechanism on the models, we visualize the fused feature maps in Fig. 6. The pixel intensity outside the ROI in the feature maps is significantly suppressed with the attention mechanism. Compared to the model with only one attention mechanism, the model with the dual-dimensional attention mechanism has a better suppression of the irrelevant features.

2) Background Weakening Constraint: As can be seen from Fig. 6, F_1^f and F_2^f can extract significant feature values in the ROIs. However, F_3^f is more abstract, and it is impossible to observe the significant characteristics of the ROIs in the images. Therefore, L_{bw} is only used for the feature maps of $d = \{1, 2\}$ in the two-stream backbone network to avoid information losses. In Fig. 7, we visualize the feature maps with $d = \{1\}$ in the two-stream backbone network before and after applying L_{bw}. Due to the shared convolution operators, the visualizations of the corresponding channel feature maps for the different modalities are similar, indicating the same semantics. Besides, the feature values of the backbone feature maps outside the ROI are significantly suppressed via L_{bw}, which is similar to the effect of the attention mechanism. The combination of the two methods has higher suppression. From the perspective of human visual perception, our improved scheme can effectively filter the noise in the feature maps.

E. Deployment Experiment

The proposed model is deployed on the edge devices to validate its feasibility for the human detection on UAVs. The

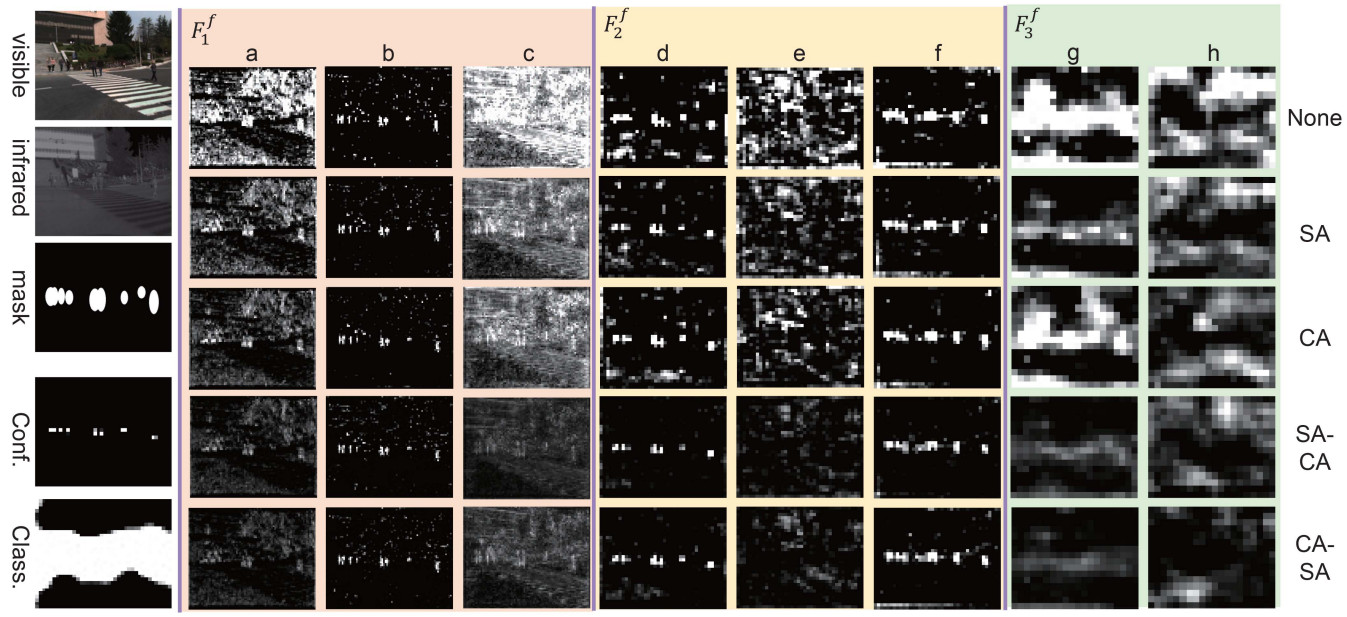

Fig. 6. Illustration of different feature maps. $\{a, b, c\}$, $\{d, e, f\}$, and $\{g, h\}$ are the selected channels from the feature maps F_1^f, F_2^f, and F_3^f, respectively. Conf. and Class. are the visualizations of the confidence and classification results outputted by the detector.

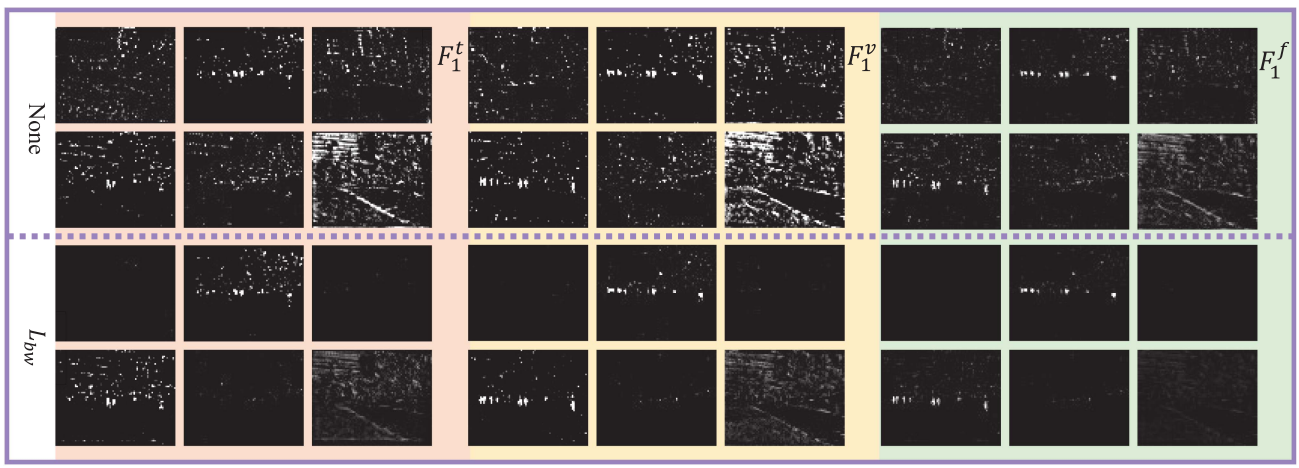

Fig. 7. Illustration of the variations among feature maps F_1^v, F_1^i, and F_1^f before and after the applied L_{bw}.

commonly adopted NVIDIA Jetson TX2 8 GB, AI performance is rated at 1.33 TFLOPS, is used as the edge device, where the model improved based on YOLOv5s can achieve hardware acceleration through the TensorRT technology. During the inference, the accelerated model takes visible and infrared images with a resolution of 640 × 512 as the input, while the batch size is set to 1.

Table VIII presents the running speeds of the model in the two modes, namely MAXQ and MAXN. The former represents the most efficient power utilization, while the latter indicates running at the maximum performance state. The proposed model can achieve a detection speed of nearly 20 frames per second (fps) in MAXQ mode, satisfactorily meeting the detection demand. Moreover, Fig. 8 illustrates the power consumption, the usage of memory, CPU and GPU during the model operation in both modes. The power consumption in the MAXN mode approaches 8 W, meaning it can run at full power for 7 h when powered by a 5000 mAh battery, exceeding the single flight time of 20–50 min under the same power supply. Furthermore, the model consumes approximately

TABLE VIII
MODEL INFERENCE SPEED ON TX2

Pipeline	MAXQ		MAXN	
	Time (ms)	FPS	Time (ms)	FPS
Data loading	11.52	86.75	8.70	114.99
Model inference	58.24	17.17	38.94	25.68
Postprocessing	12.76	78.35	6.52	153.34
All	82.54	12.12	54.16	18.47

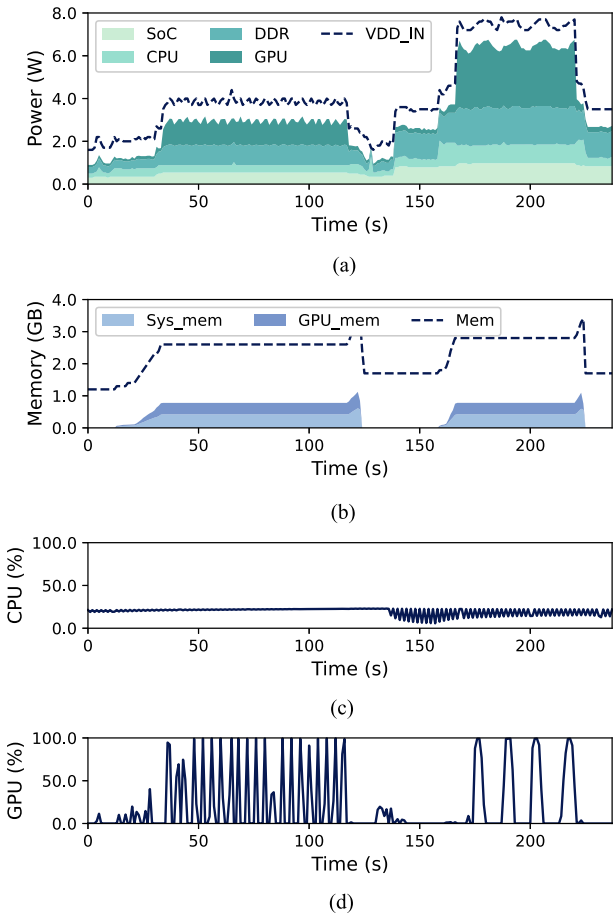

Fig. 8. Device usage of the model running on TX2 in MAXQ and MAXN modes.

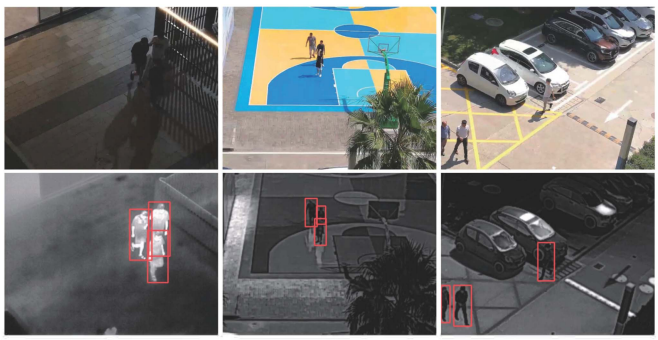

Fig. 9. Human detection with UAV deployment.

1 GB of memory and has an overall memory consumption of around 3 GB, less than half of the total 7.7 GB memory. As shown in Fig. 9, the proposed model has been validated with images collected from UAVs with satisfied human detection results.

IV. Conclusion

In this article, a novel object detection model based on YOLOv5 is proposed with complementary information of the two modalities of the visible and thermal images. The visible-thermal fusion strategy is explored to achieve the feature representation coding and a lightweight model structure is designed with only one detector. The multiscale feature maps of both modalities are extracted via the two-stream backbone with partially shared convolution operators and refined via the background weakening loss. Feature fusion is, thus, performed by exploiting the semantic similarity of the cross-modal features and attention mechanism. Then, the fused multiscale feature maps are integrated to facilitate the information interaction at multiple depths so as to improve the object detection performance. Compared with other models of the same type, the proposed model can achieve superior detection performance on the KAIST dataset with UAV-based experimental verification. Further study will investigate the lightweight implementations based on the refined feature maps and model compress technologies for the object detection.

References

[1] CRED, "Centre for research on the epidemiology of disasters (CRED)," *2021 Disasters in Numbers*. CRED, Brussels, Belgium, 2022. [Online]. Available: https://cred.be/sites/default/files/2021_EMDAT_report.pdf

[2] M. Li, X. Zhao, J. Li, and L. Nan, "ComNet: Combinational neural network for object detection in UAV-borne thermal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6662–6673, Aug. 2021.

[3] Y. Sun, S. Abeywickrama, L. Jayasinghe, C. Yuen, J. Chen, and M. Zhang, "Micro-doppler signature-based detection, classification, and localization of small UAV with long short-term memory neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6285–6300, Aug. 2020.

[4] Y. Miao, Y. Tang, B. A. Alzahrani, A. Barnawi, T. Alafif, and L. Hu, "Airborne LiDAR assisted obstacle recognition and intrusion detection towards unmanned aerial vehicle: Architecture, modeling and evaluation," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4531–4540, Jul. 2021.

[5] F. Zhao, W. Zhao, L. Yao, and Y. Liu, "Self-supervised feature adaptation for infrared and visible image fusion," *Inf. Fusion*, vol. 76, pp. 189–203, 2021.

[6] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5363–5371.

[7] T. Liu, K.-M. Lam, R. Zhao, and G. Qiu, "Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 315–329, Jan. 2022.

[8] J. U. Kim, S. Park, T. Kim, and Y. M. Ro, "Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 1157–1165.

[9] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 787–803.

[10] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5127–5137.

[11] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 73.1–73.13.

[12] K. Dasgupta, A. Das, S. Das, U. Bhattacharya, and S. Yogamani, "Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15940–15950, Sep. 2022.

[13] Y. Zhuang, Z. Pu, J. Hu, and Y. Wang, "Illumination and temperature-aware multispectral networks for edge-computing-enabled pedestrian detection," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 3, pp. 1282–1295, May/Jun. 2022.

[14] H. Zhang, E. Fromont, S. Lefèvre, and B. Avignon, "Low-cost multispectral scene analysis with modality distillation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 803–812.

[15] S. Meng and Y. Liu, "Multimodal feature fusion YOLOv5 for RGB-T object detection," in *Proc. China Autom. Congr.*, 2022, pp. 2333–2338.

[16] M. Li, B. Liu, J. Sun, G. Zhang, and W. Su, "Multimodal feature fusion YOLOv5 for RGB-T object detection," in *Proc. Int. Conf. Artif. Intell. Intell. Inf. Process.*, 2022, vol. 12456, pp. 298–303.
[17] H. Fu et al., "LRAF-Net: Long-range attention fusion network for visible-infrared object detection," *IEEE Trans. Neural Netw. Learn Syst.*, to be published, doi: 10.1109/TNNLS.2023.3266452.
[18] Y. Xie, L. Zhang, X. Yu, and W. Xie, "YOLO-MS: Multispectral object detection via feature interaction and self-attention guided fusion," *IEEE Trans. Cogn. Devlop. Syst.*, to be published, doi: 10.1109/TCDS.2023.3238181.
[19] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[20] 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 Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 390–391.
[21] 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.*, 2017, pp. 2117–2125.
[22] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.
[23] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.
[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[25] 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 Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11531–11539.
[26] A. S. Bosman, A. Engelbrecht, and M. Helbig, "Visualising basins of attraction for the cross-entropy and the squared error neural network loss functions," *Neurocomputing*, vol. 400, pp. 113–136, 2020.
[27] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.
[28] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.
[29] L. Zhang et al., "Cross-modality interactive attention network for multispectral pedestrian detection," *Inf. Fusion*, vol. 50, pp. 20–29, 2019.
[30] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 225–235.
[31] J. Kim, H. Kim, T. Kim, N. Kim, and Y. Choi, "MLPD: Multi-label pedestrian detector in multispectral domain," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7846–7853, Oct. 2021.

Xiongxin Zou received the B.S. degree in mechanical engineering from Liaoning Technical University, Fuxin, China, in 2021. He is currently working toward the M.S. degree in control science and engineering with the Hunan University of Science and Technology, Xiangtan, China.

His research interests include machine learning and computer vision.

Tangle Peng received the B.S. degree in electronics and information engineering from Hengyang Normal University, Hengyang, China, in 2021. He is currently working toward the M.S. degree in electronics and communication engineering with the Hunan University of Science and Technology, Xiangtan, China.

His research interests include UAV navigation and path planning.

Yimin Zhou (Member, IEEE) received the Ph.D. degree in control engineering from the University of Oxford, Oxford, U.K., in 2008.

She is currently a Full Professor with the Shenzhen Institute of Advanced Technology, Chinese Academy Sciences, Shenzhen, China. Her research interests include nonlinear control, fault diagnosis, robotics, machine learning, and energy management.