2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

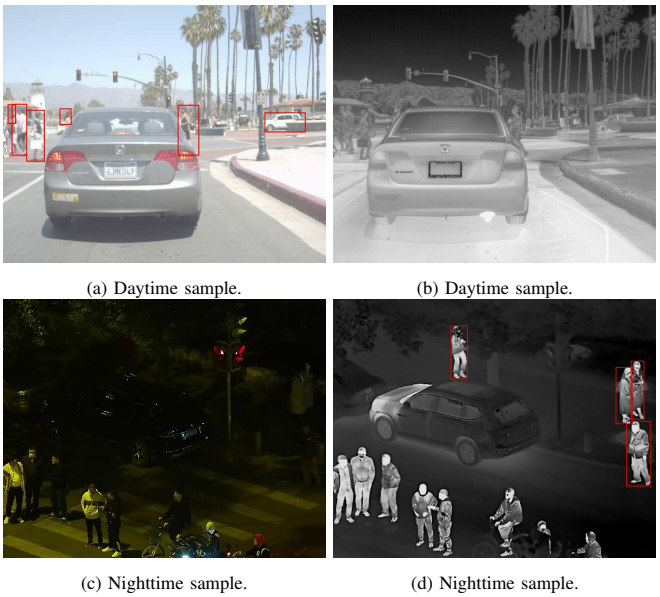
DCSANet: Dual Cross-channel and Spatial Attention Make RGB-T Object Detection Better

Xiaoxiong Lan[1], Shenghao Liu[1], Zhiyong Zhang[1], Changzhen Qiu[1,†]

***Abstract*— Multimodal image pairs can make object detection more reliable in challenging environments, so RGB-T object detection has gained extensive attention over the past decade. To alleviate the complementarity of the visible and thermal modality, we propose a novel lightweight Feature Enhancement-fusion Module (FEM), which is composed of the Channel Enhancement-fusion Unit (CEU) and Spatial Enhancement-fusion Unit (SEU) by extending the attention mechanism to operate on two modalities. CEU is used to exploit the complementarity and alleviate the data imbalance by combining internal and global channel attention. Additionally, SEU is utilized to guide the model to pay more attention to the regions of interest. By incorporating FEM, enhanced and fused features are obtained, leading to improved performance. The effectiveness and generalizability of FEM are validated by two public datasets and our proposed DCSANet achieves competitive performance while maintaining high speed (+%7.0 on LLVIP and +1.2% on FLIR in mAP). Moreover, we conducted ablation experiments to verify the effectiveness of the proposed operators.**

I. INTRODUCTION

Object detection is one of the three basic tasks in computer vision [1], which has important applications in autonomous driving [2], robot vision [3], video surveillance [4], and so forth. However, existing methods are highly dependent on the quality of visible (RGB) images. Due to the complexity of the real world, there are adverse factors such as illumination and weather, which make it difficult to design a high-performance detector. The traditional object detection algorithms rely on manually designed feature extraction methods, which are subjective and yield limited features. Consequently, they are inadequate for handling object detection tasks in various complex scenarios. Deep neural networks (DNN) can extract more effective features by automatically learning on large-scale data and have achieved some success. However, it is vulnerable to illumination or adverse weather if it only rely on the visible modality. To solve this problem, some researchers study multimodal object detection by combining the visible images and thermal images. Because the visible images and thermal images have a wide range of applications due to their inherent complementarity (Fig. 1). By introducing thermal modal as a supplement, more information can be obtained. As a result, RGB-T (RGB-Thermal) object detection in complex environments has gained extensive attention over the past decade.



(a) Daytime sample. (b) Daytime sample.

(c) Nighttime sample. (d) Nighttime sample.

Fig. 1. The visible image (left column) and thermal image (right column). During the day, visible images provide more information; During the night, thermal images provide more information. The two modalities are complementary, which can provide effective information for robust object detection from day to night.

For RGB-T object detection, how to fully explore the complementarity of the two modalities (RGB-Thermal image pairs) is a key problem. To solve it, a lot of work has been put forward. Hwang et al. [5] presented the multi-spectral aggregated channel features (ACF) for pedestrian detection to simultaneously handle RGB-Thermal image pairs and it is proved that visible-thermal image pairs fusion object detection can obtain better performance than a single modality of visible-only. Liu et al. [34] conducted an in-depth study on the structure of the fusion network and divided the fusion methods into four categories (early, middle, late, and result fusion). By conducting experiments on the paired visible-thermal pedestrian detection dataset after alignment, it was proved that the feature-level fusion in the middle stage of the network is the best. Zheng et al. [6] effectively fused the features of visible and thermal modality by a gated fusion unit and then the SSD [7] algorithm is used to detect the target, which achieves high detection speed while maintaining good accuracy. However, the above methods don't consider the differences and importance between the two modalities, which is simply done by concatenating/adding the feature

*This work was not supported by any organization
[1]School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China
†Corresponding author is Changzhen Qiu qiuchzh@mail.sysu.edu.cn

maps of the two modalities. To generate more representative features, Zhang et al. [8] presented a feature re-weighting scheme to select more reliable features and suppress the useless ones. This method has high detection accuracy and good robustness. In work [9], Zhou et al. proposed the concept of modality differentiation to address the issue of data unbalance between two modalities through the idea of differential circuits, which can realize information exchange of two modalities and obtain the weight of the two modalities through illumination perception.

However, there is still a modal imbalance problem due to the image-forming principle of visible and thermal images, resulting in the important cues of object characteristics being different in the two modalities. For example, pedestrian features like clothing are key in the visible modality, yet absent in thermal image [9]. Meanwhile, previous studies ignored the differences within a single modality. And the detection speed can not meet the requirements of real-time.

In this paper, we propose a feature enhancement-fusion module using dual cross-channel and spatial attention (DCSA) to solve the problem. Our contributions are summarized as follows:

- We design a dual cross-channel attention unit named CEU by combining intra-modality and inter-modality channel attention, which can guide the network to alleviate the channel difference within a single modality and the overall difference between the two modalities.
- We design FEM by combining CEU and SEU, a lightweight plug-and-play module that can be used for adaptive fusion and enhancement of the features from the two modalities.
- We conduct extensive experiments and achieved good results on two datasets, verifying that FEM can effectively alleviate the problem of data imbalance. And our model achieves the speed of real-time.

The structure of remainder of the paper is as follows: In section II, we introduce the related work of single-modal and multi-modal object detection. In section III, the overall architecture of the network proposed in this paper is first introduced, and then the details of the FEM are explained. In section IV, we present the experimental results and analyze them. In section V, we make a summary of this paper.

II. Related Work

We present a review of recent works related to our work.

A. Object Detetion

Object detection is an important computer vision task used to detect specific objects in images, which has made long-term progress in the past few decades [1]. According to the methods adopted, object detection algorithms can be divided into traditional object detection algorithms and deep learning-based object detection algorithms.

Traditional algorithms rely on manually designed feature extraction methods, which include V-J detector [11], [12], HOG detector [13], DPM detector [14], and so on. Although researchers have done much optimization, they are limited and cumbersome because of hand-crafted features and detection accuracy is not high enough. Meanwhile, they are difficult to optimize and accelerate, which makes it difficult to achieve real-time detection.

Recently, with the rise of deep learning, object detection algorithms based on deep learning have become mainstream. It includes the one-stage method and the two-stage method. The two-stage detector is represented by the R-CNN series, which includes R-CNN [15], Fast R-CNN [16], and Faster R-CNN [17]. Because of the intermediate region proposal network (RPN) in the two-stage detector, it is difficult to meet the real-time application requirements. Therefore, some one-stage detectors were proposed to obtain prediction results directly from images, represented by OverFeat [18], SSD [7], RetinaNet [19], YOLO [10], etc, which have achieved good performance. However, there are certain limitations of object detection algorithms that only rely on visible images. For example, in the case of low light, strong light, and illumination change, the quality of the RGB image will deteriorate sharply, which will seriously affect the performance of detectors, and even lead to a crash.

B. RGB-T Object Detetion

To address the problem, researchers have attempted to incorporate thermal modality as supplementary, leading to the emergence of RGB-T object detection. Considering the challenge of effectively using the complementarity of two modalities, many approaches have been proposed, which can be categorized into early fusion, middle fusion, and late fusion. Late fusion (decision level) is infrequently utilized. Therefore, it will not be discussed.

*1) **Early fusion**:* Early fusion includes simple concatenation and fusion network integration. . In the first category, Zhang et al. [20] concatenated the 3-channel visible image with the 1-channel thermal image by channel dimension into a 4-channel image and directly used it for object detection. In the second category, Liu et al. [21] employed a GAN [22] network to fuse pairs of visible and thermal images into a new 3-channel RGB image for object detection.

This approach allows for separate execution of image fusion and object detection, providing flexibility to employ different methods for each part. Traditional methods as well as deep learning methods can be used for two parts directly. However, compared to middle fusion, it lacks correction between two parts. Consequently, fused images may exhibit good visual performance but suffer from potential information loss during the fusion, which will degrade the performance of object detection. The optimization of this approach becomes challenging due to the separate training requirements for two parts.

*2) **Middle fusion**:* With the advent of deep learning, this paradigm has become prevailing, owing to its efficiency in extracting latent features. This approach integrates the extracted features during the detection process, ultimately enabling enhanced performance in the object detection task. It is comprehensive end-to-end, so feature extraction and

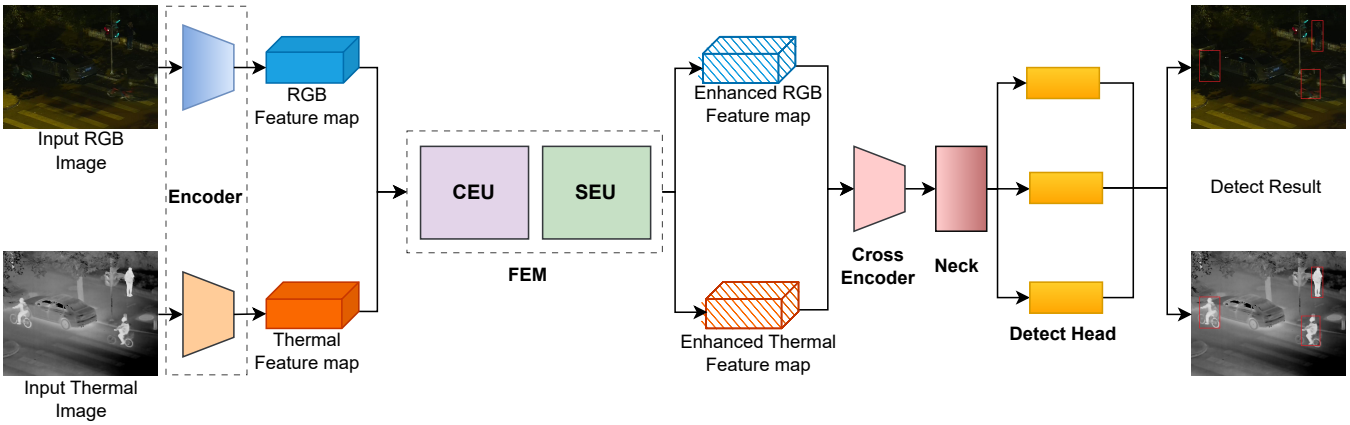


Fig. 2. Framework overview of the proposed DCSANet model. The images of the two modalities are automatically extracted by the Encoder, and then the extracted features are fused and enhanced by FEM module composed of CEU unit and SEU unit. High-level semantic features are further extracted by the CrossEncoder, and then sent it to the detector to obtain the final detection results.

object detection mutually reinforce each other, leading to better object detection performance.

Geng et al. [23] proposed a dual-branch object detection network based on Faster R-CNN [17], where paired visible-thermal images were fed into the VGG-16 backbone network for individual feature extraction, followed by fusion of the output features. However, since the used detector is a two-stage network, the speed is not fast. Based on the one-stage object detector SSD [7], Zheng et al. fuse the feature maps by using a mixture of stacked fusion and gated fusion [6]. Although the speed is fast, the detection accuracy is not high enough. Fang et al. extracted the common and unique features of multi-modal images and enhanced the unique features to improve the efficiency of feature fusion [24]. To better fuse the features from two modalities, Zhang et al. proposed a guided attention feature fusion method, which learns adaptive weight and fuses the multi-modal features by combining two attention modules [35]. Song et al. combined multi-scale feature extraction into a multi-modal feature fusion network for multi-scale pedestrian detection [25]. To solve the problem that different modal features are generated independently from each other, Hua et al. [26] proposed a multi-modal feature cross-guided learning mechanism for modeling long-term modal dependencies of two modalities, enhancing interactions between multi-modal feature generation modules and reducing inter-modal differences.

In summation, only simple concatenation or combination of the features [5], [34], [27], [28] can not fully exploit the complementarity between two modalities. It is necessary to guide the model to focus on effective features through the attention mechanism to solve the problem of data imbalance. At the same time, we need to design lightweight fusion methods to maintain high detection speed.

III. Methodology

The framework of our proposed method is shown in Fig.2.

A. Framework overview

The DCSANet inputs paired visible-thermal images simultaneously to get detection results. We use YOLOv5 as the baseline. As shown in Fig.2, the first three layers of the network are employed as an Encoder to extract the features of the visible-thermal image pair as a two-stream backbone, and FEM is used to fuse and enhance the feature from two modalities. Finally, the fused features are fed into the CrossEncoder constituted by the subsequent part of the network to refine features. The refined features are utilized by the detection head to generate final results.

We construct a siamese network as an Encoder to extract intra-modal deep semantic features and the CrossEncoder module is used to further refine the fused and enhanced features. The RGB and Thermal backbone are shown in Fig.2, and their weights are not shared. The Encoder and CrossEncoder are stacked with CBS (Conv, BN, SiLU) and CSP block(Fig.4), with short-cut only in Encoder.

Denote the input of encoder as x_i, we get output as $\hat{x}_i$.

$$\hat{x}_i = \Phi_i(x_i) \tag{1}$$

$\Phi_i(x)$ represents the computation performed by Encoder or CrossEncoder. When subscript $i = 0$ is the RGB branch (x_{rgb}), $i = 1$ is the Thermal branch (x_t), and $i = 2$ is the concatenated feature maps ($x_2 = F_{fused}$).

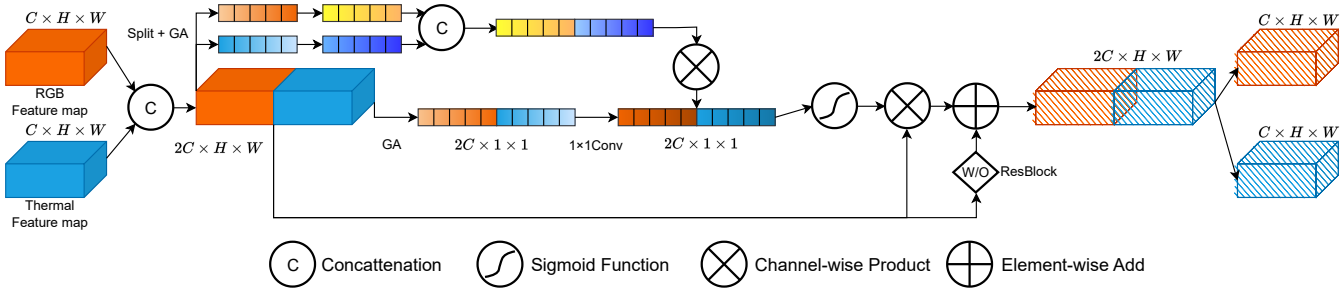
The key problem is how to effectively fuse and enhance information from visible-thermal image pairs. The problem can be defined as follows:

$$\mathbf{F}_{\text{fused}} = \mathcal{F}\left(\hat{x}_{rgb}, \hat{x}_t\right) = \mathcal{F}\left(\Phi_{rgb}\left(\mathbf{I}_{rgb}\right), \Phi_t\left(\mathbf{I}_t\right)\right) \tag{2}$$

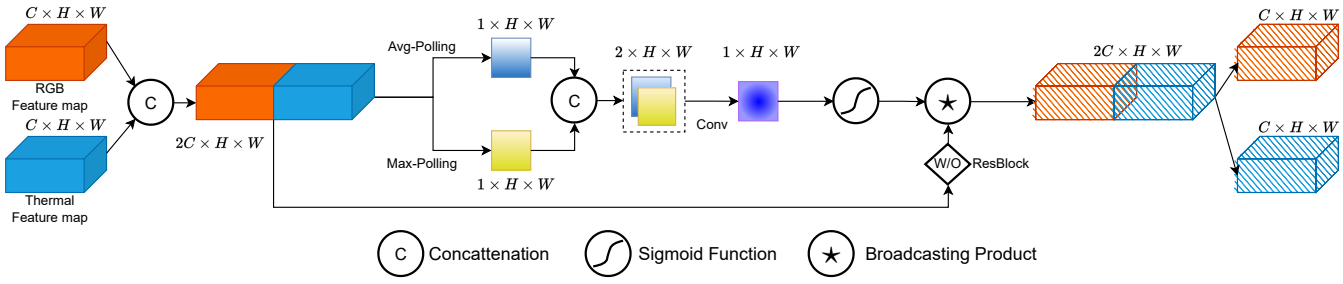
The extracted features from the RGB ($\mathbf{I}_{rgb}$) and thermal image ($\mathbf{I}_t$) are feeded into the fusion-enhancement module $\mathcal{F}$ to obtain fused feature $\mathbf{F}_{\text{fused}}$. How to design the fusion and enhancement function $\mathcal{F}$ is the problem we focus on.

B. Feature Enhancement-fusion Module

Inspired by methods [29], [30], we focus on enhancing the effectiveness features by incorporating both channel attention and spatial attention mechanisms. The channel attention can obtain the importance of each channel so that the model can focus on some more effective channels. Meanwhile, since the contribution of regions in the image to the task is not



(a) Illustration of the CEU.



(b) Illustration of the SEU.

Fig. 3. The w/o ResBlock means whether the input feature map and the output feature map are added. (a) CEU is used to effectively address differences within and between the two modalities. (b) SEU is used to make the model pay more attention to regions where objects may exist

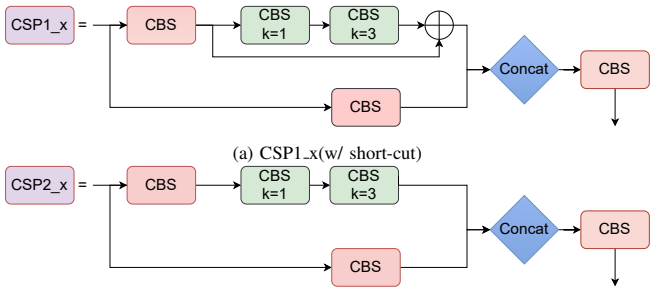


Fig. 4. CSP1_x and CSP2_x have the same structure except for short-cut (The subscript x represents the number of uses.).

equal, and some regions even have interference. The spatial attention can pay more attention to the important regions. Based on the above points, we propose FEM, which is composed of CEU and SEU (Fig.3).

Different from previous approaches, our combined attention is extended to operate on feature maps of two modalities. Since directly using the channel attention on the concatenated feature maps doesn't consider the channel imbalance problem inside each modality. Therefore, we use an internal channel attention on each modality, which is used to adjust the global attention weights, so that the data imbalance can be alleviated. Additionally, spatial attention is utilized to further guide the model to pay more attention to regions of interest. By incorporating the FEM, the fused and enhanced features are obtained.

Reference [29] proposed Channel Attention (CA), which uses two fully connected layers for encoding and decoding. Inspired by this idea, we design the CEU to enhance the feature representation by utilizing interdependencies within cross-modality and intra-modality. The parameters of the three branches are trained independently. The Inter-modal Channel Attention (Inter-CA) is used to alleviate the data imbalance of two modalities, while the Intra-modal Channel Attention (Intra-CA) is employed to determine the channel importance within a single modality and suppress noise.

We denote the input feature map as $x^c_{rgb}, x^c_t \in \mathbb{R}^{C\times H\times W}$. After the attention module, we can get the output as $\hat{x}^c_{rgb}, \hat{x}^c_t \in \mathbb{R}^{C\times H\times W}$. We concatenate x^c_{rgb} and x^c_t to form joint feature map $U^{2c} \in \mathbb{R}^{2C\times H\times W}$.

In CEU (Fig.3(a)), CA scores are divided into three branches for calculation, which can be described as:

$$\hat{x}^c_{rgb}, \hat{x}^c_t = \mathrm{CEU}(U^{2c}) = \mathrm{Split}[U^{2c} \otimes W_c + U^{2c}] \quad (3)$$

where

$$\mathrm{CA}(x) = \varepsilon(\omega(\delta(x^c))) \quad (4)$$

$$W_c = \varepsilon(\mathrm{CA}(U^{2c}) * \mathrm{Concat}(\mathrm{CA}(x^c_{rgb}), \mathrm{CA}(x^c_t))) \quad (5)$$

where $\otimes$ denotes Channel-wise Product, δ denotes global pooling, ω denotes 2-layer 1×1 2-D convolution, ε denotes Sigmoid function, $Catcat$ denotes concatenating features along channel. After adjusting by three CA modules, we add it with the original feature map. The $Split$ splits the feature maps equally along the channel dimension.

In SEU (Fig.3(b)), we suppress background noises by utilizing the inter-spatial relationship to generate a spatial attention map. We utilize two parameter-free operations, namely Channel-wise Average Pooling (CAP) and Channel-wise Max Pooling (CMP), which can be mathematically represented as follows:

$$\hat{x}^c_{rgb}, \hat{x}^c_t = \mathrm{SEU}(U^{2c}) = \mathrm{Split}[U^{2c} \circledast W_s] \quad (6)$$

where

$$W_s = \varepsilon(\varphi(Concat(\mathrm{CAP}(U^{2c}), \mathrm{CMP}(U^{2c})), W)) \quad (7)$$

$$\mathrm{CAP}(U^{2c}) = \frac{1}{2c}\sum_{c=0}^{2c-1} u_{ij}^{c} \quad (8)$$

$$\mathrm{CMP}(U^{2c}) = \max\left(u_{ij}^{0}, \ldots, u_{ij}^{2c-1}\right) \quad (9)$$

where $\circledast$ denotes Broadcasting Product, φ denotes a 2-D convolution operation (W is the parameter.).

Here, the input and output tensors of CEU and SEU have the same size. We add the residual structure to increase the flow of information, which is referred to as Pre-Res in CEU and Post-Res in SEU. Through the two units, the complementarity between the two modalities can be fully used to get more efficient feature representation.

IV. Experiments and results

We conduct experiments on two datasets, which are commonly used in RGB-T object detection.

A. Experimental Setting

The LLVIP [31] is a visible-thermal paired dataset for low-light vision. This dataset contains 15,488 image pairs taken in very dark scenes, which are strictly aligned in time and space. The same image pairs share the same annotation. The dataset contains 41,579 labels of "person", 33,648 labels in the training set, and 7,931 labels in the test set.

The FLIR dataset [32] is collected for autonomous driving. However, the original dataset is unaligned. We utilize the aligned-FLIR [38] (It is simply denoted as FLIR.), which contains 5,142 well-aligned visiblethermal image pairs and involves four types of objects, including 8,987 people, 20,608 cars, 2,566 bicycles, and 95 dogs. We removed all dog labels as they are not adequate for training.

We use YOLOv5 [33] as the baseline. The performance comparison between baseline and DCSANet is under the premise of similar parameters and computation. The code is implemented in Pytorch and we trained the model on a NVIDIA 3090 GPU. Various data augmentation methods such as Mosaic are used. We choose SGD as the optimizer, and the model is trained for 300 epochs, with a batch-size of 32, image-size of 604×512, and initial learning-rate of 0.01. To speed up the training process, warm-up and cosine annealing algorithms are used to adjust the learning rate.

We use AP50, mAP, and inference time as evaluation metrics. RGB+Thermal in baseline (TABLE I) refers to fusion by adding the two modalities with equal weight. We contrast with several state-of-the-art models, including Halfway fusion [34], GAFF [35], ProbEn [36], and CASS [37].

B. Experimental Results

1) ***Quantitative Results****:* We compare DCSANet with several SOTA models.The results are shown in TABLE I.

TABLE I shows that the AP50 and mAP achieved by our model on the LLVIP surpass other models. Since the collection time of the dataset is at night with low-light [31], the visible modal provides little complementary information. As a result, our algorithm does not achieve much performance improvement compared to thermal modality, +0.4% for AP and +0.5% in mAP. However, the performance of other models is even reduced compared with thermal modal baseline, because they can not make full use of the data complementarity. However, DCSANet considers the differences within and between the two modalities and mitigates this problem through the CEU. Meanwhile, we use the SEU to make the model pay more attention to the regions where the object exists. As a result, our model outperformers Halfway Fusion [34], GAFF [35], ProbEn [36], CSSA [37], by 5.2%, 2.6%, 3.2%, 2.3% on AP50 and by 11.1%, 10.4%, 14.7%, 7.0% on mAP, respectively. Our model achieves significant performance improvement over unimodal on the FLIR [38], which outperformers Halfway Fusion [34], GAFF [35], ProbEn [36], CSSA [37], by 8.3%, 5.2%, 4.3%, 0.6% on AP50 and by 6.7%, 5.1%, 4.6%, 1.2% on mAP, respectively. Compared to the baseline by simple addition, it achieves +%8.2 in AP50 and +%5.6 in mAP.

TABLE I shows that our model only takes 29.8ms to complete the inference of a single visible-thermal image pair, which meets the real-time requirement. Compared with other models, our model is faster than Halfway Fusion [34] and GAFF [35], and achieves comparable speed with ProbEn [36], CSSA [37]. That's because our feature fusion method is a super lightweight operation.

The results show that FEM can be used to alleviate the problem of data imbalance, provide more useful features, and allow the model to focus on the region where objects may be present. When compared with the single-modal and multi-modal object detection models, our proposed DCSANet achieves competitive performance.

2) ***Qualitative Results****:* We also perform a qualitative analysis based on the FLIR [38]. The results are shown in Fig.5, which show that in the dark environment, the quality of visible images decreases sharply, and it is difficult to obtain satisfactory detection results. However, the combination of the visible and thermal modalities can overcome the influence of adverse environments such as low-light and strong-light, and achieve better results than the CSSA [37].

Compared to CSSA [37], our regression box is more accurate on the same object. However, our algorithm successfully detects the objects that are not in the label but actually exist in the scene, which indicates that our algorithm actually learns the general feature representation and has better generalization performance. In addition, the false detection rate of our model is lower.

C. Ablation Study

The ablation study is implemented to evaluate the influence of different operations on the FLIR [38]. Five operators are considered. The checkmark indicates using the operator.

In TABLE II, we can see that both CEU and SEU can enhance the ability to fuse dual-modal features, and the combination of the CEU and SEU with Pre-Res can achieve better performance. We see that adding Intra-CA and

TABLE I: The evaluation results on LLVIP and FLIR measured by AP in percentage, Inference time, and FPS

Method	Modality	LLVIP		FLIR		Inference time(ms)↓	FPS ↑
		AP50(%) ↑	mAP(%) ↑	AP50(%) ↑	mAP(%) ↑		
baseline[G. Jocher et al., 2020 [33]]	RGB	92.2	53.0	65.6	31.2	18.6	53.76
baseline[G. Jocher et al., 2020 [33]]	Thermal	96.2	65.7	77.3	38.2	18.6	53.76
baseline[G. Jocher et al., 2020 [33]]	RGB + Thermal	95.3	63.7	71.6	36.9	18.6	53.76
Halfway fusion[Liu et al., 2016 [34]]	RGB + Thermal	91.4	55.1	71.5	35.8	42.0	23.81
GAFF[Zhang et al., 2021 [35]]	RGB + Thermal	94.0	55.8	74.6	37.4	61.0	16.39
ProbEn[Chen et al., 2022 [36]]	RGB + Thermal	93.4	51.5	75.5	37.9	**25.0**	**40.00**
CSSA[Cao et al., 2023 [37]]	RGB + Thermal	94.3	59.2	79.2	41.3	31.0	32.26
DCSANet (ours)	RGB + Thermal	**96.6**	**66.2**	**79.8**	**42.5**	29.8	33.56

Fig. 5. Visualization of results with the optimal contrast model CSSA [37]. Each row is a sample. The person, car, and bicycle are marked with bounding boxes in blue, yellow, and green. The red box is used to mark detection errors (false alarms/missed detections), and orange box is used to mark objects that are not in labels but are actually present and successfully detected in the image.

Pre-Res operations to CEU is able to improve the overall performance. However, adding Post-Res to the SEU will reduce the performance. We believe that this is because the channel attention in CEU is used to adjust the difference between the feature maps of the two modalities. The spatial attention in SEU is directly used to derive the weight of the region where the target may be, so Post-Res will interfere with the focus.

V. Conclusion

In this paper, we propose a lightweight module named FEM for RGB-T object detection to fuse and enhance feature maps from visible and thermal modalities. The detector using FEM achieves good performance on both dataset, which verifies the effectiveness of our proposed module. Furthermore, the effectiveness and necessity of the CEU and

TABLE II: The results of ablation study on FLIR dataset.

Operation					Result	
CEU	SEU	Intra-CA	Pre-Res	Post-Res	AP50	mAP
					75.5	39.9
✓					77.6	40.8
✓		✓			78.8	42.1
✓	✓	✓	✓	✓	78.5	42.2
✓	✓		✓		78.0	42.2
✓	✓	✓	✓		**79.8**	**42.5**

SEU for fusion and enhancement of features are explored through ablation experiments. In addition, FEM module can also be extended to the fusion of other modality data (such as depth images and radar images).

REFERENCES

[1] Z. Zou, K. Chen, Z. Shi, Y. Guo and J. Ye, "Object detection in 20 years: A survey," Proceedings of the IEEE, March 2023, vol. 111, no. 3, pp. 257-276.

[2] Z. Chen and X. Huang, "Pedestrian Detection for Autonomous Vehicle Using Multi-Spectral Cameras," IEEE Transactions on Intelligent Vehicles, June 2019, vol. 4, no. 2, pp. 211-219.

[3] D. H. Dos Reis, D. Welfer, M. A. De Souza Leite Cuadros, and D. F. T. Gamarra, "Mobile Robot Navigation Using an Object Recognition Software with RGBD Images and the YOLO Algorithm," Applied Artificial Intelligence, Nov. 2019, vol. 33, no. 14, pp. 1290-1305.

[4] C. T. Selvi and J. Amudha, "Automatic Video Surveillance System for Pedestrian Crossing Using Digital Image Processing," Indian Journal of Science and Technology, Jan. 2019, vol. 12, no. 02, pp. 1-6.

[5] S. Hwang, J. Park, N. Kim, Y. Choi and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015, pp. 1037-1045.

[6] Y. Zheng, I. H. Izzat, and S. Ziaee, "GFD-SSD: Gated Fusion Double SSD for Multispectral Pedestrian Detection,"arXiv preprint arXiv: 1903.06999, 2019.

[7] W. Liu et al., "SSD: Single Shot MultiBox Detector," Computer Vision-ECCV 2016, 2016, vol. 9905, pp. 21-37.

[8] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei and Z. Liu, "Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection," in 2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 5126-5136.

[9] K. Zhou, L. Chen, and X. Cao, "Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems," Lecture Notes in Computer Science, Jan, 2020, pp. 787-803.

[10] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788.

[11] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in 2001 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, USA, 2001, pp. I-I.

[12] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," International Journal of Computer Vision (IJCV), May 2004, vol. 57, no. 2, pp. 137-154.

[13] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, USA, 2005, vol. 1, pp. 886-893.

[14] P. Felzenszwalb, D. McAllester and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8.

[15] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA, 2014, pp. 580-587.

[16] R. Girshick, "Fast R-CNN," in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1440-1448.

[17] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, June 2017, vol. 39, no. 6, pp. 1137-1149.

[18] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Yann LeCun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," Dec. 2013, arXiv preprint arXiv:1312.6229.

[19] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollr, "Focal Loss for Dense Object Detection," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2999-3007.

[20] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li and Q. Du, "SuperYOLO: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery," IEEE Transactions on Geoscience and Remote Sensing, 2023, vol. 61, pp. 1-15.

[21] J. Liu et al., "Target-aware Dual Adversarial Learning and a Multi-scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection," in 2022 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), LA, USA, 2022, pp. 5792-5801.

[22] I. Goodfellow et al., "Generative adversarial networks," Communications of the ACM, Oct. 2020, vol. 63, no. 11, pp. 139-144.

[23] K. Geng et al., "Low-observable Targets Detection for Autonomous Vehicles based on Dual-modal Sensor Fusion with Deep Learning Approach," Proceedings of the Institution of Mechanical Engineers, Aug. 2019, vol. 233, no. 9, pp. 2270-2283.

[24] F. Qingyun and W. Zhaokui, "Cross-modality Attentive Feature Fusion for Object Detection in Multispectral Remote sensing imagery," Pattern Recognition, Oct. 2022, vol. 130, p. 108786.

[25] X. Song, S. Gao, and C. Chen, "A Multispectral Feature Fusion Network for Robust Pedestrian Detection," Alexandria Engineering Journal, Feb. 2021, vol. 60, no. 1, pp. 73-85.

[26] C. Hua, M. Sun, Y. Zhu, Y. Jiang, J. Yu, and Y. Chen, "Pedestrian Detection Network with Multi-modal Cross-guided Learning," Digital Signal Processing, Apr. 2022, vol. 122, p. 103370.

[27] D. Knig, M. Adam, C. Jarvers, G. Layher, H. Neumann and M. Teutsch, "Fully Convolutional Region Proposal Networks for Multispectral Person Detection," IN 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, USA, 2017, pp. 243-250.

[28] L. Ding, Y. Wang, R. Laganire, D. Huang, X. Luo, and H. Zhang, "A Robust and Fast Multispectral Pedestrian Detection Deep Network," Knowledge-Based Systems, Sep. 2021, vol. 227, p. 106990.

[29] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," In 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, 2018, pp. 7132-7141.

[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," Computer Vision-ECCV 2018, 2018, pp. 3-19.

[31] X. Jia, C. Zhu, M. Li, W. Tang and W. Zhou, "LLVIP: A Visible-infrared Paired Dataset for Low-light Vision," in 2021 IEEE International Conference on Computer Vision Workshops (ICCVW), Montreal, Canada, 2021, pp. 3489-3497.

[32] Free Teledyne FLIR thermal dataset for algorithm training, 2018. https://www.flir.ca/oem/adas/adas-dataset-form.

[33] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode, Yonghye Kwon, et al, "YOLOv5 by Ultralytics: Bug Fixes and Performance Improvements," 2020, doi: 10.5281/zenodo.4154370.

[34] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral Deep Neural Networks for Pedestrian Detection," arXiv preprint arXiv: 1611.02644, 2016.

[35] H. Zhang, E. Fromont, S. Lefevre and B. Avignon, "Guided Attentive Feature Fusion for Multispectral Pedestrian Detection," in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, USA, 2021, pp. 72-80.

[36] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal Object Detection via Probabilistic Ensembling," Computer Vision-ECCV 2022, pp. 139-158.

[37] Y. Cao, J. Bin, J. Hamari, E. Blasch and Z. Liu, "Multimodal Object Detection by Channel Switching and Spatial Attention," in 2023 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, Canada, 2023, pp. 403-411.

[38] H. Zhang, E. Fromont, S. Lefevre and B. Avignon, "Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks," in 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2020, pp. 276-280.