MCANet: Multiscale Cross-Modality Attention Network for Multispectral Pedestrian Detection

Xiaotian Wang[1,2(✉)], Letian Zhao[1,2], Wei Wu[1,2], and Xi Jin[1,2]

[1] State Key Laboratory of Particle Detection and Electronics, University of Science and Technology of China, Hefei, China
wxtdsg@mail.ustc.edu.cn
[2] Institute of Microelectronics, Department of Physics, University of Science and Technology of China, Hefei, China

Abstract. Multispectral pedestrian detection is an important and challenging task, that can provide complementary information of visible images and thermal images for high-precision and robust object detection results. To fully exploit the different modalities, we propose a Multiscale Cross-Modality Attention (MCA) module to efficiently extract and fuse features. In this module, the transformer architecture is used to extract features of two modalities. Based on these features, we design a novel spatial attention mechanism that can adaptively enhance object details and suppress background. Finally, the features of each branch are fused using the channel attention mechanism and sent to the detector. To verify the effect of the MCA module, we propose the MCANet. The MCA modules are embedded at different depths of the two-stream network and interconnected to share multiscale information. Extensive experimental results demonstrate that MCANet achieves state-of-the-art detection accuracy on the challenging KAIST multispectral pedestrian dataset.

Keywords: Multispectral detection · Cross-modality feature fusion · Attention mechanism

1 Introduction

In recent years, with the rapid development of computer technology, major breakthroughs have been made in many fields of computer vision. As one of the important object detection tasks, pedestrian detection has a wide spectrum of application prospects, including autonomous driving, surveillance, and search and rescue [1,2]. In real-world applications, the environment is often complex and changeable, such as rain, fog, and low light. In these cases, pedestrian detectors that only focus on visible images have difficulty achieving sufficient accuracy. Thermal images can overcome these difficulties because they can be imaged

D.-T. Dang-Nguyen et al. (Eds.): MMM 2023, LNCS 13833, pp. 41–53, 2023.
https://doi.org/10.1007/978-3-031-27077-2_4

clearly without relying on illumination conditions. However, thermal images lose the color and texture information, cannot be imaged through transparent objects such as glass, and do not work well when the ambient temperature is close to the target temperature. Visible and thermal images have their own advantages and disadvantages in different scenarios. Therefore, researchers have raised interest in multispectral pedestrian detection technology.

Fig. 1. It can be observed that the illumination conditions are poor and thermal images are more suitable for pedestrian detection than visible images. However, as shown in the red box, the reflection of thermal radiation on the marble surface creates a ghostly image of pedestrians. Adjusting the detector's propensity for visible or thermal images only based on illumination conditions can easily lead to false detection results. (Color figure online)

From the difference in imaging principles, it is natural to realize that different illumination conditions have a great impact on the imaging quality of visible images. Therefore, some works [3,4] design an illumination-aware network to evaluate the illumination conditions of an entire image and generate weights to fuse the features extracted from different modalities. However, there are difficulties in some scenarios, mainly for two reasons. First, the light source in the real world is complex. In addition to sunlight, there are street lights, car lighting and various reflections. In different spatial locations of an image, the illumination conditions are often inconsistent, so the contribution of visible images and thermal images to features cannot be simply represented by a single value. Second, illumination conditions are not the only standard to judge whether visible or thermal features are more conducive to pedestrian detection. For example, the specular reflection of visible light by glass, and the reflection of thermal radiation by the smooth marble surface will cause interference with pedestrian detection. A specific example is shown in Fig. 1.

Based on the above considerations, we propose a novel Multiscale Cross-Modality Attention (MCA) module to efficiently extract and fuse features. We embed the module into the two-stream backbone, and introduce MCANet for multispectral pedestrian detection. The main contributions of our work are as follows:

– We introduce an end-to-end Multiscale Cross-Modality Attention Network (MCANet) for multispectral pedestrian detection and validate the effectiveness of fusion for learning cross-modality features.
– We propose the Multiscale Cross-Modality Attention (MCA) module. We develop a novel attention fusion mechanism and combine it with transformer to enhance the saliency of objects and suppress the background. A new loss item is introduced to the loss function for training the attention weights.
– MCANet conducts extensive experiments on the KAIST dataset, obtaining state-of-the-art performance.

2 Related Work

2.1 Multispectral Pedestrian Detection

In order to fuse visible and thermal images and greatly improve the accuracy and robustness of pedestrian detection algorithms in different scenarios, researchers have made many efforts. Some challenging multispectral pedestrian detection datasets have been proposed, such as KAIST [5], LLVIP [6], etc., which have become important references to verify the performance of the algorithm. With the development of deep learning, convolutional neural networks have gradually been used in multispectral pedestrian detection tasks. Liu et al. [7] design four ConvNet fusion architectures, which fuse channel features at different ConvNet stages and prove that the halfway fusion model can achieve better performance. CIAN [8] proposes the cross-modality interactive attention to explicitly model the importance of feature channels and introduces the context enhancement blocks (CEBs) to further augment contextual information. Illumination-aware Faster R-CNN [3] introduces an illumination-aware weighting mechanism to adaptively weight the detection confidence of two modalities according to the illumination measure and adaptively merge the two sub-networks to obtain final detections. MBNet [4] designs an illumination-aware feature alignment module to align two modality features and induce the network to be optimized adaptively according to illumination conditions. Fang et al. [9] propose a transfomer-based fusion approach, named Cross-Modality Fusion Transformer (CFT), to enhance the representation capability of two-stream CNNs. Li et al. [10] propose the dense fusion strategy to fuse information at the feature level and use Dempster's combination rule to fuse the results of different branches according to the uncertainty.

2.2 Attention Mechanism

Multispectral fusion can be further divided into two questions: 1. How can complementary features be extracted between different modalities? 2. How can the extracted features be fused into the previous branch efficiently? For the first question, the previous works [7] directly use the feature maps of their respective branches for addition or concatenation operations. [4,8] adopt global pooling

and linear layers to squeeze the spatial dimension, and then channel-wise attention based weighting is applied to the cross-modality features. [11] use convolutional layers to extract features, and apply the spatial-wise attention mechanism by element-wise multiplication. Considering that it is difficult to determine whether the pixel value of the pedestrian is larger or smaller than the background, we believe that the direct use of element-wise multiplication may not be the best fusion method. In this paper, combined with the confidence information of pedestrians, a novel spatial attention mechanism using supervised learning is proposed, which enhances the saliency of objects and suppresses the background. This method can make full use of the complementary features of different modalities.

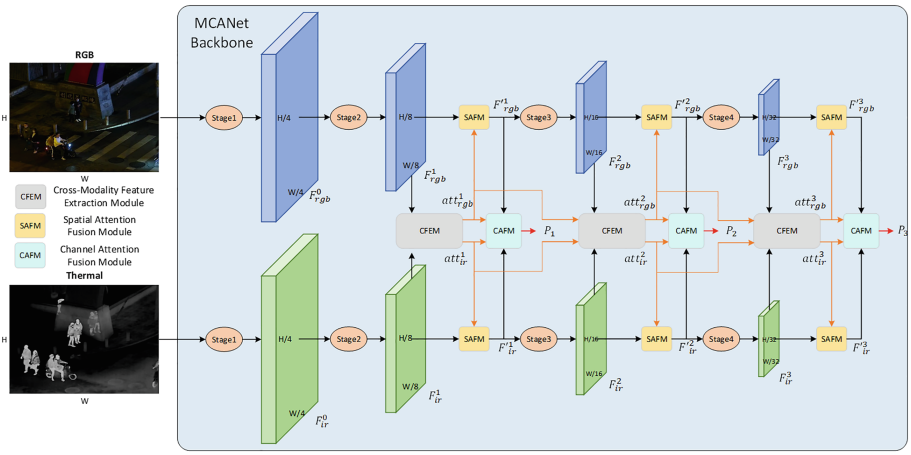

Fig. 2. Overview of Multiscale Cross-Modality Attention Network's backbone. Stage l represents the convolution module of each branch. F^l_{rgb} and F^l_{ir} are the feature maps of RGB and Thermal modalities. att^l_{rgb} and att^l_{ir} are the extracted weights after CFEM. F'^l_{rgb} and F'^l_{ir} are the fused feature maps after SAFM and P_l are the fused feature maps after CAFM. P_l will be sent to YOLOv5 detectors for prediction.

3 Proposed Method

The MCANet extends the framework of YOLOv5 [12], to enable multispectral object detection. An illustration of Multiscale Cross-Modality Attention Backbone is shown in Fig. 2. The MCA module consists of three basic components: Cross-Modality Feature Extraction Module(CFEM), Spatial Attention Fusion Module (SAFM) and Channel Attention Fusion Module (CAFM), as shown in detail in Fig. 3. They are embedded at different depths of the network and share information at different scales through interconnections.

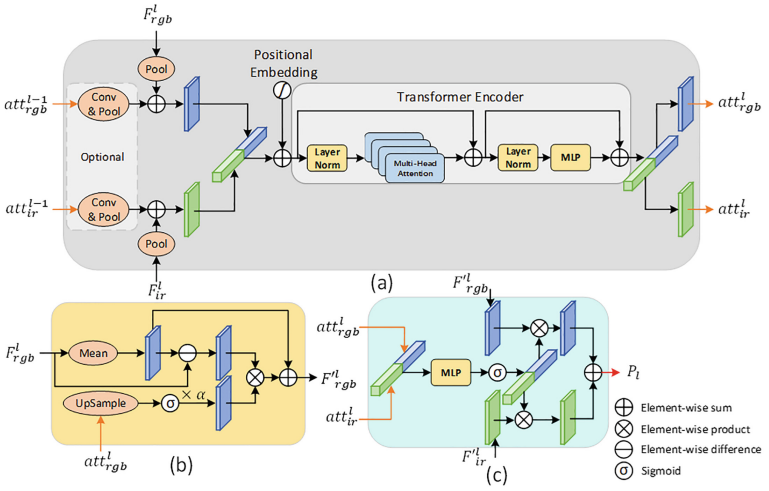

Fig. 3. The architecture of the proposed MCA module: (a) Cross-Modality Feature Extraction Module. (b) Spatial Attention Fusion Module. (c) Channel Attention Fusion Module.

3.1 Cross-Modality Feature Extraction Module

In CFEM, we want to obtain complementary features between different modalities to provide spatial weights for later attention mechanisms. Unlike convolution, which only has a local receptive field, the transformer can take into account global spatial information. Inspired by [8], we use Transformer for cross-modality feature extraction. The details are shown in Fig. 3(a).

In order to reduce the parameters and computation, taking the lth layer as an example, the RGB feature map F^l_{rgb} and thermal feature map F^l_{ir} are first downsampled to $f^l_{rgb} \in R^{C\times H\times W}$ and $f^l_{ir} \in R^{C\times H\times W}$ by max pooling and average pooling. If the previous CFEM exists, a convolution and a pooling operation will be performed on the previous outputs att^{l-1}_{rgb} and att^{l-1}_{ir}. Then the results will be added to f^l_{rgb} and f^l_{ir} respectively. The sequence of embedded patches $x^l_{p-rgb} \in R^{HW\times C}$ and $x^l_{p-ir} \in R^{HW\times C}$ is achieved by flattening the spatial dimensions of the feature map and projecting to the transformer dimension. We concatenate the patch embeddings of each modality and position embeddings are added to the patch embeddings to retain positional information. The fusion patch embeddings x^l_p are obtained and passed to the Transformer Encoder which consists of alternating layers of multiheaded self-attention (MSA) (1) and MLP (2) blocks:

$$Z'_l = x^l_p + MSA(LayerNorm(x^l_p)) \tag{1}$$

$$Z_l = Z'_l + MLP(LayerNorm(Z'_l)) \tag{2}$$

Layernorm (LN) is applied before every block, and residual connections are applied after every block. The MLP contains two layers with a GELU non-linearity. Finally, exploiting the inverse operation of the first step, the output sentences Z_l are converted into the results att^l_{rgb} and att^l_{ir}.

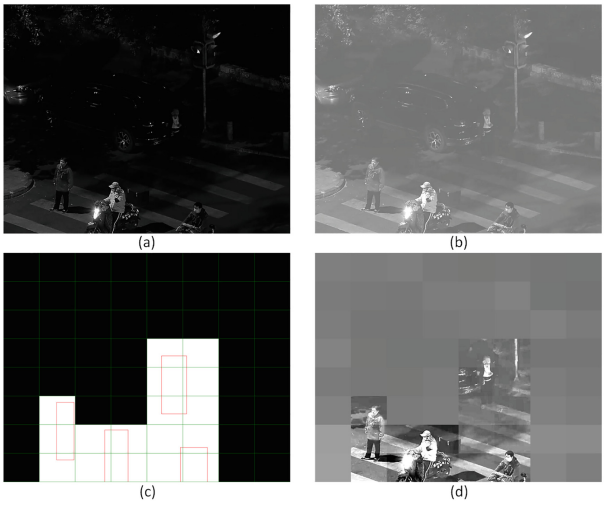

Fig. 4. (a) The grayscale image example. (b) The mean of image (a) is shifted to 127. (c) The partitioned grids and bboxes. (d) The fused image after SAFM and the mean of it is also shifted to 127.

Specifically, we want att^l_{rgb} and att^l_{ir} to give the approximate position where the object is located in the feature map. For the sake of illustration and visualization, we take Fig. 4 as an example. We convert the image from RGB to gray as the extracted feature map F^l_{rgb} in Fig. 4(a). Suppose we resize it to f^l_{rgb} with shape $h \times w$ and obtain output att^l_{rgb}. Then each element of att^l_{rgb} corresponds to the region with shape $H/h \times W/w$ in feature map F^l_{rgb}, i.e. the green grid in Fig. 4(c). The red box represents the bounding box of the object. The value of the green grids that intersect with the bounding box is set to 1, and the rest are set to 0. We take it as the ground truth, denoted as gt_att^l. During training, we apply a piecewise activation function $\sigma'(x)$ on att^l_{rgb} to restrict its range to (0,1). The $\sigma'(x)$ is formulated as:

$$\sigma'(x) = \begin{cases} 1 & \sigma(x) > 0.7 \\ \sigma(x) & 0.3 < \sigma(x) < 0.7 \\ 0 & \sigma(x) < 0.3 \end{cases} \tag{3}$$

where $\sigma(x)$ denotes the Sigmoid function. When $\sigma(x)$ is greater than 0.7 or less than 0.3, the value is directly set to 1 or 0, respectively. In this way, during backward propagation, while having the ability to roughly describe the object

location, att^l_{rgb} can concentrate on receiving the gradients propagated by the backbone network and learn more representations.

Then we compute a Binary Cross Entropy Loss with gt_att^l. The same operation is performed for att^l_{ir} and the final loss is the average of them.

$$BCELoss(x, y) = -\frac{1}{whc}\sum_{i=1}^{w}\sum_{j=1}^{h}\sum_{k=1}^{c}-[y_{i,j}\cdot\log x_{i,j,k} + (1-y_{i,j})\cdot\log(1-x_{i,j,k})] \tag{4}$$

$$L_{att} = \sum_{l=1,2,3}\frac{1}{2}[BECLoss(\sigma'(att^l_{rgb}), gt\_att^l) + BECLoss(\sigma'(att^l_{ir}), gt\_att^l)] \tag{5}$$

The total loss function uses 4 components: box, class, objectness and attention as follows:

$$L = \lambda_{obj}L_{obj} + \lambda_{box}L_{box} + \lambda_{cls}L_{cls} + \lambda_{att}L_{att} \tag{6}$$

The first three items are the same as in YOLOv5. Four parameters are used to balance different losses. In the task of multispectral pedestrian detection, the classification loss is equal to 0. In our experiment, λ_{obj}, λ_{box}, and λ_{att} are set to 0.05, 1, and 0.015, respectively.

3.2 Spatial Attention Fusion Module

In the spatial attention fusion module (SAFM), we do not use the traditional spatial attention mechanism in which the feature maps are multiplied directly with the weights. We note that pedestrians in visible images may have lower pixel values than the background due to wearing dark clothes, or higher pixel values than the background at night due to insufficient illumination conditions. In the thermal image, the pixel values of pedestrians are generally higher than that of the background. However, it is also possible that the pixel values of pedestrians are lower due to wearing thick clothes, or the ground and vehicle shell being exposed to high temperature for a long time, especially in summer. The pixel values of different parts of pedestrians may also vary greatly. Therefore, it is not always possible to extract better features if the spatial weights are multiplied with feature maps directly. We think it may be better to zoom in on the difference between the object and the background.

Therefore we design the SAFM as shown in Fig. 3(b). Similar to Sect. 3.1, we divide the input feature F^l_{rgb} into $w \times h$ green grids, as shown in Fig. 4(c). Within each grid, we average all pixel values and reassign all pixels with the mean. The offsets are obtained by subtracting the mean from the original feature maps as follows:

$$Offset^l_{rgb} = F^l_{rgb} - Mean(F^l_{rgb}) \tag{7}$$

Table 1. Comparisons with the state-of-the-art methods on the KAIST dataset

Methods	Reasonable			All		
	Rea.	Rea.Day	Rea.Night	All	Day	Night
ACF [5]	47.32	42.57	56.17	67.74	64.31	75.06
Halfway Fusion [7]	25.75	24.88	26.59	49.18	47.58	52.35
FusionRPN+BF [13]	18.29	19.57	16.27	51.70	52.33	51.09
IAF R-CNN [3]	15.73	14.55	18.26	44.23	42.46	47.70
IATDNN+IASS [14]	14.95	14.67	15.72	48.96	49.02	49.37
CIAN [8]	14.12	14.77	11.13	35.53	36.02	32.38
MSDS-RCNN [15]	11.34	10.53	12.94	34.15	32.06	38.83
AR-CNN [16]	9.34	9.94	8.38	34.95	34.36	36.12
MBNet [4]	**8.13**	**8.28**	7.86	31.87	32.37	30.95
CMPD [10]	8.16	8.77	7.31	28.98	28.3	30.56
MCANet	8.24	8.97	**7.00**	**26.07**	**27.07**	**24.3**

In Sect. 3.1, we already have the information about the possible locations of the objects. Then we resize it to $W \times H$ by nearest-neighbor sampling and apply a Sigmoid activation to it, denoting the result as Att^{l}_{rgb}.

$$Att^{l}_{rgb} = \sigma(UpSampling(att^{l}_{rgb})) \tag{8}$$

The offset is scaled by multiplying by Att^{l}_{rgb} and coefficient α, and then added to the mean of the input feature to obtain the final output $F^{'l}_{rgb}$.

$$F^{'l}_{rgb} = \alpha \cdot Att^{l}_{rgb} \cdot Offset^{l}_{rgb} + Mean(F^{l}_{rgb}) \tag{9}$$

Take Fig. 4(b) and Fig. 4(d) as an intuitive comparison of the fusion effect. Figure 4(b) is the grayscale image and Fig. 4(d) is the ideal fusion result $F^{'l}_{rgb}$. Due to the poor illumination conditions, we move the mean values of both Fig. 4(b) and Fig. 4(d) to 127 for easier comparison. Ideally, the value of pixels which represent the background in gt^{l}_{att} is 0. As a result, the background values after fusion are all equal to the mean of each grid. In contrast, the variance of pixel values becomes larger for regions containing objects. The pedestrians in Fig. 4(d) are more salient than those in Fig. 4(b), and the background regions are greatly suppressed. The two pedestrians on the left are much brighter than the background, and the clothes of the pedestrian in the middle of the picture are much darker and easier to distinguish. It should be noted that since the pixel values of Fig. 4(d) may be out of range (0–255) after being scaled up, in order to display normally, the excess parts have to be truncated, resulting in some details in Fig. 4(d) becoming blurred. In fact, there is no truncation operation during the training and inference, so all texture details are preserved without concerns about the loss of information.

3.3 Channel Attention Fusion Module

In the Channel Attention Fusion Module (CAFM), we make a small modification to the traditional channel attention mechanism for multi-modality fusion, as shown in Fig. 3(c). We apply the Global Average Pooling (GAP) to att^l_{rgb} and att^l_{ir} and concatenate them. After that, they are sent to the MLP block and a Sigmoid function to generate the weights of channels as follows:

$$ch\_att^l = \sigma(MLP(Concat(GAP(att^l_{rgb}), GAP(att^l_{ir})))) \tag{10}$$

The ch_att^l is separated into two parts $ch\_att^l_{rgb}$ and $ch\_att^l_{ir}$. The fusion result P_l for later prediction can be formulated as:

$$P_l = ch\_att^l_{rgb} \cdot F^{'l}_{rgb} + ch\_att^l_{ir} \cdot F^{'l}_{ir} \tag{11}$$

Table 2. Evaluations on the KAIST dataset under six subsets.

Methods	Near	Medium	Far	None	Partial	Heavy	All
ACF [5]	28.74	53.67	88.2	62.94	81.40	88.08	67.74
Halfway Fusion [7]	8.13	30.34	75.70	43.13	65.21	74.36	49.18
FusionRPN+BF [13]	0.04	30.87	88.86	47.45	56.10	72.20	51.70
IAF R-CNN [3]	0.96	25.54	77.84	40.17	48.40	69.76	44.23
IATDNN+IASS [14]	0.04	28.55	83.42	45.43	46.25	64.57	48.96
CIAN [8]	3.71	19.04	55.82	30.31	41.57	62.48	35.53
MSDS-RCNN [15]	1.29	16.19	63.73	29.86	38.71	63.37	34.15
AR-CNN [16]	0.00	16.08	69.00	31.40	38.63	55.73	34.95
ASPFFNet [11]	0.01	16.27	45.42	25.60	34.90	57.53	–
MBNet [4]	0.00	16.07	55.99	27.74	35.43	59.14	31.87
CMPD [10]	0.00	12.99	51.22	24.04	33.88	59.37	28.98
MCANet	**0.00**	**12.22**	**42.40**	**21.37**	**29.78**	**56.52**	**26.07**

4 Experiments

4.1 Dataset and Metric

The KAIST multispectral pedestrian detection dataset [5] contains 95,328 pairs of aligned visible and thermal images. It contains a variety of scenes acquired during the day and night to cover changes in diverse lighting conditions. The test set consists of 2, 252 frames sampled every 20th frame from video, among which 1,455 images are captured during daytime and the remaining 797 images are captured during nighttime. Due to the problematic annotations in the original training data, we adopt the annotations improved by Zhang et al. [16] for

training. All the detection performances are evaluated on the KAIST test set with annotations improved by Liu et al. [7]. The evaluation metric follows the standard KAIST evaluation [5]: log-average Miss Rate over False Positive Per Image (FPPI) range of $[10^{-2}, 10^{0}]$ (denoted as MR^{-2}). A lower score indicates better performance.

4.2 Implementation Details

Throughout this paper, we extend the framework of YOLOv5l, to enable multispectral object detection. The anchors are set to [10,13, 16,30, 33,23], [30,61, 62,45, 59,119], and [116,90, 156,198, 373,326] on three detectors with different scales. We use the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 1e-2, a momentum of 0.937, and a weight decay of 0.0005. To avoid optimization instabilities, we use the first three epochs for warmup. The warmup initial momentum is set to 0.8 and the warmup initial bias learning rate is set to 0.1. For data augmentation, we use the mosaic method which mixes four training images into one image. The MCANet is developed on an Ubuntu 18.04 platform with PyTorch 1.12.0 and two NVIDIA 3090 GPUs. The network is trained for 20 epochs and the batch size is 32.

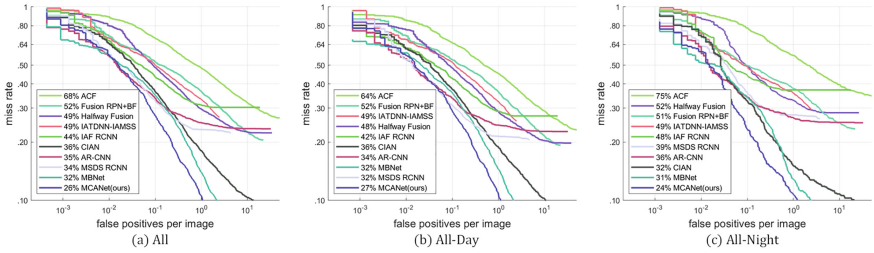

Fig. 5. Comparisons of detection results on KAIST.

4.3 Quantitative Evaluation

As shown in Table 1, we present the experimental results in terms of MR^{-2} under reasonable and all-dataset settings, respectively, as in existing works. In reasonable setting, only pedestrians taller than 55 pixels under no or partial occlusions are considered in the evaluation. Instead, all the labels, including small pedestrians and heavy occlusions, are used for evaluation in all-dataset setting. Therefore, it is obvious that all-dataset setting is more challenging than the reasonable setting. In the all-dataset setting, Table 1 shows that our proposed method achieves approximately 2.91% lower on MR^{-2} which implies that the MCANet has a substantially better localization accuracy compared with CMPD. The results show that our model can better extract and fuse complementary

features of multiple modalities to improve the detection accuracy. The FPPI-MR curve on the all-dataset setting is shown in Fig. 5, which also demonstrates the superiority of our method.

In order to have a comprehensive understanding of detector performance, we also make an evaluation under other six subsets including the pedestrian distances and occlusion levels. As shown in Table 2, the MCANet ranks first in all six subsets. Especially in the Far subset, MCANet achieves 8.82% lower on MR^{-2}, which indicates that MCANet performs satisfactorily in detecting small targets.

Fig. 6. Comparisons of detection results on KAIST with MBNet.

Table 3. The ablation experiments of the MCANet on the KAIST dataset

Method	All	Day	Night
CFEM+element-wise product+element-wise sum	9.39	10.29	7.45
CFEM+SAFM+element-wise sum	8.55	9.11	7.24
CFEM+SAFM+CAFM	**8.24**	**8.97**	**7.00**

4.4 Qualitative Evaluation

In order to further demonstrate the effectiveness of the proposed MCANet, the detection results are compared with MBNet, which is one of the state-of-the-art algorithms that have been open sourced, as shown in Fig. 6. The first row is the original images and the green rectangles are manually labeled ground truth. The second and third rows are the detection results of MBNet and our method, respectively. It can be clearly seen from Fig. 6 that MCANet can effectively solve the problem of missed detection. Due to the efficient feature extraction and saliency enhancement functions of MCANet, the detection accuracy of small targets can be effectively improved.

4.5 Ablation Study

Ablation experiments are performed on the KAIST dataset to demonstrate the effectiveness of the components of our MCA module. We test three different fusion strategies as shown in Table 3. All of them use CFEM to extract the features. The first method replaces the SAFM and CAFM with element-wise product and element-wise sum respectively. And the second method replaces the CAFM with element-wise sum. The third method uses all of the proposed components. The experiments show that the final version with all three designed components outperforms the other versions. The results of the ablation study demonstrate the effectiveness of the proposed components.

5 Conclusion

In this work, we propose a novel MCA module to efficiently extract and fuse features. The MCA modules are embedded into two-stream backbone and the MCANet is introduced. We explore how to effectively enhance the saliency of objects and suppress the background. Specifically, the transformer architecture is used to extract Cross-Modality complementary features. Then a novel attention fusion mechanism is developed. The multiscale information is shared between different depths of the network to ensure the robustness of the detector. The experiments demonstrate that the proposed MCANet outperforms the state-of-the-art on the challenging KAIST dataset in terms of the accuracy.

References

1. Torabi, A., Massé, G., Bilodeau, G.-A.: An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications. Comput. Vis. Image Underst. **116**(2), 210–221 (2012)
2. Wu, B., Iandola, F., Jin, P.H., Keutzer, K.: SqueezeDet: unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 129–137 (2017)
3. Li, C., Song, D., Tong, R., Tang, M.: Illumination-aware faster R-CNN for robust multispectral pedestrian detection. Pattern Recogn. **85**, 161–171 (2019)
4. Zhou, K., Chen, L., Cao, X.: Improving multispectral pedestrian detection by addressing modality imbalance problems. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12363, pp. 787–803. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58523-5_46
5. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: benchmark dataset and baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1037–1045 (2015)
6. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: LLVIP: a visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3496–3504 (2021)
7. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. arXiv preprint arXiv:1611.02644 (2016)

8. Zhang, L., et al.: Cross-modality interactive attention network for multispectral pedestrian detection. Inf. Fusion **50**, 20–29 (2019)
9. Qingyun, F., Dapeng, H., Zhaokui, W.: Cross-modality fusion transformer for multispectral object detection. arXiv preprint arXiv:2111.00273 (2021)
10. Li, Q., Zhang, C., Hu, Q., Fu, H., Zhu, P.: Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. In: IEEE Trans. Multimedia (2022)
11. Fu, L., Gu, W.-B., Ai, Y.-B., Li, W., Wang, D.: Adaptive spatial pixel-level feature fusion network for multispectral pedestrian detection. Infrared Phys. Technol. **116**, 103770 (2021)
12. Jocher, G.: YOLOv5 release v5.0 (2022). https://github.com/ultralytics/yolov5/releases/tag/v5.0
13. Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M.: Fully convolutional region proposal networks for multispectral person detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 49–56 (2017)
14. Guan, D., Cao, Y., Yang, J., Cao, Y., Yang, M.Y.: Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Inf. Fusion **50**, 148–157 (2019)
15. Li, C., Song, D., Tong, R., Tang, M.: Multispectral pedestrian detection via simultaneous detection and segmentation. arXiv preprint arXiv:1808.04818 (2018)
16. Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5127–5137 (2019)