Attention-based Cross-Modality Multiscale Fusion for Multispectral Pedestrian Detection

Zhou Hui
School of Electrical and Information Engineering
TongJi University
Shanghai, China

***Abstract*—Multispectral pedestrian detection has wide applications in fields such as autonomous driving and intelligent surveillance. Mining complementary information between modalities is one of the most effective approaches to improve the performance of multispectral pedestrian detection. However, the inevitable introduction of redundant information between modalities during the fusion process leads to feature degradation. To address this challenge, we propose a multiscale differential fusion algorithm that leverages complementary information between modalities to suppress feature degradation caused by noise propagation along the network. We compare our algorithm with other cross-modal fusion pedestrian detection algorithms on the LLVIP and cleaned KAIST datasets. Experimental results demonstrate that our algorithm outperforms others, particularly in nighttime scenes where our algorithm achieves a 7.28% improvement in recall rate compared to the baseline on the cleaned KAIST dataset.**

***Keywords*—Pedestrian detection; multispectral pedestrian detection; attention mechanism; cross-modal fusion.**

I. Introduction

Pedestrian detection plays an important role in autonomous driving systems. In well-illuminated conditions, pedestrian detection achieves high precision. In poor lighting conditions, the appearance of pedestrians becomes blurred. Obstacles, overlapping figures, and varying distances contribute to these differences. As a result, nighttime pedestrian detection currently faces significant challenges [1].

Many advanced algorithms based on visible light images achieve notable performance improvements. Recent studies involving these images have validated their effectiveness, including in nighttime environments [2]. Due to the poor quality of nighttime visible light images, deep convolutional neural networks struggle to learn effective features. Image enhancement techniques show remarkable performance in enhancing the contrast between the foreground and background of an image. Some studies utilize enhanced image for feature extraction [3]. However, the majority of machine vision and deep learning models tend to perform poorly in highly challenging low-light scenarios [4]. Infrared images can highlight the thermal radiation characteristics of target objects, allowing for the capture of details such as human contours. Therefore, it possesses unique advantages in scenarios with insufficient lighting, adverse weather conditions, or concealed surveillance.

Despite the significant advantages of multimodal input data, effectively fusing information between modalities has become the core challenge and focus of algorithmic research. Li [5] et al. compared six fusion architectures which integrate color and thermal modalities at different position. Based on different fusion stages, it can be classified into early fusion, halfway fusion, and late fusion. Late fusion is currently the more commonly employed method, capable of mitigating the influence of modality and feature misalignment. However, it encounters challenges in network convergence and high computational complexity. We observed that discussions rarely address both the redundancy and complementarity of modalities. Crucially, the spread of redundant information can have detrimental effects in networks. This paper focuses on examining and mitigating these negative impacts by leveraging differential information of modalities in the backbone network to reduce redundancy.

The Non-Local neural network (Non-Local) [6] enhances inputs by calculating similarity in the channel direction. We conjecture that constructing an attention map by calculating similarity between pedestrian features could effectively allocate increased attention to those with blurred characteristics. In multispectral scenarios, there also exists a certain level of correlation both between channel dimensions and between spatial dimensions. Therefore, this paper proposes a dual-branch attention mechanism, named Dual Non-Local, which is based on both channel and spatial information. It establishes long-range dependencies between channels and spaces. Simultaneously, we utilize bright channel prior (BCP) algorithm to address low-light image compensation issues, and employ a multiscale feature fusion module to integrate visible and infrared modalities. Our work achieves superior results compared to some methods on the public available datasets KAIST and LLVIP.

We summarize the contributions of our work as follows:

1) A novel fusion approach is proposed for mining complementarity and reducing feature degradation. This technique involves the cross-fusion of complementary information from different modalities within the backbone network. The outputs of the backbone network for each modality is effectively integrated through this method.
2) A dual-branch attention mechanism based on channel and spatial attention. We embed positional information into attention map, and reduced the computational complexity of spatial attention. Ultimately, we build a dual-branch 3D attention mechanism that collaborates between spatial and channel dimensions.

II. Related Work

A. Pedestrian Detection

Pedestrian detection has high practical value in various applications, eg., autonomous driving and video surveillance. It receives extensive research attention in the field of computer vision. Pedestrian detection has undergone a significant transformation from handcrafted features to depending on deep convolutional networks for feature extraction [7]. Based on channel features or Deformable Part Models (DPM), there are two approaches to pedestrian detection that rely on handcrafted features. In 2009, P. Dollar et al. offered a fresh approach Integral Channel Features (ICF) [8], which utilized integral images for rapid feature computation. By combining channel feature pyramids with a cascaded classifier, they achieved faster detection results. ICF was the basis of channel features. Filtered Channel Features (FCF) [9] was optimization methods derived from ICF. Conventional algorithms were contingent upon manual design and frequently yielded diminished levels of detection accuracy. With Convolutional Neural Networks (CNNs) demonstrating outstanding feature extraction capabilities across various object detection tasks, pedestrian detection methods focused on leveraging deep learning techniques to enhance detection performance recently. The emergence of single-stage and two-stage algorithms, such as Faster R-CNN [10], [11], was a substantial potential for advancing accuracy and speed in pedestrian detection. Once in all weather conditions, especially during nighttime scenes, visible-light-based detection methods struggle to be effective. Simultaneously, infrared images complements visible light images, enabling the capture of pedestrian contours even in nighttime conditions. Detecting pedestrians in all weather conditions using multispectral images of color-thermal pairs has become a research hotspot.

B. Multispectral Pedestrian Detection

Effectively integrating infrared and visible light modalities is a challenging problem. In 2015, Hwang [12] et al. collected multispectral datasets, KAIST. The authors proposed the multispectral Aggregated Channel Features (ACF) method, incorporating intensity and gradient information from the thermal channel as additional channel information. An increasing number of multispectral pedestrian detection algorithms emerged based on this dataset. Liu [13] confirmed that multimodal pedestrian detection outperforms single-modal detection in terms of performance. Liu also investigated four fusion architectures: early fusion, mid-fusion, late fusion, and confidence fusion. They concluded that halfway fusion is the most effective fusion architecture. Inspired by Faster R-CNN, Konig [14] et al. proposed an effective multispectral RPN (Region Proposal Network)+BDT (Enhanced Decision Tree) model. In addition to investigating the fusion stages of multispectral images, another research approach involved using an illumination-aware network to weight the two modalities. Illumination-aware Faster R-CNN (IAF R-CNN) [5] introduced an illumination-weighting mechanism, forming a unified detection framework with separate subnetworks for visible light and infrared, along with a weighting layer. That means in low-light conditions, the network emphasized the features learned from the infrared sub network. In well-illuminated conditions, it focused on the visible light subnetwork. Our work is closely related to the conclusions drawn in [14]. We employed YOLOv7 [15] as our baseline and investigated the positive impact of low-light image enhancement techniques on the performance of multispectral pedestrian detection.

C. Attention Mechanism

In deep learning, the attention mechanism emulates the human visual and cognitive system, enabling neural networks to focus attention on relevant parts. Due to its outstanding performance, the attention mechanism is widely utilized in machine vision. Squeeze-and-Excitation Networks (SENet) [16] achieved adaptive channel-wise feature recalibration by modeling interdependencies between channels. Convolutional Block Attention Module (CBAM) [17] combined a channel attention module with a spatial attention module, allowing channel attention and spatial attention to operate sequentially. This enabled the network to simultaneously learn dependencies between channels and positional information. Non-Local neural network [6] combined self-attention with the general non-local mean method, establishing a long-range dependency model for transmitting long-range information. Non-Local maintained consistent feature scales between input and output, so it can be employed without modifying the network architecture. Criss-Cross Attention Network (CCNet) [18] and Global Context Network (GCNet) [19] were improvements derived from Non-Local. Similarity-based Attention Module (SimAM) [20] suggested that attention in the human brain often work in synergy, thus a unified attention mechanism was more in line with the working mechanism of neurons in the human brain. This paper introduces a new attention mechanism that combines the ideas of SimAM and Non-Local.

III. Proposed Method

The structure of this paper is depicted in Fig. 1. This model utilizes YOLOv7 as the baseline and integrates it with an illumination compensation module, a multiscale fusion module, and a detection module, forming a unified detection architecture. Our model consolidates the methods of image enhancement and differential fusion into a cohesive framework, thoroughly addressing the redundancy and complementarity across different modalities.

A. Illumination Compensation Network

The atmospheric scattering model is commonly employed to represent the degradation process of hazy images and is sometimes used for image enhancement tasks in low-light conditions as well [21]. Original image captured by the camera can be expressed as:

$$I = tJ + A\,(1 - t) \tag{1}$$

where, I is original image, J is restoration function of the image, A is environmental light description function, and t is medium propagation description function.

Wang [22] demonstrated that well-exposed images have at least some pixels with high illumination, unless these pixels are in shadow or covered by a black object. We visualize the Bright Channel on the KAIST dataset in Fig. 2. Most visible light images on the KAIST dataset are in underexposed scenes. We adopt unsupervised low-light Image enhancement network

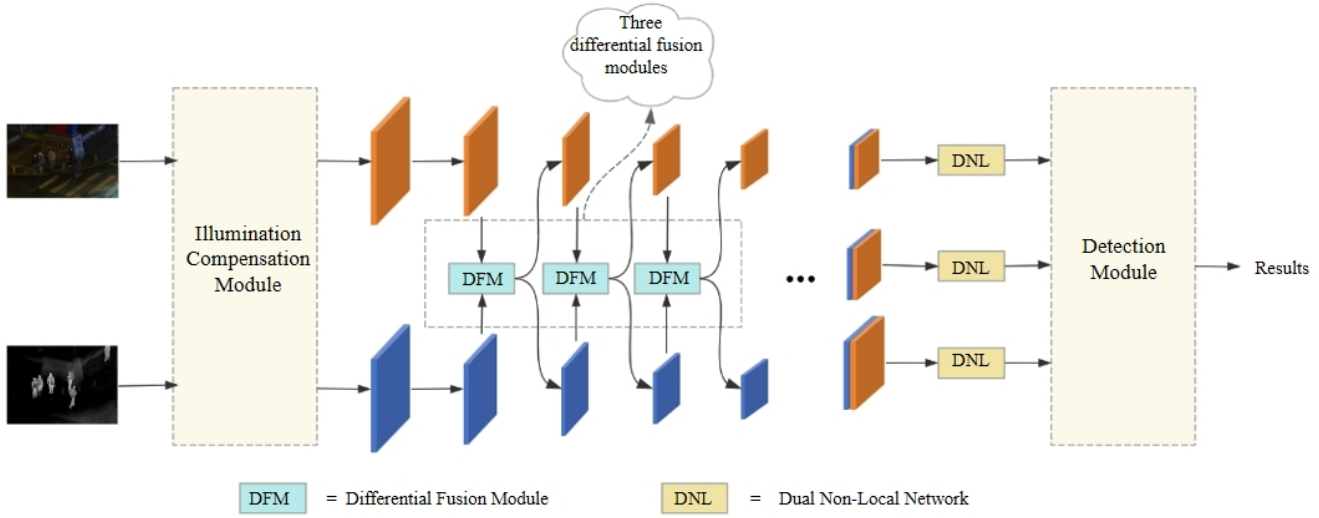


Fig. 1. The overall structure of proposed method. The network takes multispectral images of color-thermal pairs as inputs.

Fig. 2. The bright channel visualization of KAIST. We selected several low-light images and calculated the brightest pixels in R, G, B channels for each image, denoted as bright channel.

for outputing adjusted image that is guided by a unsupervised loss L_{BCP}. The parameters in Eq. (1) are reinterpreted as Eq. (2):

$$I_p = t_p J_p + A\,(1 - t_{\mathrm{p}}\,) \tag{2}$$

where, A and t_p is environment light and the illumination map, respectively. J_p represents enhanced output images and I_p is observed images. According to BCP [22], We adopted Eq. (3) as the brightest intensity:

$$J_p^{bright} = \max_{c\in r,g,b}\left(\max_{q\in\Phi(p)} J_q^c\right) \tag{3}$$

where, J_p^{bright} represents the brightest intesity in r,g,b channels, q is the pixels centered at p, and c is different channels in RGB images. Additionally, the brightest intesity becomes $J_p^{bright} \to 1$. Assuming that A is known, $\widetilde{t}$ represents the illumination map, which is considered as a constant within a patch.

By taking the maximum operator between the left and right sides of Eq. (3), we obtain an initial illumination map formulation as shown in Eq. (4):

$$\tilde{t}_p = 1 - \max_{c,q}\left(\frac{1 - I_q^c}{1 - A^c}\right). \tag{4}$$

where $\tilde{t}_p$ is the illumination value at pixel p, I_q^c is observed image, q is pixel centered at p, and c is different channels in RGB images. Under the supervision of the initial illumination map, we obtain enhanced illumination map t_p through Illumination Compensation Network. Substituting t_p into Eq. (2), we get enhanced output images as Eq. (5):

$$J_p = \frac{I_p - A}{t_p} + A \tag{5}$$

The darkest pixels in the bright channel of the image can be considered as environment light. To adjust dark spots and black objects in real life, we take the average value of the darkest 0.1% pixels (denoted as K) in the bright channel of the image as the environment light, as shown in Eq. (6):

$$A = \frac{1}{|K|}\sum_{p\in K} I_p \tag{6}$$

To address oversaturation, this paper similarly utilizes the output from Eq. (7) as the attention map:

$$T_{attention} = T^{\gamma} \tag{7}$$

where T is thermal image, and $\gamma(\gamma > 1)$ controls the curvature of the attention map.

The enhancement network alters the feature scale and utilizes the attention map to optimize spatial weights. In summary, our final illumination compensation network is illustrated in Fig. 3. The visible light features are compensated by the infrared images, effectively enhancing the distinguishability of the RGB images.

B. Multiscale Fusion Module

In all weather conditions, visible light images provide more information about pedestrians in well-illuminated conditions while in low-light conditions, thermal images provide more information. Most multispectral approaches extract features from

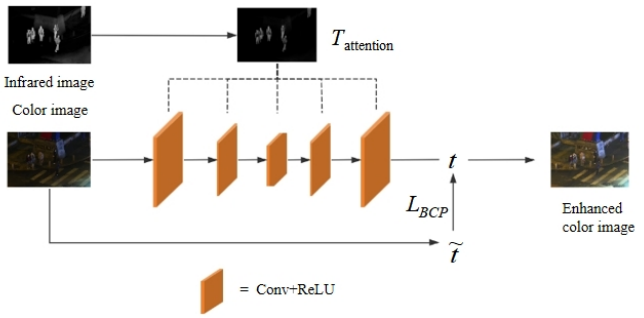


Fig. 3. The structure of Illumination Compensation module. $T_{attention}$ fed to five convolutional layers to adjust visible light feature. $\tilde{t}$ is an initial illumination map.

two streams and directly combine them either by element-wise addition or by channel concatenation. However, these mehods overlook the complementarity between the two modalities. The propagation of redundant information between modalities through the network also have adverse effects.

Inspired by the differential modality information [23], we propose a fusion module: Differential Fusion Module(DFM), to enhance the mutual suppression and enhancement, as shown in Fig. 4. The features obtained by element-wise subtraction of the two modalities reflect their complementary information, ingeniously excluding redundant information from feature fusion. This element-wise subtraction also prevents interference from features learned from another modality in the previous fusion from affecting the next fusion. Integrated within the architecture of YOLOv7, we perform multiscale feature fusion at the position illustrated in Fig. 1. In multispectral pedestrian

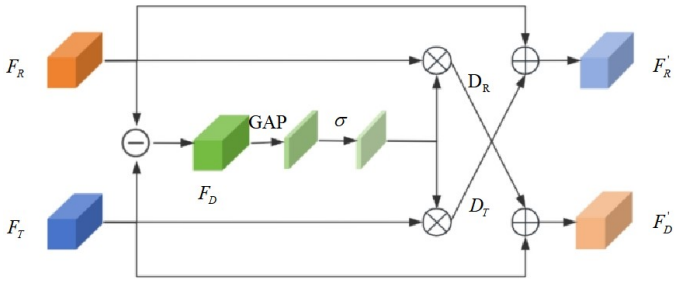


Fig. 4. The structure of differential fusion module. F_R and F_T are visible light features and thermal features. We obtain F_D by subtracting element-wise.

detection task, it is crucial to effectively integrate valuable information between modalities and mitigate interference caused by redundant information.

We applied DFM at Conv3, Conv4, Conv5 layers of the Backbone. The outputs feed into multiscale feature fusion network. To further enhance crucial features, we employ the Dual Non-Local attention mechanism before pyramid feature network. This helps to improve pedestrian feature expression effectively. DFM involves using a differencing mechanism, where F_R and F_T are subtracted element-wise to obtain the feature F_D. The equation for F_D is as follows:

$$F_{D1} = F_R - F_T \tag{8}$$

$$F_{D2} = F_T - F_R \tag{9}$$

where, F_R and F_T represent the extracted features from the visible light image and infrared image, respectively. F_D is the difference between F_R and F_T.

Subsequently, F_D is obtained through a global average pooling layer(GAP) and a tanh activation layer in order to get the attention map in the channel direction. That attention map is multiplied with the input visible light features and infrared features separately, producing D_R and D_T. The cross addition is applied to F_T and D_R, as well as F_R and D_T. Differential features yields the output features.

GAP computes the mean of the two-dimensional images within each channel, obtaining an attention map in the channel direction that contains global information. D_R is present in the visible light features but absent in the infrared image features while D_T is present in the infrared image features but absent in the visible light features. The formulation of differential feature is as follows:

$$F_R^{'} = F_R + D_T \tag{10}$$

$$F_T^{'} = F_T + D_R \tag{11}$$

where $F_R^{'}$ is the output of F_R, and $F_T^{'}$ is the output of F_T.

After cross-complementary feature fusion at three scales in the Backbone, the deep semantic information is concatenated. The deep semantic information needs to be fed into the Dual Non-Local module to enhance crucial information before the feature pyramid network. There is a high degree of correlation between pedestrian features, so establishing long-range dependencies is beneficial for modeling the similarity relationships between pedestrian features.

C. Dual Non-Local Attention Mechanism

Capturing long-range dependencies is crucial in pedestrian detection task. Long-range dependencies in image can only be formed through the successive convolutional layers in deep neural networks, named large receptive field. Inspired by SENet and Non-Local, we proposed a 3D attention mechanism called Dual Non-Local, as illustrated in the Fig. 5. Does long-range dependency work effectively in multi-pedestrian scenes? Undoubtedly, there is some correlation among the extracted features of pedestrians. The feature similarity matrices between each pixel and the feature similarity matrices between each channel assist in providing blurred pedestrian features with the weights of clear pedestrian features. Spatial attention and channel attention were applied concurrently for enhancing pedestrian feature. Dual Non-Local Network consists of a Spatial Attention Module (SAM) and a Channel Attention Module (CAM), which share the same input, denoted as $x \in \mathbb{R}^{C\times H\times W}$. There is a similar attention map computed for each position in Non-Local Network [19], thus, we use global attention maps to reduce computational cost.

In spatial attention module, we reshape the input into m, $m \in \mathbb{R}^{1\times C\times (H*W)}$. A 1×1 convolutional operation is imposed to x for getting global information of channels, denoted as n,

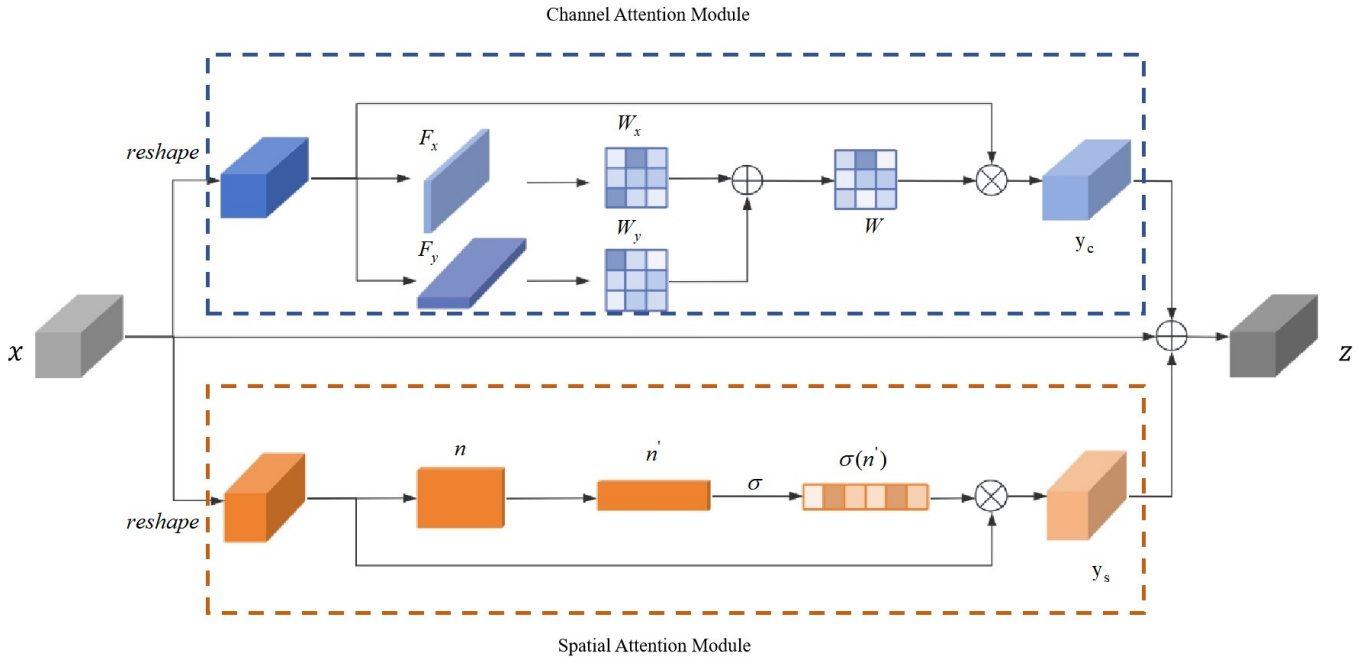


Fig. 5. The structure of dual non-Local network consists two branch attention mechanisms. SAM uses a shared attention map $\delta(n^{'})$ globally for reducing computation, while CAM calculates attention maps in different dimensions.

$n \in \mathbb{R}^{1\times H\times W}$. Before multipling with m, $n^{'}$ passes through a softmax function. Global channel information was reshaped to $n^{'} \in \mathbb{R}^{1\times(H*W)\times 1}$. In order to get the similarity between $H*W$ pixels, we multiply m and $n^{'}$. y_s is the output of spatial attention module, $y_s = m \bigotimes \delta(n^{'})$, where δ is softmax function. y_s is formulated as follows:

$$\mathbf{y}_{\mathrm{s}} = \sum_{j=1}^{N} \mathbf{x}_j \delta\left(\mathbf{n}'\right) \tag{12}$$

where, N is the number of pixels and y_s is shared globally as a spatial attention map.

In channel attention module, this paper innovatively performs pooling operations separately in the x and y dimensions, injecting positional information into channel attention map. After pooling in the x dimension, features are denoted as F_x, $F_x \in \mathbb{R}^{C\times H\times 1}$, and in the y dimension are denoted as F_y, $F_y \in \mathbb{R}^{C\times 1\times W}$, correspondingly. F_x performs a 1×1 filter, followed by a reshape function, to obtain θx, φx, where $\theta x \in \mathbb{R}^{C\times 1\times H}$ and $\varphi x \in \mathbb{R}^{C\times H\times 1}$. Dot product is carried out between θx and φx to obtain the weights between channels refered to W_x, $W_x \in \mathbb{R}^{C\times C}$ in dimension y. In a similar way, W_y, $W_y \in \mathbb{R}^{C\times C}$, represents the weights between channels in dimension x.

We suggest x and y dimensions play the same important role in channel attention, so the total channel weight distribution is considered to be the sum of the weight distribution in both dimension x and dimension y, denoted as W, $W = W_x + W_y$. Followed by a 1×1 filter and reshape operation, x generates g_x, $g_x \in \mathbb{R}^{C\times(H*W)}$, in order to obtain the final channel attention output y_c by applying the learned channel attention weights (W). To recover the features to their original input dimensions, we use a 1×1 filter to generate weights W_z. The final channel attention output is formulated as follows:

$$y_c = W_z(W \bigotimes g_x) \tag{13}$$

Finally, there is a shortcut between input and output as a residual structure. The network retains the input x, only learns the difference between output and input, and the data flows across layers to avoid the gradient disappearing during the training process. We denoted the final response of x as z, and we summarized the formulation between every pixel as follows:

$$\begin{aligned} z_i =& W_z \sum_{j=1}^{N_C} \left(\frac{f\left(C_{W_i}, C_{Wj}\right) + f\left(C_{Hi}, C_{Hj}\right)}{N_C} \right) \left(g_x \cdot x_j\right) \\ &+ m \bigotimes \delta(n^{'}) + x_i \end{aligned} \tag{14}$$

where i represents a position in image, and j is all possible positions. N_C is number of channels, $f\left(C_{W_i}, C_{Wj}\right)$ and $f\left(C_{Hi}, C_{Hj}\right)$ are the similarity between channels calculated by dot product.

The total loss is defined as (16):

$$\begin{aligned} L_{BCP} =& \frac{1}{N} \sum_{p} \{ \left(t_p - \tilde{t}_p\right)^2 \\ &+ \lambda \sum_{i,j\in\Phi(p)} w_{ij} \left(t_i - t_j\right)^2 \} \end{aligned} \tag{15}$$

$$L = L_{BCP} + \alpha L_{YOLOv7} \tag{16}$$

where N is the number of pixels, $\Phi(p)$ is the pixels within a 3×3 patch centered at p, w_{ij} represents affinity matrix between $\Phi(p)$, and λ controls balance between the data term and the smoothing term. Integrating detection and enhancement

loss during training is beneficial for obtaining image with prominent pedestrian features for pedestrian detection.

IV. Experiments

In order to demonstrate the effectiveness of the model we proposed, we present the detection results on two datasets, cleaned KAIST and LLVIP.

A. Dataset

1) KAIST: The KAIST dataset, proposed by Hwang [12] et al., consists of multispectral pedestrian data captured by specialized hardware with a beam splitter. It comprises 95,328 pairs of color and thermal images. However, this dataset is derived from consecutive frames of a video causing a high similarity in adjacent images, so we perform data clean. Finally, we get 7601 pairs of images as training set, and 2252 pairs of images as testing set. Additionally, we adopted the re-annotated labels by Li [24] and Hangil [25] for the training and test sets, respectively, to enhance label quality.

2) LLVIP: The LLVIP [26] dataset consists of rigorously aligned pairs of images in both time and space, which is used for pedestrian detection in low-light conditions. The entire dataset comprises 15,488 pairs of color-thermal images.

B. Evaluation Metrics

We use the Recall and Average Precision (AP) as evaluation metrics to evaluate the proposed model effectively. Here, we use TP (True Positive), FP (False Positive) to represent true positive predictions and false positive predictions, respectively. Recall is the ratio of detected pedestrians in ground truth. $Recall = \frac{TP}{TP+FP}$.

C. Implementation Details

In this paper, we built our network based on YOLOv7 and added a illumination compensation network at the input, which enhances the visible light by using the bright channel prior. In the Backbone, differential fusion module was performed on the feature inputs of Conv3, Conv4, and Conv5 to reduce redundancy in modal fusion. Finally, an innovative attention mechanism was added for long-term dependencies, facilitating direct transmission of high-level semantics.

The experiments were conducted on an NVIDIA GeForce RTX 4080 GPU, Intel(R) Core(TM) i7-13700F CPU, using the PyTorch framework and public code YOLOv7. We set the batch size to 8, epoch to 100, and resize input images to 640×640. K-means clustering provided nine anchor boxes for the KAIST dataset: [44,65], [26,111], [33,141], [41,117], [43,153], [58,116], [52,146], [59,178], [71,152]. We used some training tricks such as mosaic augmentation and random cropping to enhance the network's generalization.

D. Results Analysis

We conduct a comparison between our algorithm, Halfway Fusion and IAF R-CNN on the cleaned KAIST dataset. Here, we primarily discuss the potential advantages of our method, such as how our framework utilizes a Halfway Fusion architecture for integration, and we identify key methodologies that are beneficial in enhancing detection performance. The pedestrian detection results are presented in Table I. For the cleaned KAIST dataset, proposed method achieves the best detection performance in terms of Recall 64.17%.

TABLE I. Comparison on Cleaned KAIST Dataset in Terms of Recall

Method	DAY	NIGHT	ALL
Halfway Fusion [13]	59.98	50.77	58.27
IAF R-CNN [5]	65.22	56.62	62.14
YOLOv7 [15]	66.11	49.89	60.98
Ours	68.23	57.17	64.17

Compared to IAF R-CNN, our method is equally competitive, maintaining a high recall rate while our inference time is only 0.096s/image, as opposed to 0.210s/image for IAF R-CNN. We record comparison of inference time using an NVIDIA GeForce RTX 4080 GPU in Table II. This advantage is attributed to the real-time nature of the single-stage object detection algorithm, but the improvement in recall rate is due to our differential fusion module effectively mining the complementary features of pedestrian characteristics, reducing redundant noise interference in feature propagation. However, the effectiveness of DFM is limited by the requirement that the input pairs of visible and infrared images must be strictly aligned. Misaligned image pairs transmit incorrect differential information, and the noise is amplified by the network. This limitation calls for the use of more sophisticated image acquisition instruments to be adequately addressed.

TABLE II. Comparision of Inference Time using an NVIDIA GeForce RTX 4080 GPU

Method	IAF R-CNN [5]	YOLOv7 [15]	Ours
Time(s.)	0.210	0.041	0.096

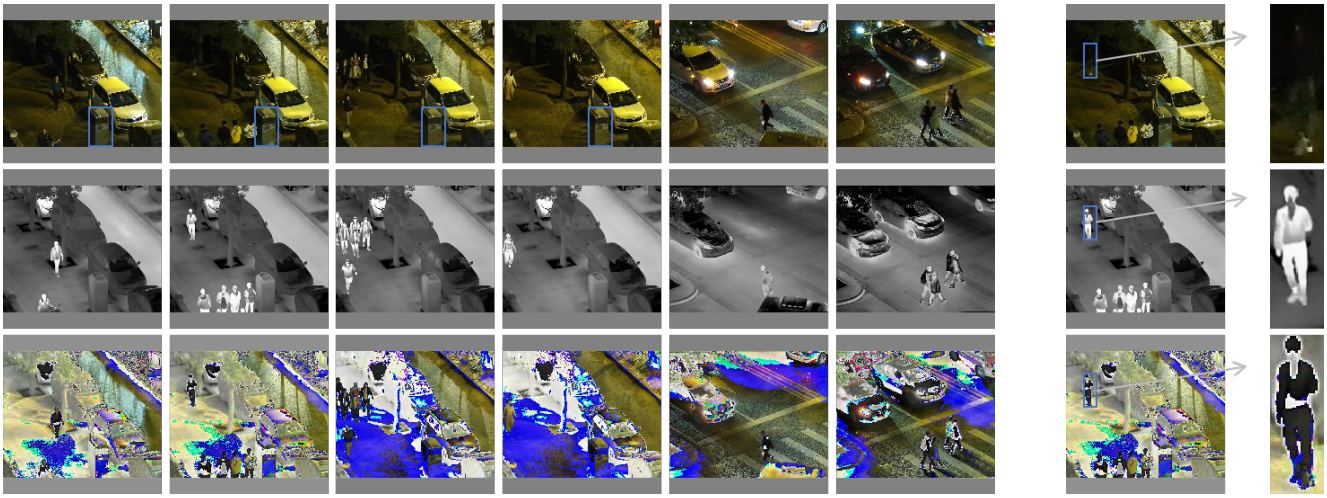
In Table III, IV and V, we compare our algorithm with YOLOv7. Moreover, we explored three versions input of YOLOv7: a) RGB branch; b) thermal branch; c) concat thermal image and visible light image as input. Directly concatenat-

TABLE III. Comparison on Cleaned KAIST Dataset for Nighttime Scenes in Terms of Average Precision, Recall, and Accuracy.

Method	AP/%	Recall/%	Accuracy/%
YOLOv7 RGB	39.73	35.54	88.03
YOLOv7 T	49.24	18.74	92.65
YOLOv7 T+RGB	55.58	49.89	80.35
Ours	64.20	57.17	86.50

TABLE IV. Comparison on Cleaned KAIST Dataset for Daytime Scenes in Terms of Average Precision, Recall, and Accuracy

Method	AP/%	Recall/%	Accuracy/%
YOLOv7 RGB	60.97	60.26	80.05
YOLOv7 T	44.57	14.68	89.89
YOLOv7 T+RGB	66.05	66.11	72.73
Ours	63.83	68.23	81.47



(a) some examples (b) detail comparsion

Fig. 6. The visualizations of enhanced features from Illumination compensation network a) some examples; b) detail comparison.

TABLE V. COMPARISON ON CLEANED KAIST DATASET FOR ALL WEATHER SCENES IN TERMS OF AVERAGE PRECISION, RECALL, AND ACCURACY

Method	AP/%	Recall/%	Accuracy/%
YOLOv7 RGB	54.30	52.44	81.63
YOLOv7 T	46.01	15.96	90.90
YOLOv7 T+RGB	62.60	60.98	76.01
Ours	63.94	64.17	82.96

ing the visible light image and thermal image did not lead to a significant improvement. We achieve improvements of 3.19%, 2.12%, and 7.28% on the all weather, daytime, and nighttime test sets in terms of Recall, respectively. Our method demonstrate better performance in both accuracy and Recall, indicating the effectiveness of our fusion strategy. Illumination compensation network enhances pedestrian features in low-light scenarios. Thus, we obtained the highest performance in nighttime scenes.

There are some visualizations for enhancced images as the outputs of Illumination Compensation Network in Fig. 6. We observed that obstacles in blue boxes have a high similarity to pedestrians, especially in low-light scenarios. The illumination compensation network is advantageous in suppressing background features and enhancing foreground characteristics. That enables the network to concentrate more on pedestrian targets, free from background interference. In the third line of Fig. 6b, pedestrian feature is clearer in blue box. However, the multispectral images of color-thermal pairs must be aligned. When misalignments occur, our model leads to worse results, which requires more sophisticated image acquisition instruments.

In addition to the quantitative analysis,we also provide several qualitative results on the cleaned KAIST dataset in Fig. 7. Upon observation, it is evident that our method excels in generating precise bounding boxes and accurately detecting pedestrians, especially in challenging scenarios when compared to the baseline model.

Gradient-weighted Class Activation Mapping (Grad-CAM) is a method for visualizing the attention mechanisms of deep neural networks. Our Dual Non-Local module constructs a unified attention framework based on the similarity of channel and spatial features, making it particularly suitable for single-object detection scenarios. The feature similarity between different types of targets may cause confusion in the attention map. In single object detection tasks, our Dual Non-Local module demonstrates superior performance compared to Non-Local and SimAM. These three similar attention mechanisms have consistent input and output dimensions. We removed all other modules in Fig. 1, retaining only the attention module. We replaced this position with different attention mechanisms and trained using the RGB images from LLVIP.

Using Grad-CAM, we visualized the outputs of Dual Non-Local, Non-Local, and SimAM, as shown in Fig. 8. Our Dual Non-Local model focuses more attention on the entirety of pedestrians, while Non-Local and SimAM distribute attention more precisely, but they both have the issue of some pedestrian regions not receiving attention. Comparatively, although the attention regions of Dual Non-Local are less precise, all pedestrian regions receive attention intensely. This also validates that features of clear pedestrians can rectify those pedestrians with blurred features.

E. Ablation Experiment

Our model achieves a leading performance. Nevertheless, the specific contributions of each module to the results remained uncertain. To address this, we design some ablation experiments to verify it. The comparsion results are presented in Table VI.

1) Illumination Compensation Module(IC): To figure out the impact of the illumination compensation network, we design a new network that uses the concatenated outputs of the illumination compensation network as the input of backbone, by removing the multiscale fusion network and

(a) visible light images (b) baseline detection results (c) our detection results (d) ground truth

Fig. 7. The visualizations of baseline and our algorithm. It contains a) visible light images; b) baseline detection results; c) our detection results; d) ground truth. According to the results, our method generate more target boxes correctly. Our method performs better when visible light images are in low-light condition.

TABLE VI. ABLATION RESULTS ON THE CLEANED KAIST DATASET IN TERMS OF PRECISION

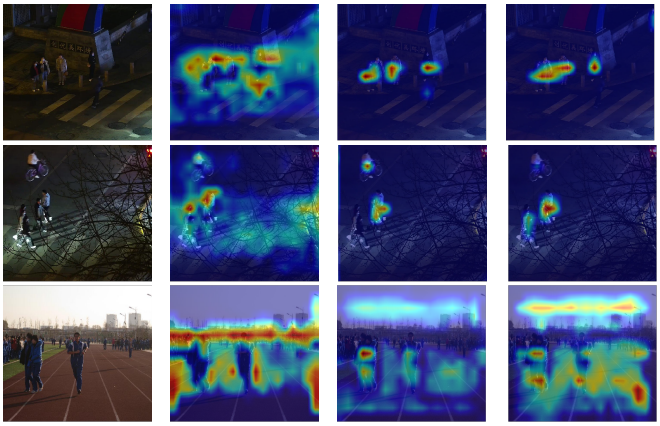
Method	DAY/%	NIGHT/%	ALL/%
YOLOv7_4c	72.73	80.35	76.01
YOLOv7+IC	73.98	84.09	79.90
YOLOv7+MFM	74.20	84.61	78.45
YOLOv7+DNL	73.10	80.53	76.56
YOLOv7+IC+MFM	79.87	86.11	81.39
YOLOv7+IC+DNL	74.41	84.05	80.04
YOLOv7+DNL+MFM	74.57	85.26	79.48
YOLOv7+DNL+MFM+IC(Ours)	81.47	86.50	82.96

attention mechanism. Thanks to the (15), the proposed L_{BCP} loss also has contributions to performance improvement, by distinguishing foreground from background as effectively as possible.

2) Multiscale Fusion Module (MFM): As shown, the one-branch methods are undoubtedly inferior to the two-branch approach. However, the crucial factor is fusion stage while halfway fusion structure achieves the best performance. For the two-branch method, we created two separate backbones to process visible and infrared images. It's worth noting that, during this experiment, we omitted the illumination compensation network. Both modalities were fed directly into their respective backbones. According to the results, it is evident that DFM plays a crucial role in improving detection performance, which resonates with our initial conjecture. Although DFM requires strictly aligned image pairs as input, this outcome provides strong experimental support for future research on pedestrian detection in more challenging environments.

F. Other Dataset

To demonstrate the generalization capability of our algorithm, we conducted experiments not only on the cleaned KAIST dataset, but also on another multispectral pedestrian detection benchmark called LLVIP. The majority of the LLVIP dataset were captured in low-light nighttime conditions. We



(a) original image (b) Dual Non-Local (c) Non-Local (d) SimAM

Fig. 8. The Grad-Cam visualizations between non-local, SimAM, and our dual non-local.

recorded the performance of the LLVIP dataset in Table VII, with mAP as the evaluation metric.

TABLE VII. COMPARISON ON LLVIP DATASET FOR ALL WEATHER SCENES IN TERMS OF AVERAGE PRECISION, RECALL, AND ACCURACY

Method	AP/%	Recall/%	Accuracy/%
YOLOv4 T+RGB	50.90	57.10	74.00
YOLOv7 T+RGB	79.60	71.89	94.34
Ours	83.76	78.16	97.81

V. CONCLUSION AND FUTURE WORK

In this paper, we investigated to integrate color-thermal image pairs effectively, leveraging the complementarity and exclusivity between modalities to enhance detection performance. We proposed an algorithm based on multiscale feature fusion. Specifically, we performed image enhancement on the input visible light image and simultaneously improved the Backbone network through integrating two modalities using differential information in Conv3, Conv4, and Conv5 convolutional layers. Our approach demonstrated outstanding performance on the cleaned KAIST and LLVIP datasets. Particularly in nighttime scenarios, we achieved a improvement of 7.28% in terms of Recall compared to the baseline on the cleaned KAIST dataset. We suggested that the proposed Dual Non-Local attention mechanism is also effective for other single object detection tasks, which is part of our future work. The findings of this paper offer a novel approach to combine image enhancement techniques and feature fusion for multispectral pedestrian detection, with potential applications beyond pedestrian detection. In our future work, we aim to further explore the complementarity between modalities and reduce redundant information between modalities in more challenging weather conditions, such as rain and snow scenarios.

REFERENCES

[1] F. Xu and K. Fujimura, "Pedestrian detection and tracking with night vision," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1, 2002, pp. 21–30 vol.1.

[2] A. Ćorović, V. Ilić, S. Đurić, M. Marijan, and B. Pavković, "The real-time detection of traffic participants using yolo algorithm," in *2018 26th Telecommunications Forum (TELFOR)*, 2018, pp. 1–4.

[3] W. Wang, Y. Peng, G. Cao, X. Guo, and N. Kwok, "Low-illumination image enhancement for night-time uav pedestrian detection," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5208–5217, 2021.

[4] S. S. S. Kruthiventi, P. Sahay, and R. Biswal, "Low-light pedestrian detection from rgb images using multi-modal knowledge distillation," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4207–4211.

[5] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster r-cnn for robust multispectral pedestrian detection," *Pattern Recognition*, vol. 85, pp. 161–171, 2019.

[6] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[7] T. Liu, K.-M. Lam, R. Zhao, and G. Qiu, "Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 315–329, 2021.

[8] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," 2009.

[9] S. Zhang, R. Benenson, B. Schiele *et al.*, "Filtered channel features for pedestrian detection." in *CVPR*, vol. 1, no. 2, 2015, p. 4.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[11] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2017.

[12] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[13] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[14] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 49–56.

[15] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[18] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[19] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[20] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *International conference on machine learning*. PMLR, 2021, pp. 11 863–11 874.

[21] H. Lee, K. Sohn, and D. Min, "Unsupervised low-light image enhancement using bright channel prior," *IEEE Signal Processing Letters*, vol. 27, pp. 251–255, 2020.

[22] Y. Wang, S. Zhuo, D. Tao, J. Bu, and N. Li, "Automatic local exposure correction using bright channel prior for under-exposed images," *Signal processing*, vol. 93, no. 11, pp. 3227–3238, 2013.

[23] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 787–803.

[24] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," *arXiv preprint arXiv:1808.04818*, 2018.

[25] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," in *2016 23rd International conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 621–626.

[26] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3496–3504.