electronics

MDPI

Article

Illumination-Aware Cross-Modality Differential Fusion Multispectral Pedestrian Detection

Chishe Wang [1,2], Jinjin Qian [1,*], Jie Wang [2] and Yuting Chen [1]

[1] School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China; wyxcs@163.com (C.W.)
[2] Jinling Institute of Technology, Nanjing 210001, China
* Correspondence: 2021201256@aust.edu.cn

Abstract: Multispectral information fusion technology is a practical approach to enhance pedestrian detection performance in low light conditions. However, current methods often overlook the impact of illumination on modal weights and the significance of inter-modal differential information. Therefore, this paper proposes a novel illumination-aware cross-modality differential fusion (IACMDF) model. The weights of the different modalities in the fusion stage are adaptively adjusted according to the illumination intensity of the current scene. On the other hand, the advantages of the respective modalities are fully enhanced by amplifying the differential information and suppressing the commonality of the twin modalities. In addition, to reduce the loss problem caused by the importance occupied by different channels of the feature map in the convolutional pooling process, this work adds the squeeze-and-excitation attention mechanism after the fusion process. Experiments on the public multispectral dataset KAIST have shown that the average miss rate of our method is substantially reduced compared to the baseline model.

Keywords: pedestrian detection; cross-modality; illumination aware; multispectral fusion; deep learning

Citation: Wang, C.; Qian, J.; Wang, J.; Chen, Y. Illumination-Aware Cross-Modality Differential Fusion Multispectral Pedestrian Detection. *Electronics* **2023**, *12*, 3576. https://doi.org/10.3390/electronics12173576

Academic Editor: Donghyeon Cho

Received: 8 August 2023
Revised: 22 August 2023
Accepted: 22 August 2023
Published: 24 August 2023

1. Introduction

Due to the influence of low light, backlight, uneven light, and dim light, the quality of images/video generation is poor, seriously affecting the performance of image-based object detection [1]. One solution is to improve the quality of the image using low-light image enhancement algorithms such as LLNet [2], ABSGNet [3], and URetinex-Net [4]. Another method utilizes multiple sensors to obtain information about detected objects in a light-varying environment. Pedestrian detection is one of the critical application fields of computer vision. However, traditional computer-vision-based pedestrian detection models have relatively high requirements for the circumstances of the input images or videos [5,6]. With the spread of intelligent surveillance and autonomous driving, the demands for pedestrian detection in complex scenarios (e.g., low illumination conditions) are becoming increasingly urgent [7]. Integrating another modality can significantly improve the efficiency of the pedestrian detection task compared to a single visible modal information input. For example, spectral images can detect a substance's light radiation and reveal the target object's basic colour properties. In contrast, thermal images can be acquired based on the thermal radiation difference of an object without relying on an external light source [8–11]. Consequently, it is complementary to visible modal detection due to its ability to capture thermal features under poor lighting conditions. As shown in Figure 1, the visible light sensor does not detect all pedestrians on a poorly lit road at night. In contrast, in the same scenario, using an infrared camera can identify pedestrians on the dark side of the road.

Figure 1. Images with different modes in the same scene. **Left:** image acquired in visible light mode. **Right**: image captured in infrared light mode in the same scene.

The critical issue in the current research on multispectral pedestrian detection is how to fuse multimodal information. At what stage to fuse visible and thermal image information is the first concern to be considered for combined multimodal detection [12,13]. According to the stages of model processing, there are three major types: early fusion, halfway fusion, and late fusion [14]. Early fusion at the model level refers to integrating features immediately after they are extracted. Since deep learning will inherently involve learning specific representations of features from raw data, this leads to the fact that fusion may sometimes be required before features are extracted, i.e., data fusion. Therefore, both feature-level and data-level fusion is referred to as early fusion. Halfway fusion is performed during the feature extraction process by the model to feed into the detector. Late fusion is also called decision-level fusion, where a deep learning model is first trained on different modalities and then fuses the outputs of multiple models. Ref. [12] elaborated on four convolutional neural network fusion architectures that integrate dual branches of convolutional neural networks in different DNN (deep neural network) stages. The experimental results demonstrate that halfway fusion is superior to other fusion approaches (e.g., early fusion, late fusion, and score fusion). Early multimodal data fusion does not fully demonstrate the complementarity between the modalities and may lead to redundant vector inputs. In contrast, late fusion has high requirements for the strategy of fusion. Six fusion architectures that fuse visible and thermal modalities at different stages were compared in [15], leading to the same conclusion. In the following years, the advantages of halfway fusion were demonstrated in [16–20]. Therefore, the fusion strategy of the model in this paper is also chosen in this way. Integrating multispectral information more effectively in the halfway fusion process has recently become the research focus. In [21], the region proposal network (RPN) was utilised as a feature extraction module based on halfway fusion and classification using a boost decision tree (BDT) to improve the performance of the pedestrian detector. A gated fusion unit between the two single detectors was proposed in [22] to fuse visible and thermal spectral image features efficiently. Nevertheless, most existing methods integrate the information of two modalities without considering the influence of current scene factors (e.g., illumination) on inter-modal fusion. In well-lit scenes, the visible modal information plays a more significant role than the infrared mode. Conversely, the infrared light information should dominate in low-illumination scenes. Therefore, the integration mechanism should be adaptive rather than static. In addition, since the visible and infrared light image pairs are in the same scene, there is a large amount of redundant information. We argue that the fusion of features between modes should not just be superimposed or concatenated; more attention should be paid to the differences between the two, since differential information from the same scene can better show the strengths of the respective modalities. Furthermore, involving the lighting conditions of the current scene can further extend the advantages of the differentiating information of the modes.

From this perspective, this paper proposes the illumination-aware cross-modality differential fusion model. Specifically, information from multispectral images is extracted

separately by a dual-branch convolutional network, with halfway fusion having been selected. The features of different modalities are passed through the cross-modal differential fusion module to fully amplify the differential information of sensors and suppress the same information. We trained the illumination intensity classification sub-network in advance (see Section 3.3 for details). The illumination intensity information of the current scene is obtained by passing the visible image of the scene through this pre-trained sub-network. The illumination information adds a weight value to the features of both modalities. The processed multi-channel feature map is followed by adding the SE [23] attention mechanism to mitigate the loss caused by the different importance occupied by different channels of the feature map during the convolution pooling process and, finally, fed to the detector. This paper's main contributions can be summarised as follows:

1. We explore a new mechanism for the adaptive calculation of the two modal weights required depending on the different illumination levels of different scenery.
2. This paper proposes the IACMDF algorithm that combines the illumination information of a scenario with the differential information of different modalities of the same scene.
3. The experimental results show that the proposed method is competitive and performs better than the baseline methods.

The remainder of this paper is as follows. Section 2 presents an overview of related work on multispectral pedestrian detection. We present a detailed description of our proposed IACMDF model in Section 3. Section 4 illustrates how the proposed method compares with other methods on a benchmark dataset, as well as some exploratory and ablation experiments. Section 5 summarises our conclusions.

2. Related Work

In this section, we first review the main multispectral pedestrian methods based on visible and infrared light. Afterwards, this work focus on analysing the illumination-aware modal weight recalculation strategy during network fusion and exploring the potential scope for improvement.

2.1. Multispectral Pedestrian Detection

In recent years, multispectral pedestrian detection has received increasing attention from researchers. In 2015, Hwang et al. [24] proposed the KAIST multispectral dataset. They extended the aggregated channel feature pedestrian detector to extend the ACF [25] method by enhancing the thermal intensity of the thermal images and HOG (histogram of oriented gradient) features as additional channel features. Liu et al. [12] demonstrated that halfway fusion is more advantageous than other fusion strategies by designing four CNN-based architectures that fuse features extracted from two subnets. MSDS-RCNN [26] adopted a halfway fusion architecture with dual streams to combine the pedestrian detection task with the semantic segmentation task to improve detection accuracy. Since then, halfway fusion has become the default strategy in deep-learning-based multispectral work [27–29]. Zheng et al. [22] proposed a novel gated fusion unit (GFU) based on halfway fusion, which learns a combination of feature maps generated by coloured SSD [30] and thermal SSD. CIAN [31] proposed a cross-modal interaction attention mechanism to exploit modal correlations and adaptive fusion features. Kim et al. [32] deployed EfficientDet as a backbone and proposed a fusion framework for multispectral pedestrian detection based on EfficientDet, which improves the detection accuracy of pedestrians in visible and thermal images by adding and cascading visible and thermal features. Zhang et al. [33] proposed guided attention feature fusion (GAFF) to guide multispectral feature fusion. GAFF achieves a fully adaptive fusion of thermal and visible features without hand-made assumptions or additional annotations. Kim et al. [34] proposed a novel single-stage detection framework that used multi-label learning to learn input-state-aware features by assigning an individual label based on a given state of the input image pair. Uncertainty-aware cross-modal guidance (UCG) [35] modules have been proposed to encode the similarity of feature distributions

by guiding the two modalities to obtain more distinguishing features. Zhang et al. [36] introduced the regional feature alignment (RFA) module, which adaptively compensates for feature mapping misalignment in two ways. To handle the position shift problem in visible and thermal image alignment, AR-CNN designed a region feature alignment module to align the region features of both modalities. Shojaiee et al. [37] extracted different feature vectors from the visible and thermal domains of a nighttime pedestrian head image. The features from both parts were fused at the feature level. Roszyk et al. [38] designed various fusion structures and demonstrated the advantages of halfway fusion. Yang et al. [39] proposed a cascaded information enhancement module. From the perspective of feature fusion, the interference of colour and thermal modal backgrounds on pedestrian detection was reduced, and an attention mechanism enhanced pedestrian features of colour and thermal modal backgrounds. Most of these methods are based on a simple dual-branch architecture. They do not explore the correlation between modalities at a deeper level nor associate the scene's illumination with multimodal fusion.

2.2. Illumination Awareness in the Fusion Model

Several researchers have realised that the illumination factor plays a substantial role in multispectral fusion models. Different illuminance values determine new fusion weights during the fusion process. Chen et al. [15] introduced an illumination-aware network to give the illumination metric of the input image, adaptively merging the visible and thermal networks employing a gate function defined over the illumination values. Guan et al. [40] proposed a multispectral pedestrian detection framework based on light-aware pedestrian detection and semantic segmentation. They used a novel light-aware weighting mechanism to describe the light conditions and integrated illumination information into a dual-stream CNN to obtain human-related features under different light conditions. MBNet [41] facilitates the optimisation process in a more flexible and balanced way, improving the detector's performance, utilising a light-aware feature alignment module that adaptively selects complementary information based on lighting conditions. To overcome the effects of environmental factors such as real time and resistance to low light conditions, Zhuang et al. [42] proposed a lightweight one-stage illumination and temperature-aware multispectral network (IT-MN). Yang et al. [29] proposed an efficient cross-modal fusion module called bidirectional adaptive attention gate (BAA-Gate). Based on the adaptive interaction of BAA-Gate with the illumination weighting strategy, it adaptively adjusts the recalibration and aggregation intensity in BAA-Gate to enhance the robustness to illumination changes. Tang et al. [43] utilise the light probability to construct an illumination aware loss to guide the training of the fusion network, the cross-modal differential aware fusion module and the semi-modal fusion strategy fuse entirely the joint and complementary information under the light perception loss constraint. Yan et al. [14] incorporate an adaptive illumination-aware weight generation module to lift the contributions of visible and thermal modalities by extracting the illumination information from the two modalities. These integrate scene illumination information directly into the fused network, taking little account of the variable weighting of the scene information obtained by different sensors at different illumination levels.

3. Methods

In this section, details of the proposed illumination-aware cross-modality differential fusion (IACMDF) multispectral pedestrian detection framework are provided. The overall architecture of the model is first described, followed by the IACMDF module formed by combining the illumination-aware (IA) module with the cross-modality differential fusion (CMDF) module. Finally, the work introduces the illumination intensity classification sub-network required in the IACMDF.

3.1. Overall Architecture

How to adaptively fuse multispectral image pairs has been the most widely researched topic in multispectral pedestrian detection. Researchers have experimentally demonstrated that halfway fusion is more effective than early and late fusion. Meanwhile, most researchers have adopted some representative detection models as the baseline network for fusion algorithms. The choice of a representative baseline algorithm makes it easy to conclude whether or not the proposed model is suitable. The proposed detection model is also based on an SSD-like halfway fusion model.

Figure 2 illustrates the overall architecture of the proposed dual-stream multispectral pedestrian detection model based on halfway fusion. The model consists of feature extraction, feature fusion, and a detection head. The input to the model requires pairs of visible and thermal images of the same scene. Features are extracted separately using independent dual-branch backbone networks. The backbone network selected for comparison in this work was a VGG16 network pre-trained on ImageNet. The extracted features are fed into the IACMDF module for fusion at the conv4_3, conv6, conv7, conv8, conv9, and conv10 layers. The SE attention mechanism is added after all the fused feature maps, and finally, the SSD-like detection head is input to obtain the result.

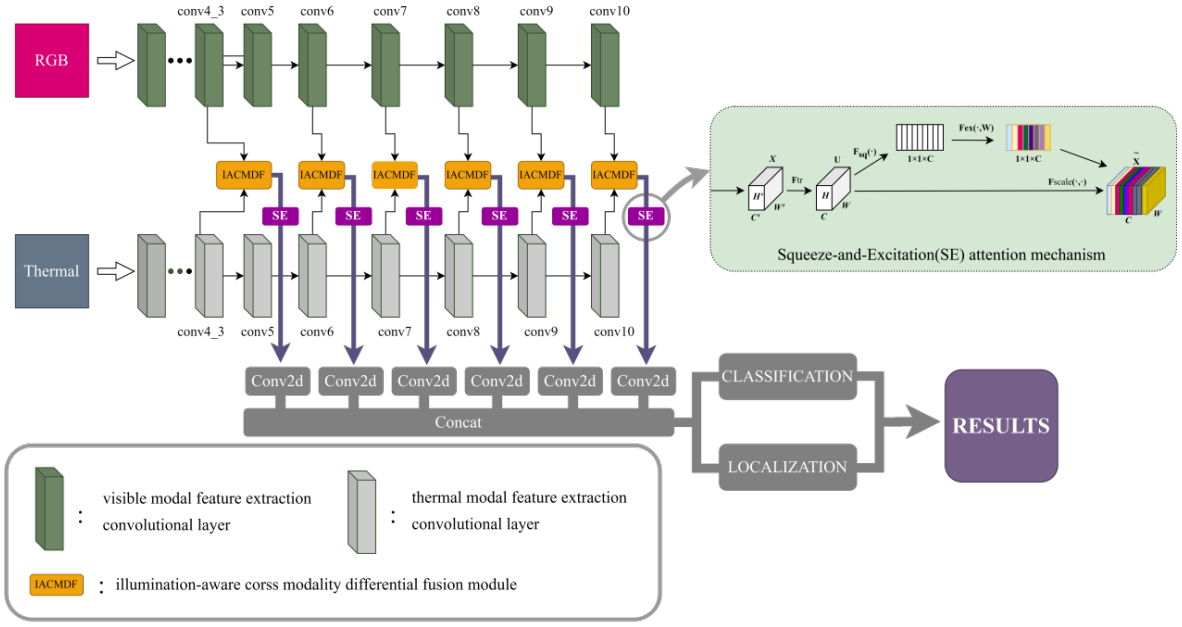

Figure 2. The overall architecture of the proposed model.

In some previous multispectral pedestrian detection models, convolutional layers with shared parameters were used in the feature extraction stage to improve detection efficiency, which reduced the feature extraction capability. So, this paper uses mutually independent dual-branch networks using independent convolution to extract features from multispectral images. As shown in Equation (1), where F represents the feature map after the fusion of the two modalities, f_s and f_i denote the mode-shared and mode-independent convolution parts, respectively. I_V and I_T indicate the image information of visible and thermal optics, respectively, and $\otimes$ denotes a fusion mechanism. This model architecture reduces the number of parameters in the model to a certain extent but at the expense of detection accuracy. As shown in Equation (2), the proposed model is more concerned with the mechanism of fusion, focusing on making its fusion strategy more effective, which means it performs better in cross-sectional comparisons with other baseline models using the same evaluation metrics on public data sets.

$$F = f_s[f_i(I_v) \otimes f_i(I_t)] \quad (1)$$

$$F = f_i(I_v) \otimes f_i(I_t) \tag{2}$$

3.2. IACMDF Module

Effectively fusing the two information modes becomes the most critical task for multimodal pedestrian detection. Different scenes have different illumination rates. Thus, the contribution of the information in the two image modes is not the same. Furthermore, the contribution of the individual channels of the picture varies at different illumination levels from a microscopic point of view. Adding or multiplying functionality can lead to worse results than in a single mode. Inspired by MBNet [41], we take the principle of utilising differential amplification circuits. By complementing the two modal features and augmenting one modality with the other, the algorithm can increase its sensitivity to the information of the other modal feature and improve the degree of information interaction between the two channels.

The proposed IACMDF module is a further exploration on this basis. As shown in Figure 3, the IACMDF module consists of a hot-swappable combination of a cross-modality differential fusion module (CMDF) and an illumination-aware module (IA). The differential information is a well-represented difference between the two modalities. The CMDF module is to amplify the difference between the two modes to improve the overall performance. Furthermore, the CMDF module links the *RGB* feature extractor and the thermal feature extractor. It should play a balanced role in the impact of both modes. F_D represents the difference between the visible light's and infrared light's modal picture features and is designed for the symmetry of the CMDF module. To enable the integration of the differential information into the corresponding channels, the new fused features are obtained by global average pooling of the differential information, activating it, and then dot-multiplying it with the modal differences, and finally, summing it with the respective channels. Global average pooling converts the feature map into a global representation to apply attention at the channel level. So the purpose of using global averaging in pooling here is to compress the F_D into a global difference vector, which can be interpreted as a channel descriptor with statistics representing the difference between the *RGB* and thermal modes. Equations (3) and (4) show that $F_R^{'}$ and $F_T^{'}$ represent the features of the processed visible and thermal images, respectively. *GAP* represents global average pooling, and σ is the *sigmoid* activation function. F_R and F_T denote the features of the visible and thermal images, respectively.

$$F_R^{'} = F_R + [\sigma(GAP(F_T - F_R)) \times (F_T - F_R)] \tag{3}$$

$$F_T^{'} = F_T + [\sigma(GAP(F_R - F_T)) \times (F_R - F_T)] \tag{4}$$

Since the illumination intensity of different scenes is different, the weights of the two models should be determined according to the current scene illumination information. The overall role of the IA module is to obtain the sizes of the weights required for each modality to be fused by the illumination of the current scene. The IA module comprises a pre-trained illumination intensity sub-network and an illumination gate. Using the pre-trained illumination intensity classification sub-network, we can obtain two values from 0 to 1 representing the light intensity from the input visible light modal image. These two values represent the probability that the current scene is strongly or weakly illuminated.

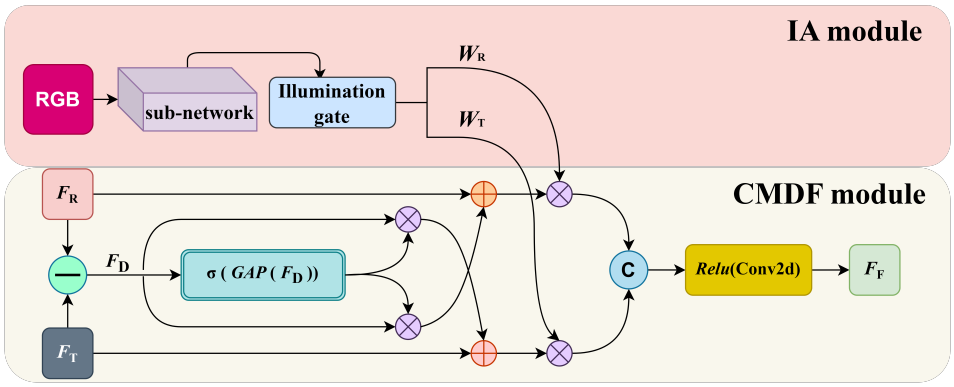

Figure 3. Illustration of the proposed IACMDF module.

As shown in Figure 4, the illumination gate calculates the required weighting based on the values obtained from the illumination intensity sub-network. The value of I_C is the intensity classification, and I_V denotes the intensity value of the illumination. If the light intensity strong value is higher than the weak, I_C is positive and the value of I_V is equal to the strong value. Conversely, I_C is minus one, and I_V is the weak value. The weight values for the two modalities are mapped through the value of I_V, essentially the probability value of classifying the current scene as strong or weak in light intensity. W_R and W_T represent the weight of the visible and infrared light modes at the fusion time, respectively. Equation (5) shows the procedure for obtaining the W_R values. This work has controlled the final output value to sbetween 0.85 and 1.15 by adding a limiting factor α. Specifically, the range of I_V values is [0, 1]. The value of the limiting factor α has been set to 0.15 based on extensive experiments. The possible values of I_C are positive or negative, multiplying them together and adding 1 to the range of values [0.85, 1.15]. Equation (6) shows the procedure for obtaining the W_T values.

$$W_R = 1 + \alpha \times I_C \times I_V \tag{5}$$

$$W_T = 2 - W_R \tag{6}$$

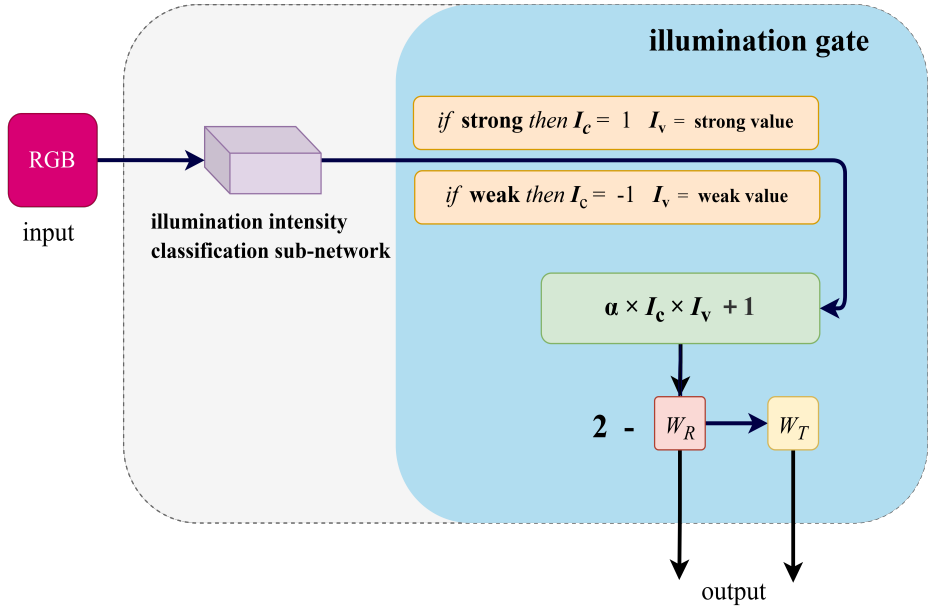

Figure 4. Details of the illumination gate interior.

The obtained weight values are multiplied with the processed visible and thermal light image features and then connected to obtain the fused feature map. Finally, we do not feed this feature map directly into the detector but perform another 3×3 convolution with *ReLU* activation. As shown in Equation (7), where the ϕ denotes concatenating the re-weighted features and F_F is the final weighted fusion feature map.

$$F_F = Relu(Conv2d(\phi(W_R F_R^{'}, W_T F_T^{'}))) \tag{7}$$

The detection head uses multiple features fused with different resolution maps as input to detect pedestrians of different sizes. The loss term for localisation is the same as SSD [30]. For classification loss, the network adopted a binary cross-entropy loss function in an end-to-end manner. The prediction scores in the classification task are calculated by taking the average *GRB* and hot confidence scores corresponding to the same bounding box. The final loss item is the sum of the two loss items in a ratio of 1:1.

3.3. Illumination Intensity Classification Sub-Network

The objective of the illumination intensity classification sub-network is to estimate the light intensity of a scene. The input to the network is the visible light image. The output is two values, i.e., the probability of the current scene being strongly or weakly illuminated, normalised by the softmax function to [0,1]. The architecture of the network is shown in Figure 5. In order not to consume too many computational resources on this sub-network, the illumination intensity classification sub-network consists of four convolutional layers, a global average pooling layer, and two fully connected layers. The numbers of convolutional kernels in the four convolutional layers are 16, 32, 64, and 128, respectively, and the size of the convolutional kernels is 3×3. The stride size is set to 2. This work used *Leaky ReLU* as the activation function. A global average pooling operation is used in the middle of the convolution and fully connected layers to integrate image feature information. Finally, two fully connected layers calculate the illumination intensity probability. The illumination intensity sub-network is essentially a classifier. Therefore, this work used cross-entropy loss to constrain the training process of the sub-network, as shown in Equation (8), where $\widetilde{y}$ represents the label of the input image, encoded by one-hot, the training labels with intense illumination are set to [1,0] and the labels with weak illumination are set to [0,1], the y denotes the sub-network probability value of the solid or weak illumination output. The Θ refers to the *softmax* function.

$$L = -\widetilde{y} \log \Theta(y) - (1 - \widetilde{y}) \log(1 - \Theta(y)) \tag{8}$$

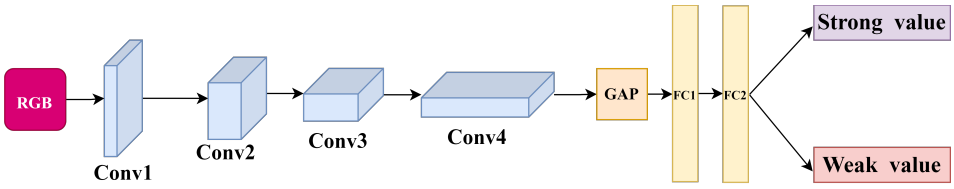

Figure 5. Illumination intensity sub-network structure.

4. Experiments

In this section, we first present the multispectral pedestrian dataset and evaluation metrics and give the implementation details. The experimental results on the public dataset are then compared with the baseline multispectral detection models. Finally, the validity of each proposed component is verified by an ablation study.

4.1. Dataset and Metrics

This study requires a specific dataset of visible images paired with thermal images. Therefore, the KAIST [24], CVC-14 [44], and LLVIP [45] multispectral pedestrian datasets

were chosen to evaluate our model. The KAIST multispectral pedestrian dataset comprises 95,328 aligned visible and thermal paired images of the same scene in different urban environments. A total of 1182 pedestrian instances with 103128 bounding boxes were included. In the training phase, one frame was taken from every two frames to obtain 25,076 images as the training subset. In the test phase, one frame was extracted from every 20 frames, and a total of 2252 frames were obtained, of which 1455 frames were obtained during the day and 797 frames during the night. Considering that the original dataset was sampled to some extent as the images were taken from successive frames of the video and there was no difference between adjacent images, and some of the annotations in the original dataset were incorrect, the improved annotations proposed in [36] were used for training. The evaluation process used the improved test annotations from ref. [26].

The CVC-14 dataset contains visible and thermal paired images and is a multispectral pedestrian dataset taken with a stereo camera configuration. This dataset's training and test sets contain 7085 and 1433 images, respectively. The training set contains 3695 daytime and 3390 nighttime images, including annotated pedestrians in 1500 daytime and nighttime images. The test set contains 1433 images, of which 706 were taken during the day and 727 at night. The annotations are provided separately for each mode as the camera is not calibrated.

The LLVIP dataset is a recently proposed visible–infrared paired pedestrian dataset from Jia et al. Most of this dataset was taken in low light conditions. It contains 33,672 images. All images are strictly aligned. The pedestrians in the dataset are labelled. Each image pair is collected by a binocular camera consisting of a visible light camera and an infrared camera.

The evaluation metric is the log-average miss rate (LAMR), which represents the averaging of the miss rate of FPPI (false positive per image) in the range of [0.01, 1] for nine points uniformly taken in logarithmic space, as suggested by Dollar et al. [46]. The lower the average miss rate, the better the algorithm's performance. LAMR is one of the most popular metrics for pedestrian detection tasks. This work furthermore evaluates the model using mean average precision (mAP), where higher scores indicate better performance. The mAP represents the average of the mAP with an IoU threshold of 0.5 to 0.95 at intervals of 0.05. mAP50 denotes the mAP at an IoU threshold of 0.5, and mAP75 refers to the mAP with an IoU threshold of 0.75.

4.2. Implementation Details

The experimental test environment was the Ubuntu 16.04 operating system, and the server hardware configuration was NVIDIA P100 graphics cards. The illumination intensity sub-network model was first trained and supervised using visible light images from the training set. The classification categories were divided into strong and weak categories based on the light intensity in the photo scene. The Adam optimizer updated the model parameters, and the learning rate was first initialised to 0.001 and then decayed exponentially. We trained the sub-network batch size set to 16 for 20 epochs. When training the detection model, classification results were obtained directly from the input visible light image for that image's illumination.

The detection model adopted was a modified PyTorch-based SSD algorithm model, and the backbone network used VGG16 pre-trained on ImageNet. We have redesigned the feature aggregation module by adding a fusion layer. Based on the features of human appearance, we set the predefined anchor frame parameters to have aspect ratios of $\frac{1}{1}$ and $\frac{1}{2}$, fine scales of (2^0,$2^{\frac{1}{3}}$,$2^{\frac{2}{3}}$), and scale levels of 40, 80, 160, 200, 280, and 360. The initial learning rate, momentum, and weight decay were 0.0001, 0.9, and 0.0005, respectively. The model was trained using SGD (stochastic gradient descent) with a batch size set to 16 for 50 epochs. The image input size was adjusted to 512 (H)×640 (W). To prevent distortion in the augmented images, we used geometric transformations such as horizontal flips and random resizing crops with a probability of 0.5.

4.3. Comparison Experiments

***KAIST*:** The proposed model was trained and tested on the KAIST dataset and compared with other excellent multispectral pedestrian detection models, including ACF + T + THOG (optimised) [24], halfway fusion [12], IAF-RCNN [16], CIAN [31], MSDS-RCNN [26], AR-CNN [36], MBNet [41], and MLPD [34]. As shown in Table 1 and Figure 6, the proposed model is superior in terms of detection performance when compared to current excellent multimodal algorithms. In the case of an IoU (intersection over union) threshold of 0.5, compared to previous methods such as MLPD, the average miss rate of the proposed IACMDF fusion model is 1.41% lower overall, 0.89% lower during the day, and 2.62% lower at night. Furthermore, some qualitative results on KAIST are shown in Figure 7. The proposed IACMDF locates pedestrians more accurately than other pedestrian detectors. In multispectral pedestrian detection, a suitable fusion strategy plays a vital role in obtaining the maximum performance gain. We argue that the combination of scene illumination intensity information with inter-modal differential information can fully amplify the advantages of each modality. To explore the possibility of obtaining better detection results, we tested the VGG16 and ResNet50 backbone networks separately in our model. The experiments showed that VGG16 was slightly better than the ResNet50 network on the proposed model. We also evaluated the proposed model on different subsets, which included different scale levels and occlusion levels of the KAIST dataset, as shown in Table 2. Compared to the baseline model, the proposed method still achieves relatively better results. This also validates the robustness and generality of the proposed method.

Table 1. Evaluation results on the KAIST dataset.

Method	Backbone	Miss Rate (IoU = 0.5) (%)			Platform	Speed (s)
		All	Day	Night		
ACF [24]	-	47.32	42.57	56.17	MATLAB	2.73
Halfway fusion [12]	VGG16	25.75	24.88	26.59	TITAX	0.43
IAF R-CNN [16]	VGG16	15.73	14.55	18.26	TITAX	0.25
CIAN [31]	VGG16	14.12	14.77	11.13	1080 Ti	0.07
MSDS-RCNN [26]	VGG16	11.34	10.53	12.94	TITAX	0.22
AR-CNN [36]	VGG16	9.34	9.94	8.38	1080 Ti	0.12
MBNet [41]	ResNet50	8.13	8.28	7.86	1080 Ti	0.07
MLPD [34]	VGG16	7.58	7.95	6.95	2080 Ti	**0.012**
IACMDF (ours)	ResNet50	6.93	7.36	6.11	P100	0.048
IACMDF (ours)	VGG16	**6.17**	**7.06**	**4.33**	P100	0.034

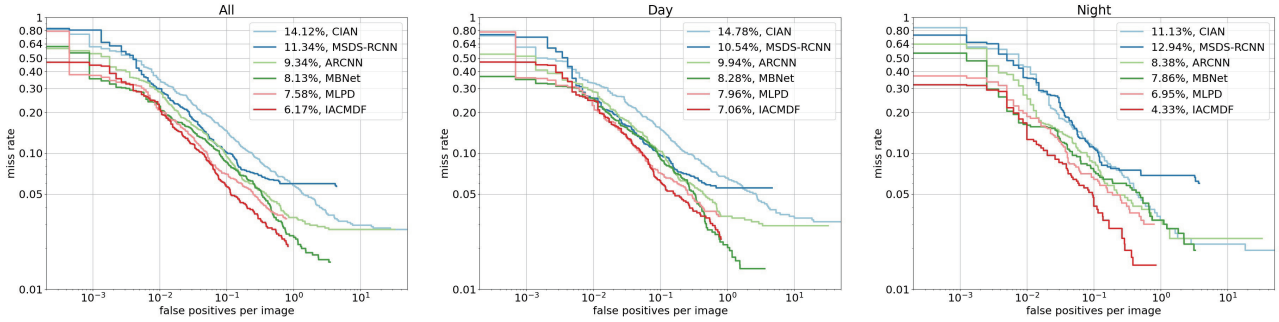

Figure 6. Comparison of detection results on the test set of the KAIST multispectral benchmark, in terms of all-day (**left**), daytime (**middle**), and nighttime (**right**). Compared to other methods, the proposed model provides consistently stable performance across all scenarios.

Table 2. Evaluation results on different scale and occlusion levels of the KAIST dataset.

Method	Scale (MR%)			Occlusion (MR%)		
	Near	Medium	Far	No	Part	Heavy
ACF [24]	28.74	53.67	88.20	62.94	81.40	88.08
Halfway fusion [12]	8.13	30.34	75.70	43.13	65.21	74.36
IAF R-CNN [16]	0.96	25.54	77.84	40.17	48.40	69.76
CIAN [31]	3.71	19.04	55.82	30.31	41.57	62.48
MSDS-RCNN [26]	1.29	16.19	63.73	29.86	38.71	63.37
AR-CNN [36]	**0.00**	16.08	69.00	31.40	38.63	55.73
MBNet [41]	**0.00**	16.07	55.99	27.74	35.43	59.14
MLPD [34]	**0.00**	15.89	57.45	26.46	35.34	60.28
IACMDF (ours)	**0.00**	**15.15**	**55.16**	**24.97**	**33.22**	**53.88**

In addition, we provide the mAP50 metrics on the KAIST dataset in Table 3 to further demonstrate the effectiveness and progressiveness of the proposed method. The results of the proposed method and the benchmark method on the KAIST dataset show that the proposed method obtains better results.

Table 3. The mAP50 evaluation metrics on the KAIST dataset.

Method	mAP50 (%)
ACF [24]	40.00
Halfway fusion [12]	57.24
IAF R-CNN [16]	56.62
CIAN [31]	69.00
MSDS-RCNN [26]	67.19
AR-CNN [36]	75.31
MBNet [41]	75.93
IACMDF (ours)	**76.22**

***CVC-14*:** In order to verify the robustness of the model, we used the trained KAIST model on the CVC-14 dataset to fine-tune the training and compare it with other methods. We adopted the protocol introduced by [9,34,36,41] for these methods to make a fair comparison. As shown in Table 4, the proposed method still achieves competitive results compared to suitable benchmark methods.

Table 4. Evaluation results on the CVC-14 dataset.

Method (Grey + Thermal)	Miss Rate (%)		
	All	Day	Night
MACF [9]	69.71	72.63	65.43
Halfway fusion [9]	31.99	36.29	26.29
Park et al. [9]	26.29	28.67	23.48
AR-CNN [36]	22.1	24.7	18.1
MBNet [41]	**21.1**	24.7	13.5
MLPD [34]	21.33	**24.18**	17.97
IACMDF (ours)	21.15	25.06	**13.31**

***LLVIP*:** Table 5 shows the detection results of the proposed IACMDF and other unimodal networks. By fusing the complementary features of RGB and thermal modes based on the IACMDF module, the mAP is higher than that of the unimodal-based detection model at different thresholds of the IoU. We also compare with the multimodal-based CFT model, which IACMDF improves on by 1.2% at mAP75 and 1.5% at mAP. This shows that the proposed method can be well generalised to different types of images.

Table 5. Evaluation results on the LLVIP dataset.

Method	Dataset	mAP50 (%)	mAP75 (%)	mAP (%)
SSD [47]	RGB	82.6	31.8	39.8
SSD [47]	Thermal	90.2	57.9	53.5
YOLOv3 [45]	RGB	85.9	37.9	43.3
YOLOv3 [45]	Thermal	89.7	53.4	52.8
YOLOv5 [45]	RGB	90.8	51.9	50.0
YOLOv5 [45]	Thermal	94.6	72.2	61.9
CFT [47]	RGB + Thermal	**97.5**	72.9	63.6
IACMDF (ours)	RGB + Thermal	97.3	**74.1**	**65.1**

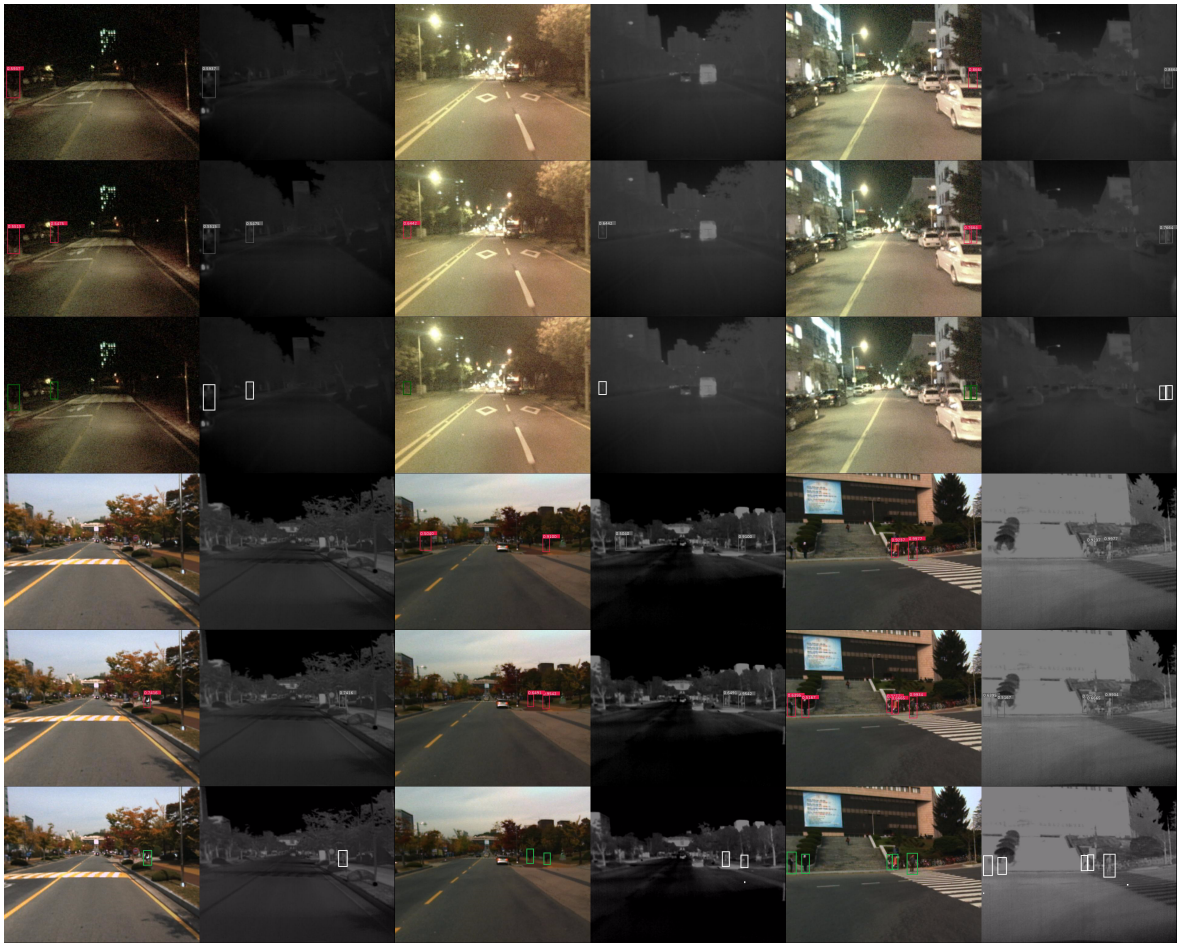

Figure 7. Qualitative comparisons of detection results on the test set of KAIST multispectral datasets. The top three rows are at night, and the bottom are during the day. Each group from top to bottom: MLPD, IACMDF, ground truth.

4.4. Ablation Study

To analyse the proposed model in more detail and to verify the effectiveness of the IACMDF module added in this paper, we conducted ablation experiments on the KAIST dataset. Table 6 gives the experimental results for the baseline dual-branch network, the addition of the SE attention mechanism module only, the addition of the CMDF module, and the addition of the complete IACMDF module. There was some improvement in the detection performance after adding only the SE attention mechanism module, with an overall MR reduction of 0.31%. The value of the average miss rate decreases significantly with the addition of the CMDF module, and the model shows excellent detection performance with the addition of the entire IACMDF module. The value of the miss rate for pedestrians

was reduced by 2.02% compared to the baseline network in the overall situation, with a 1.39% decrease during the day and 2.68% decrease at night. This ablation experiment proves the effectiveness of the added module.

Table 6. Ablation study on the KAIST dataset.

IA	CMDF	SE	Miss Rate (%)		
			All	Day	Night
-	-	-	8.19	8.45	7.01
-	-	✓	7.88	7.95	6.49
-	✓	✓	6.53	7.18	5.07
✓	✓	✓	**6.17**	**7.06**	**4.33**

5. Conclusions

In this paper, we proposed a framework for the illumination-aware cross-modality differential fusion model. The framework extracts the features of different modal images' information through a dual-branch deep learning network first, and adaptively calculates the weight values of the two modalities utilising a scene light intensity classification sub-network with light values, and combines the difference information of the features of different modalities of the same scene using a halfway fusion method. The fused features are added to the SE attention mechanism before being fed into the final detection head, ultimately improving the robustness of pedestrian detection in scenes with different illumination scales, to reduce the losses caused by the different importance of the different channels of the feature map during convolutional pooling. The experimental results show that the proposed IACMDF model has a significant advantage in the average miss rate metric compared to other baseline detectors. In practical applications, it is first necessary to use two different modalities of the camera to capture information from the same scene. Moreover, the information from the different modalities needs to be strictly matched. However, due to various practical reasons such as camera failure, shake, etc., modal photographs from multiple sensors may not be strictly paired, which can have an impact on the actual detection results. In future work, we will explore solutions to the unpaired problem.

Author Contributions: Formal analysis, C.W.; project administration, C.W.; software, J.Q; writing—original draft, J.Q; writing—review, J.W.; investigation, J.Q. and Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Ministry of Transport's Industry Key Science and Technology Project (Project No. 2020-ZD3-029), and 2021 Nanjing Municipal Industry and Information Technology Development Special Fund Project (Project name: Construction of 5G-based Application Scenarios for Digital Operation and Control of Intelligent Transportation).

Data Availability Statement: All implementation details, sources, and data are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, C.; Guo, C.; Han, L.; Jiang, J.; Cheng, M.M.; Gu, J.; Loy, C.C. Low-Light Image and Video Enhancement Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 9396–9416. [CrossRef] [PubMed]
2. Lore, K.G.; Akintayo, A.; Sarkar, S. LLNet: A Deep Autoencoder Approach to Natural Low-light Image Enhancement. *Pattern Recognit.* **2017**, *61*, 650–662. [CrossRef]
3. Chen, Z.; Liang, Y.; Du, M. Attention-based Broad Self-guided Network for Low-light Image Enhancement. In Proceedings of the 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 31–38. [CrossRef]
4. Wu, W.; Weng, J.; Zhang, P.; Wang, X.; Yang, W.; Jiang, J. URetinex-Net: Retinex-Based Deep Unfolding Network for Low-Light Image Enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5901–5910. [CrossRef]

5. Wagner, J.; Fischer, V.; Herman, M.; Behnke, S. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 27–29 April 2016; Volume 587, pp. 509–514.
6. Liu, T.; Lam, K.M.; Zhao, R.; Qiu, G. Deep Cross-Modal Representation Learning and Distillation for Illumination-Invariant Pedestrian Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 315–329. [CrossRef]
7. Dasgupta, K.; Das, A.; Das, S.; Bhattacharya, U.; Yogamani, S. Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15940–15950. [CrossRef]
8. Xu, D.; Ouyang, W.; Ricci, E.; Wang, X.; Sebe, N. Learning cross-modal deep representations for robust pedestrian detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5363–5371. [CrossRef]
9. Park, K.; Kim, S.; Sohn, K. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognit.* **2018**, *80*, 143–155. [CrossRef]
10. Dai, X.; Yuan, X.; Wei, X. TIRNet: Object detection in thermal infrared images for autonomous driving. *Appl. Intell.* **2021**, *51*, 1244–1261. [CrossRef]
11. Cao, Y.; Luo, X.; Yang, J.; Cao, Y.; Yang, M.Y. Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection. *Inf. Fusion* **2022**, *88*, 1–11. [CrossRef]
12. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. *arXiv* **2016**, arXiv:1611.02644.
13. Song, X.; Gao, S.; Chen, C. A multispectral feature fusion network for robust pedestrian detection. *Alex. Eng. J.* **2021**, *60*, 73–85. [CrossRef]
14. Yan, C.; Zhang, H.; Li, X.; Yang, Y.; Yuan, D. Cross-modality complementary information fusion for multispectral pedestrian detection. *Neural Comput. Appl.* **2023**, *35*, 10361–10386. [CrossRef]
15. Chen, Y.; Xie, H.; Shin, H. Multi-layer fusion techniques using a CNN for multispectral pedestrian detection. *IET Comput. Vis.* **2018**, *12*, 1179–1187. [CrossRef]
16. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171. [CrossRef]
17. Zhang, H.; Fromont, E.; Lefevre, S.; Avignon, B. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; IEEE: New York, NY, USA, 2020; pp. 276–280. [CrossRef]
18. Wolpert, A.; Teutsch, M.; Sarfraz, M.S.; Stiefelhagen, R. Anchor-free small-scale multispectral pedestrian detection. *arXiv* **2020**, arXiv:2008.08418.
19. Pei, D.; Jing, M.; Liu, H.; Sun, F.; Jiang, L. A fast RetinaNet fusion framework for multi-spectral pedestrian detection. *Infrared Phys. Technol.* **2020**, *105*, 103178. [CrossRef]
20. Cao, Z.; Yang, H.; Zhao, J.; Guo, S.; Li, L. Attention fusion for one-stage multispectral pedestrian detection. *Sensors* **2021**, *21*, 4184. [CrossRef]
21. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 49–56. [CrossRef]
22. Zheng, Y.; Izzat, I.H.; Ziaee, S. GFD-SSD: Gated fusion double SSD for multispectral pedestrian detection. *arXiv* **2019**, arXiv:1903.06999. [CrossRef]
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
24. 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, Boston, MA, USA, 7–12 June 2015; pp. 1037–1045. [CrossRef]
25. Dollár, P.; Appel, R.; Belongie, S.; Perona, P. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1532–1545. [CrossRef]
26. Li, C.; Song, D.; Tong, R.; Tang, M. Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv* **2018**, arXiv:1808.04818. [CrossRef]
27. Ding, L.; Wang, Y.; Laganiere, R.; Huang, D.; Fu, S. Convolutional neural networks for multispectral pedestrian detection. *Signal Process. Image Commun.* **2020**, *82*, 115764. [CrossRef]
28. Deng, Q.; Tian, W.; Huang, Y.; Xiong, L.; Bi, X. Pedestrian Detection by Fusion of RGB and Infrared Images in Low-Light Environment. In Proceedings of the 2021 IEEE 24th International Conference on Information Fusion (FUSION), Sun City, South Africa, 1–4 November 2021; IEEE: New York, NY, USA, 2021; pp. 1–8. [CrossRef]
29. Yang, X.; Qian, Y.; Zhu, H.; Wang, C.; Yang, M. BAANet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; IEEE: New York, NY, USA, 2022; pp. 2920–2926. [CrossRef]

30. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016, Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October* 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]
31. Zhang, L.; Liu, Z.; Zhang, S.; Yang, X.; Qiao, H.; Huang, K.; Hussain, A. Cross-modality interactive attention network for multispectral pedestrian detection. *Inf. Fusion* **2019**, *50*, 20–29. [CrossRef]
32. Kim, J.; Park, I.; Kim, S. A Fusion Framework for Multi-Spectral Pedestrian Detection using EfficientDet. In Proceedings of the 2021 21st International Conference on Control, Automation and Systems (ICCAS), Jeju, Republic of Korea, 12–15 October 2021; IEEE: New York, NY, USA, 2021; pp. 1111–1113. [CrossRef]
33. Zhang, H.; Fromont, E.; Lefèvre, S.; Avignon, B. Guided attentive feature fusion for multispectral pedestrian detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 72–80. [CrossRef]
34. Kim, J.; Kim, H.; Kim, T.; Kim, N.; Choi, Y. MLPD: Multi-label pedestrian detector in multispectral domain. *IEEE Robot. Autom. Lett.* **2021**, *6*, 7846–7853. [CrossRef]
35. Kim, J.U.; Park, S.; Ro, Y.M. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1510–1523. [CrossRef]
36. Zhang, L.; Liu, Z.; Zhu, X.; Song, Z.; Yang, X.; Lei, Z.; Qiao, H. Weakly Aligned Feature Fusion for Multimodal Object Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–15. [CrossRef] [PubMed]
37. Shojaiee, F.; Baleghi, Y. Pedestrian head direction estimation using weight generation function for fusion of visible and thermal feature vectors. *Optik* **2022**, *254*, 168688. [CrossRef]
38. Roszyk, K.; Nowicki, M.R.; Skrzypczyński, P. Adopting the YOLOv4 architecture for low-latency multispectral pedestrian detection in autonomous driving. *Sensors* **2022**, *22*, 1082. [CrossRef] [PubMed]
39. Yang, Y.; Xu, K.; Wang, K. Cascaded information enhancement and cross-modal attention feature fusion for multispectral pedestrian detection. *Front. Phys.* **2023**, *11*, 1121311. [CrossRef]
40. Guan, D.; Luo, X.; Cao, Y.; Yang, J.; Cao, Y.; Vosselman, G.; Yang, M.Y. Unsupervised Domain Adaptation for Multispectral Pedestrian Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 434–443. [CrossRef]
41. Zhou, K.; Chen, L.; Cao, X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Computer Vision – ECCV 2020, Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Part XVIII; Springer: Cham, Switzerland, 2020; pp. 787–803. [CrossRef]
42. Zhuang, Y.; Pu, Z.; Hu, J.; Wang, Y. Illumination and temperature-aware multispectral networks for edge-computing-enabled pedestrian detection. *IEEE Trans. Netw. Sci. Eng.* **2021**, *9*, 1282–1295. [CrossRef]
43. Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; Ma, J. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion* **2022**, *83–84*, 79–92. [CrossRef]
44. González, A.; Fang, Z.; Socarras, Y.; Serrat, J.; Vázquez, D.; Xu, J.; López, A.M. Pedestrian Detection at Day/Night Time with Visible and FIR Cameras: A Comparison. *Sensors* **2016**, *16*, 820. [CrossRef]
45. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 3489–3497. [CrossRef]
46. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761. [CrossRef]
47. Qingyun, F.; Dapeng, H.; Zhaokui, W. Cross-Modality Fusion Transformer for Multispectral Object Detection. *arXiv* **2021**. [CrossRef]