electronics

MDPI

Article

Multispectral Pedestrian Detection Based on Prior-Saliency Attention and Image Fusion

Jiaren Guo [1,2], Zihao Huang [1,2] and Yanyun Tao [1,2,3,*]

[1] School of Rail Transportation, Soochow University, Suzhou 215005, China; 20214246019@stu.suda.edu.cn (J.G.); 20215246008@stu.suda.edu.cn (Z.H.)
[2] Suzhou Transportation Big Data Innovation & Application Laboratory, Suzhou 215005, China
[3] Key Laboratory of Information Processing and Intelligent Control, Shanghai Jiaotong University, Shanghai 350121, China
[*] Correspondence: taoyanyun@suda.edu.cn

Abstract: Detecting pedestrians in varying illumination conditions poses a significant challenge, necessitating the development of innovative solutions. In response to this, we introduce Prior-AttentionNet, a pedestrian detection model featuring a Prior-Attention mechanism. This model leverages the stark contrast between thermal objects and their backgrounds in far-infrared (FIR) images by employing saliency attention derived from FIR images via UNet. However, extracting salient regions of diverse scales from FIR images poses a challenge for saliency attention. To address this, we integrate Simple Linear Iterative Clustering (SLIC) superpixel segmentation, embedding the segmentation feature map as prior knowledge into UNet's decoding stage for comprehensive end-to-end training and detection. This integration enhances the extraction of focused attention regions, with the synergy of segmentation prior and saliency attention forming the core of Prior-AttentionNet. Moreover, to enrich pedestrian details and contour visibility in low-light conditions, we implement multispectral image fusion. Experimental evaluations were conducted on the KAIST and OTCBVS datasets. Applying Prior-Attention mode to FIR-RGB images significantly improves the delineation and focus on multi-scale pedestrians. Prior-AttentionNet's general detector demonstrates the capability of detecting pedestrians with minimal computational resources. The ablation studies indicate that the FIR-RGB+ Prior-Attention mode markedly enhances detection robustness over other modes. When compared to conventional multispectral pedestrian detection models, Prior-AttentionNet consistently surpasses them by achieving higher mean average precision and lower miss rates in diverse scenarios, during both day and night.

Keywords: multispectral; pedestrian detection; feature fusion; computer vision; prior-attention

check for updates

Citation: Guo, J.; Huang, Z.; Tao, Y. Multispectral Pedestrian Detection Based on Prior-Saliency Attention and Image Fusion. *Electronics* **2024**, *13*, 1770. https://doi.org/10.3390/electronics13091770

Academic Editor: George A. Tsihrintzis

Received: 11 April 2024
Revised: 26 April 2024
Accepted: 30 April 2024
Published: 3 May 2024

1. Introduction

Traffic object (pedestrians, motor vehicles, non-motor vehicles, road signs, etc.) detection under varying illuminance conditions has a wide range of applications in road traffic, such as obstacle detection on roads, traffic flow monitoring at intersections, and unmanned driving during the day and at night. In these scenarios, high-precision and -reliability object detection methods are required [1]. During daylight hours, contemporary object detection algorithms such as Yolo [2] and the fast R-CNN series [3] demonstrate commendable performance when operating on data from visible cameras. However, as illuminance conditions deteriorate, the information gleaned from visible images weakens, often becoming indistinguishable amidst background noise. In contrast, infrared (IR) images, generated by capturing the heat radiating from objects, exhibit higher resilience to low visibility and adverse weather conditions. They encapsulate vital contour information of objects that visible images struggle to preserve under low-illuminance conditions. Nevertheless, IR images exhibit limitations in terms of resolution and detailed information. Conversely, RGB images excel in capturing object details and texture information during daylight hours. Combining

IR and RGB images facilitates mutual feature compensation, effectively enhancing object features in the fused images [4]. Thus, under varying illuminance conditions, multispectral information fusion plays a pivotal role in enhancing object detection.

At present, the fusion based on multi-scale decomposition (MSD) remains the main image fusion method that is widely studied and applied [5]. MSD decomposes a source image into high- and low-frequency information and combines the decomposed frequency information through fusion rules. How to improve MSD is thus the critical issue to the MSD-based fusion framework. For this purpose, researchers have proposed new methods such as Laplacian pyramid [6], low-pass pyramid [7], discrete wavelet pyramid [8], curvelet transform [9], non-subsampling contourlet transform (NSCT) [10], and so on. Under low illuminance, the fusion of original images generates the detailed information and contour of objects integrated, but the background noise in visible images is also compensated by IR images. Thereby, it renders the difference between contour in IR images and background noise insignificant. If the area of the objects cannot be accurately acquired and located, the objects are interfered with by the background noise in the fusion. This main issue affects the accuracy of weak-light object detection [11,12]. Current multispectral detection methods [13–15] are targeted to detect objects all day long by finding the appropriate proportion of fusion, but they are not intently designed to detect objects at night. Therefore, how to focus on the object area during fusion so as to reduce the effect of background noise on detection is the key issue for low-illuminance object detection.

Recently, research on the human visual system (HVS) has become increasingly mature [16]. Visual attention is an important mechanism of HVS that helps extract complex and important visual information by quickly selecting the most significant area [17]. Usually, the heat of important objects (pedestrians and moving vehicles) is higher than that of the surrounding environment. The high-temperature area in an FIR image is typically the region of important objects. Under low illuminance, the objects in the FIR image must be clearer than those in an RGB image because of the heat difference. With this feature of FIR images, the attention mechanism on the thermal objects of FIR images is adaptive to varying-illuminance object detection. However, in FIR images, the scale of vehicles must be larger than that of pedestrians in the same position. The size of near-end pedestrians is also larger than that of far-end ones. The multiple scales of objects lead to difficulty in attention region acquisition. Therefore, a multi-scale region attention mechanism is required.

This study introduces Prior-AttentionNet, an innovative object detection framework designed to address varying-illuminance challenges. The model encapsulates a mechanism of saliency attention and segmentation prior, meticulously focusing on multi-scale objects to elevate detection efficiency. The core of this approach lies in the Prior-Attention module, a novel attention mechanism that amalgamates segmentation prior knowledge with saliency attention extraction, thereby streamlining the identification of critical attention regions and enriching feature diversity across scales. Additionally, we pioneer an FIR-RGB fusion technique grounded in illuminance levels, which, by merging infrared and visible spectrum information, markedly bolsters object details and contours, thus ensuring high detection accuracy under diverse lighting scenarios. Our comprehensive evaluation of Prior-AttentionNet across KAIST, OTCBVS, and CVC-14 datasets, and its subsequent comparison with leading multispectral detection models, underscores the model's efficacy and potential in enhancing object detection performance.

2. Related Work

Under varying illuminance, multispectral object detection attracts extensive attention due to the great advantages of multispectral data in all-day visual displays. Hwang collected and marked the first dataset called KAIST for pedestrian detection and then proposed the ternary histogram of oriented gradient (T-HOG) operator with a multispectral auto-correlation function (ACF) to expand the gradient information of the IR image channel, process FIR-RGB images in parallel, and detect pedestrians through the AdaBoost classifier. Afterwards, ACF+T+THOG was comprehensively used as the baseline in this field, which encouraged researchers to improve the relative technologies in this field. Alzate et al. mainly selected a group of the most commonly used and highest-scoring features from the current pedestrian detection works. They then accessed the deformable part model (DPM) and random forest (RF) of local experts with HOG and local binary pattern (LBP) on benchmarks [18].

Some researchers adopted segmentation, enhancement, and edge computation to improve object detection. A self-balanced sensitivity segmentationer (SuBSENSE) makes camouflaged foreground objects easier to detect through a pixel-level segmentation method [19]. Another foreground segmentation method using a neural network (FgSegNet) is also applied in moving object detection [20]. Illuminance-invariant structural complexity (IISC) is employed to implement background subtraction in outdoor scenes for moving object detection [21]. Kim et al. [22] reconstructed moving objects by computing the edges on the result of frame differencing under varying illuminances. Gautam et al. [23] proposed a style-transferred enhanced image module (STEIM) with an EfficientDet module (EDM) to improve the resolution of IR input frames.

In terms of fusion methods, among the varied MSD methods for image fusion, NSCT has a better frequency selectivity and regularity than other methods. Thus, NSCT-based fusion methods are widely used in fusion applications. For instance, Chen et al. [24] proposed a pulse convolution neural network (PCNN) combined with non-subsampled shearlet transform (NSST) for multi-source image fusion. They employed a CNN to extract the features of visible and IR images, which calculated the fusion weights. Su et al. [25] proposed a compression fusion of FIR and visible images based on the combination of the robust principal component analysis (RPCA) and NSCT. It aims to make up for the loss of detailed texture information in fusion images. However, the principal component analysis (PCA) that reduces the dimension information of original images unintentionally increases the loss of useful information. Wagner et al. [26] developed two fusion architectures (LateFusion and EarlyFusion) with a deep CNN, and their multispectral pedestrian detector achieved more true detection.

Compared with the algorithm of single-mode pedestrian detection, the multispectral pedestrian detection algorithm based on the R-CNN series and a deep neural network (DNN) has stronger robustness and accuracy. Ding et al. [27] employed a network in network (NIN) for FIR-RGB fusion and a selective kernel network to adaptively adjust the receptive field size in detection. Liu et al. [28] developed a Halfway model of intermediate fusion that achieved a balance between visual details and semantic information and a missed detection rate (MDR) of 37% on KAIST. He et al. [29] proposed the regional proposal network (RPN) from fast R-CNN for single-mode pedestrian detection and used a decision tree for classification. Based on the concept, Konig et al. combined visible and IR images into the RPN [30] and reduced the MDR to 29.8% on KAIST. A multispectral simultaneous detection and segmentation R-CNN called MSDS-RCNN is explored, which adds a pixel-level segmentation module to the detection method that splits the background and objects [31].

The mode difference remains a difficult issue in fusion. Guan et al. [14] and Li et al. [32] introduced the light perception network to weight day and night networks by using predicted illuminance values. Xu et al. [33] proposed a cross-modality transfer CNN (CMT-CNN) to learn the non-linear mapping from FIR-RGB images so as to overcome adverse lighting. Zhang et al. [34] improved the effectiveness of multi-mode feature fusion by encoding the correlation between the two modes through a cross-modal interactive attention network (CIAN). Then, they proposed a novel aligned region CNN (AR-CNN) to capture position offset and solve the problem of position offset between multi-modal images through a regional feature alignment module [35]. Zhou et al. [13] proposed the mode balance network (MBNet), which adaptively selects two mode features through the differential mode perception fusion module. Jiang et al. proposed a cross-modality fusion framework based on Yolov5. The backbone network is used for multi-scale feature fusion, which enables the network to detect objects with different scales, thus improving detection accuracy. Park et al. [36] verified the effectiveness of the proposed framework with extensive experiments, and it achieved state-of-the-art pedestrian detection performance on thermal infrared images.

3. Materials and Methods

3.1. Network Structure

In this study, we introduce a deep neural detection network named Prior-AttentionNet, which is based on a Prior-Attention module and image fusion for pedestrian detection under varying illumination conditions. The architecture of Prior-AttentionNet is depicted in Figure 1 and primarily consists of two parts: the image fusion algorithm framework and the detector framework. Within the image fusion algorithm, this paper focuses on adopting superpixel segmentation to extract prior features from infrared images. These features are then integrated into the UNet [37] saliency attention extraction module to obtain feature maps, which are combined with the FIR-RGB fused image.

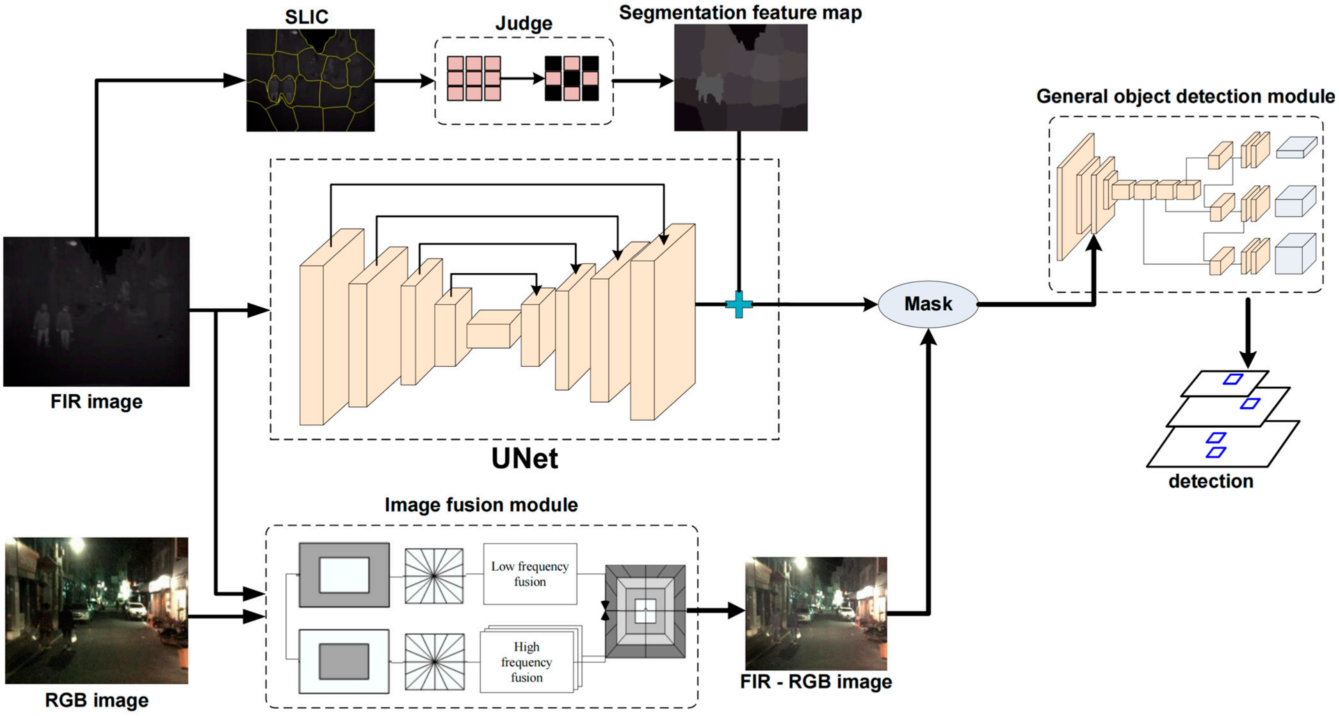

Figure 1. Overall framework of the proposed network: Prior-AttentionNet.

The Prior-Attention module and the image fusion module are integrated using a masking approach, where the feature map obtained from the Prior-Attention module is masked onto the FIR-RGB multispectral fusion image. The method of merging the feature map with the FIR-RGB fusion image is illustrated in Equation (1).

$$I_{out} = M_{att} + I_{FIR\text{-}RGB} \tag{1}$$

Here, M_{att} represents the feature map output by the Prior-Attention module, $I_{FIR\text{-}RGB}$ represents the fused image after image fusion, and I_{out} represents the FIR-RGB image after being masked by the feature map.

Masking can filter out background noise in the FIR-RGB image, reducing interference in the detector's feature-learning process while simultaneously enhancing the edge contour features of pedestrians, making pedestrians in the image easier to detect. This complies with the requirements for scientific research documentation.

For detection, the network employs YOLOv5, into which the combined images are fed to realize end-to-end training and detection.

3.2. Prior-Attention Module

In this section, we introduce a Prior-Based Saliency Attention module, consisting of feature extraction using SLIC (Simple Linear Iterative Clustering) superpixel segmentation and saliency attention derived from UNet.

3.2.1. Segmentation Feature Extraction Module

The segmentation feature extraction module is based on the superpixel segmentation. The Simple Linear Iterative Clustering can be divided into the following steps: First, initialize cluster centers uniformly on the image. Then, assign pixels to the nearest cluster centers based on a comprehensive measure of color similarity and spatial proximity. Afterward, recalibrate each cluster center based on the mean of the pixels assigned to them. This assignment and update process iterates until the cluster centers stabilize, achieving convergence.

The specific implementation steps of this method are as follows:

Step 1: Choose vector X as the feature vector of the pixel.

Step 2: Initialize cluster centers: Distribute initial cluster centers evenly based on the set number of superpixels (K). Assuming the total number of image pixels is N and the distance between neighboring centers is S, the calculation formula is Equation (2).

$$S = sqrt(N/K) \tag{2}$$

Step 3: Move the initial cluster centers to the position of the minimum gradient within a 3 × 3 neighborhood around each initial cluster center.

Step 4: Assign labels to each pixel based on the distance D from the center pixel within a 2S × 2S neighborhood around each seed point. The distance measurement for superpixel segmentation quality includes color distance and spatial distance, as shown in Equations (3)–(5).

$$D = \sqrt{\left(\frac{d_c}{m}\right)^2 + \left(\frac{d_s}{\mathrm{s}}\right)^2} \tag{3}$$

$$d_c = \sqrt{\left(R_j - R_i\right)^2 + \left(G_j - G_i\right)^2 + \left(B_j - B_i\right)^2 + \left(D_j - D_i + \left(H_j - H_i\right)^2\right)} \tag{4}$$

$$d_s = \sqrt{\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2} \tag{5}$$

Here, d_c is the color distance, d_s represents the spatial distance, and m and s denote the maximum allowable values for color distance and spatial distance, respectively.

Step 5: Update cluster centers, repeating the above steps iteratively until convergence of the error.

The number K of superpixels in segmentation varies for different images. For accurate segmentation, we adopt a pre-trained detection (e.g., YOLOv5) model to estimate the appropriate number of superpixels. The ratio of the FIR image size to the largest detection frame is used as the number of superpixels.

Through the segmentation, the image can be split into pieces of superpixels. Each superpixel comprises pixels with similar thermal features, and the gray-scale value of each superpixel corresponds to the level of thermal radiation. The superpixels with a gray level higher than the setting threshold are the attention area, which retains the objects. We expect that a small- or middle-scale object can be entirely involved in one single superpixel block. In some cases, small objects may be divided into multiple disconnected superpixels. To solve this issue, we use an iterative process to update the cluster centers until the error converges. This enhances the connectivity among superpixels. In Figure 2a,b, we can see that an FIR image decomposes to superpixels. Two pedestrians are included in one superpixel.

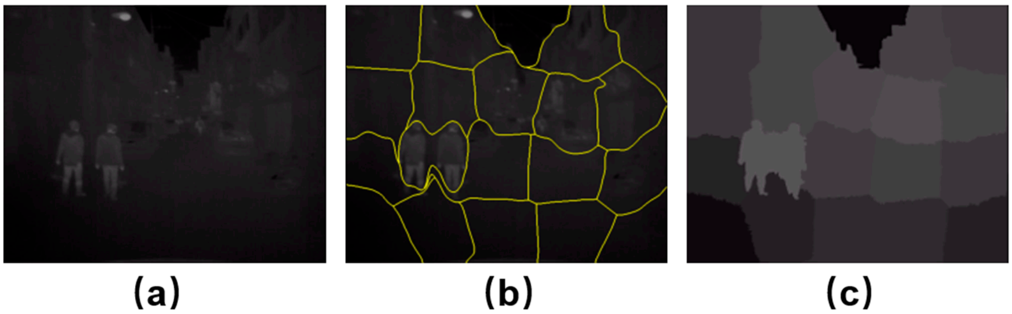

Figure 2. Segmentation-based extraction on FIR image: (**a**) FIR image; (**b**) segmentation superpixels; (**c**) segmentation feature map.

Segmentation enhances the superpixels of the foreground area prior to filtering while weakening the superpixels of the background area. In Figure 2, SLIC segments the FIR image into a set of superpixels, forming a segmented image, which is ultimately transformed into a segmentation feature image based on the average gray-scale values of the superpixels.

3.2.2. Prior-Based Saliency Attention

The segmentation prior prefers to concentrate on the connected regions, which contain the objects. For middle- or large-scale objects, the richer the details of the object, the more likely it is that the object is divided into different connected regions. When the integrity of the object is damaged, the accuracy of object detection declines.

For the larger-scale objects with rich details, we need another attention mode. UNet is a well-known deep learning network, originally designed for the segmentation of medical images. Its characteristic feature is the adoption of a symmetrical Encoder–Decoder structure, and the introduction of Skip connections between the encoder and decoder, to preserve detailed information lost during the down-sampling process. This structure enables UNet to excel in image segmentation tasks that require precise localization.

The UNet network is used for saliency attention extraction. The input is an FIR image, and the trained UNet saliency extraction network can output high-quality feature maps. As illustrated, the feature map distinctly extracts two pedestrians, eliminating a large amount of background noise. The UNet network is capable of precisely capturing the subtle temperature differences between pedestrians and the background in infrared images, and through deep learning, it automatically learns to differentiate features of pedestrians and the background from the data. UNet's strong adaptability, scalability, and flexibility, with proper training, enable it to effectively locate the salient regions of pedestrians in infrared images, providing powerful support for pedestrian detection and related applications. Joint training with subsequent detection models can help the network focus more quickly on learning and detecting pedestrian features, improving the performance of the detection network.

From Figure 3, fusing segmentation prior with saliency attention at the decoding layer significantly accentuates pedestrian features while mitigating background noise. This adaptation allows subsequent detection models to more efficiently focus on and extract pedestrian features, thereby enhancing both learning efficiency and detection accuracy. The incorporation of SLIC superpixel segmentation prior features enables the UNet saliency attention extraction network to converge more swiftly under the guidance of prior features.

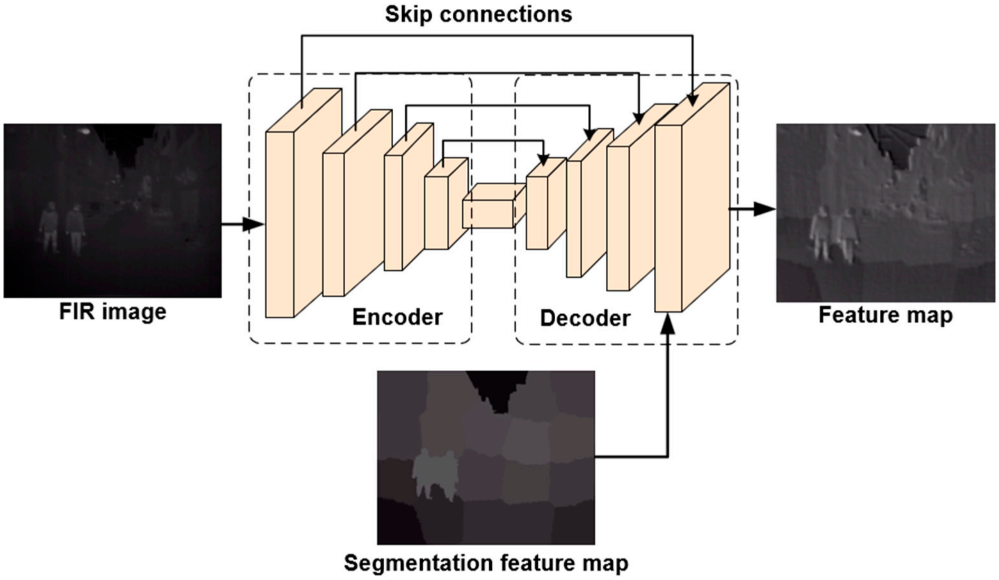

Figure 3. Saliency attention extraction using UNet.

3.3. Image Fusion Module

Fusions at the pixel, feature, and decision levels are the main methods of image fusion. The fusions at the feature and decision levels often lose the details of objects. In particular, before different-sized images are fed to the network, down-sampling usually causes information loss. When the image resolution is low, the loss is not conducive to object detection. The resolution of FIR images is typically low, and RGB images under low illuminance have a small amount of information. Hence, fusion either at the feature level or decision level is unsuitable. Fusion at the pixel level has high efficiency in data processing, rich image details, and strong robustness. Thus, pixel-level image fusion is adopted.

In pixel-level fusion, the registration method matches the feature points and utilizes the affine matrix of image transformation. NSCT is adopted to implement the fusion in multi-scale transformation. In NSCT, the non-subsampling (NSP) tower performs multi-scale decomposition to obtain low- and high-frequency information. After this, the non-down-sampling filter bank (NSDFB) constructs multi-direction and multi-scale representations of the image. Finally, the fusion of high- and low-frequency information is based on rules and the inverse NSCT.

The low-frequency part indicates the smooth area (e.g., background) of the image. It contains the spectral information and most of the energy in the image. The fusion in the low-frequency part is the weighted sum of the low frequencies of both the FIR and RGB images. Generally, the edge information of FIR images is highly significant under varying illuminances, especially at night. Hence, the intensity distribution of FIR images dominates the weight of low-frequency fusion. This guarantees that high-contrast features are preserved in FIR-RGB images at night. The fusion in the low-frequency part is calculated using Equation (6).

$$LF_N = F \cdot LA_N + (1 - F) \cdot LB_N \tag{6}$$

Here, L_{AN} and L_{BN} indicate the low-frequency information of FIR and RGB images, respectively. LF_N indicates the low-frequency information after fusion. F is the weight for fusion.

$$P = \frac{R}{\max_{x \in \Omega}\{R(x)\}} \tag{7}$$

$$F = \frac{\arctan(\lambda P)}{\arctan(\lambda)} \tag{8}$$

In Equations (7) and (8), R represents the pixel values of the original image, max refers to the maximum pixel intensity value within the pixel region Ω, $P \in [0, 1]$ maps the pixel intensity distribution in the infrared image, $F \in [0, 1]$ represents the distribution of infrared features, and λ is the tuning parameter of the function. Through this method, greater weight is assigned to the infrared information during the fusion process, thereby helping to avoid the problem of reduced target contrast in the fused image due to the loss of high-frequency information.

When λ is relatively large, according to the formula for F, the corresponding non-linear transformation becomes more pronounced. Since infrared images contain more information under low-illumination conditions compared to RGB images, F must be greater than 0.5; under high illumination, F is less than 0.5. Through experimental testing, we determined that when $\lambda = 30$, it satisfies the requirements for the magnitude of F under different illumination conditions.

The high-frequency information usually contains the edges and contours of objects. In the image decomposition, a very high frequency typically represents sharp changes in brightness and edges with large contrast changes in the image, such as borders, bright lines, and regional outer lines. The larger the absolute value of the high frequency, the richer the details the objects have. For pedestrians, the absolute value is large. The maximum absolute value rule for the fusion of the high-frequency band can be calculated using Equation (9).

$$LH_l = \begin{cases} LA_l, & |LA_l > LB_l| \\ LB_l, & otherwise, \end{cases} \tag{9}$$

Here, LA_l and LB_l are the high-frequency parts decomposed in the FIR and RGB images, respectively. LH_l indicates the high-frequency information after fusion. The contrast of the fused high-frequency image is close to that of the source sub-image.

3.4. Complexity and Running Time

The prior-saliency attention is the main part of Prior-AttentionNet. The complexity of Prior-AttentionNet depends on the segmentation and saliency attention.

The time complexity for the segmentation calculation mainly comes from SLIC superpixel segmentation. SLIC is based on a k-means algorithm with a time complexity of $O(N \times c)$, where N is the number of pixels in the image and c is the number of clusters. The time complexity of SLIC is proportional to the number of pixels and clusters. In this study, SLIC is dedicated to the superpixel clustering problem, and its complexity is linear with the number of pixels and independent of c. Therefore, the time complexity of SLIC is $O(N)$.

The saliency attention is calculated by UNet, and the time complexity of UNet depends on the depth and width of the network as well as the size of the input image. Specifically, assuming that UNet has L layers, each layer has C convolutional kernels with a size of K^2, and the input image size is $W \times H = N$. Therefore, the time complexity of UNet is $O((LCK^2) \times W \times H) = O(NLCK^2)$. Here, LCK^2 represents the time complexity of the convolutional layers, and N is the number of pixels in the input image. The time complexity of UNet increases with the depth and width of the network.

The calculations for the segmentation and saliency attention are parallel, so they do not affect each other. The time complexity of UNet is larger than that of SLIC, i.e., $O(NLCK^2) > O(N)$. In terms of time complexity, the calculation of saliency attention is dominant in Prior-AttentionNet. The time complexity of dual attention is $O(NLCK^2)$. The object detection algorithm used in this paper primarily runs on GPU for most of its computational power, so the computation time is mainly governed by the GPU's processing time. The runtime depends on the size of the dataset, and on our dataset, it operates at a speed of 52 frames per second (fps) on an RTX 4080 graphics card.

4. Experiment

4.1. Datasets

KAIST [38], OTCBVS [39], and CVC-14 are benchmarks adopted for model validation. Table 1 lists the number of images derived from different datasets for training and validation.

Table 1. Number of images for training and validation from day and night.

Datasets	Day		Night	
	Training	**Validation**	**Training**	**Validation**
KAIST	2254	225	3500	350
OTCBVS	1054	105	400	106
CVC-14	706	71	1386	139

In KAIST, we adopted the pre-process that removed the images without pedestrian instances and those with inconsistent numbers of pedestrian instances. A total of 6329 RGB and FIR images were derived from the frames for detection during the day and at night. To improve the training, we re-labeled some images that contained wrong labels.

In OTCBVS, we used the OSU RGB-thermal database, which supports the detection of objects in thermal FIR and RGB imagery. We extracted 1454 and 211 frames of RGB and FIR videos for training and testing, respectively. We re-labeled some images without labels during the pre-processing stage. In this dataset, the size of the moving objects is very small.

CVC-14 contains FIR and RGB images of high quality. We selected 777 and 1535 FIR-RGB images of day and night scenes for training and validation, respectively.

4.2. Training Setting

In this study, we conducted experiments using NVIDIA RTX4080 GPU with PyTorch. The parameter setting was achieved by conducting pretests on the datasets for ten runs. Then, we took the value of parameters under the highest accuracy of detection as the setting in the uniform experiment.

To observe the effect of the detector in Prior-AttentionNet, we adopted YOLOv5 to implement pedestrian detection. The parameter settings for the detector used in this paper are shown in Table 2.

Table 2. Parameter settings of the detector.

Detector	Backbone	Training Epoch	Batch Size	Learning Rate	Weight Attenuation	IoU
YOLOv5	CSPDarknet53	100	4	10^{-3}	10^{-4}	0.5

4.3. Evaluation Metric

In evaluating object detection models, including pedestrian detection, we focus on three critical metrics: the miss rate (MR), mean average precision (mAP), and the F1 Score. The MR, sampled for a false positive rate per image (FPPI) within the range [0.01, 1], as proposed by Dollar et al. [40], is the most popular metric for pedestrian detection tasks, emphasizing high-accuracy areas, making it highly relevant for commercial applications. The mAP, on the other hand, calculates the average precision at varying recall levels, providing a comprehensive single-figure measure of a model's quality across different confidence thresholds, widely accepted for its effectiveness in assessing overall detection precision. Lastly, the F1 Score harmonizes precision and recall into a single metric, offering a balanced view of the model's detection accuracy by equally weighting the importance of precision (the quality of the detected objects) and recall (the completeness of the detection). Together, these metrics offer a multifaceted evaluation of detection models, ensuring both the effectiveness and reliability of the detection tasks.

4.4. Comparison with State-of-the-Art Multispectral Pedestrian Detection Methods

4.4.1. KAIST Dataset

To validate the effectiveness of our proposed method, we compared it with the current state-of-the-art multispectral pedestrian detectors, including RF HOG+LBP [18], LateFusion CNN, EarlyFusion CNN [26], CMT-CNN [33], Halfway fusion [28], MSDS-RCNN [31], CIAN [34], AR-CNN [35], MBNet [13], and ICAFusion [41]. Table 3 shows the detection results of our method and the state-of-the-art detectors on the KAIST dataset. Across all scenes, including day and night, we achieved a miss rate of 6.13% at night and 7.04% during the day, which is a reduction of 0.69% at night and 0.81% during the day compared to the previous best method, ICAFusion [41]. From these results, our method demonstrates effective feature fusion of the two spectra, maintaining good detection accuracy in both daytime and nighttime scenarios.

Table 3. MR of comparison models and Prior-AttentionNet on KAIST dataset.

Method	Low Illuminance (Night)	High Illuminance (Day)
RF HOG+LBP [18]	29.4	28.7
LateFusion CNN [26]	37.0	46.2
EarlyFusion CNN [26]	51.8	50.9
CMT-CNN [33]	54.8	47.3
Halfway fusion [28]	26.59	24.88
MSDS-RCNN [31]	10.60	13.73
CIAN [34]	11.13	14.77
AR-CNN [35]	8.38	9.94
MBNet [13]	7.86	8.28
ICAFusion [41]	6.82	7.85
Ours	6.13	7.04

4.4.2. OTCBVS Dataset

To verify the effectiveness of our proposed method, we compared it with the current state-of-the-art multispectral pedestrian detectors, including SuBSENSE [19], IISC [21], Kim's method [22], FgSegNet [20], SRCNN+EDM, EDSR+EDM, and STEIM+EDM [23]. Table 4 presents the detection results of our method and these advanced detectors on the OTCBVS dataset. Across all scenes, encompassing both day and night, our method achieved an F1 Score of 0.678 and a mean average precision (mAP) of 94.54%, which is a 7.03% improvement over the previous best method, STEIM+EDM [23]. Given that pedestrian targets in OTCBVS are primarily small-scale, the aforementioned results indicate that our method also exhibits commendable performance in handling small-scale targets.

Table 4. mAP and F1 Score of models tested on OTCBVS dataset.

Method	F1 Score			mAP		
SuBSENSE [19]		0.638			-	
IISC [21]		0.674			-	
Kim's method [22]		0.569			-	
FgSegNet [20]		0.077			-	
SRCNN+EDM [23]		-			65.45	
EDSR+EDM [23]		-			76.01	
STEIM+EDM [23]		-			87.51	
Ours	all	day	night	all	day	night
	0.678	0.699	0.639	94.54	97.65	92.62

4.4.3. CVC-14 Dataset

Table 5 shows the performance of models tested on CVC-14; we compared our model with RF HOG+LBP [18], Halfway fusion [28], AR-CNN [35], and MBNet [13]. Across all scenes, including day and night, we achieved a miss rate of 17.6% at night and 11.0% during the day. Our model achieves the best detection performance during the daytime. While there is not a significant improvement in detection performance compared to MBNet during nighttime, our model significantly outperforms MBNet and other methods during daytime.

Table 5. MR of comparison models and Prior-AttentionNet on CVC-14 dataset.

Method	Low Illuminance (Night)	High Illuminance (Day)
RF HOG+LBP [18]	26.6 (RGB) 16.7 (FIR)	81.2 (RGB) 24.8 (FIR)
Halfway fusion [28]	34.4	38.1
AR-CNN [35]	18.1	24.7
MBNet [13]	13.5	24.7
Ours	17.6	11.0

4.5. Ablation Study

In this ablation study, we delve into the intricate interplay between two pivotal modules: the FIR-RGB Fusion Module and the Prior-Attention module. These modules represent critical components of our pedestrian detection framework, each contributing distinct capabilities to enhance the overall performance. Our primary objective is to comprehensively assess and quantify the impact of these modules on pedestrian detection accuracy, particularly when confronted with varying illuminance conditions. We conducted ablation experiments on the KAIST and OTCBVS datasets.

4.5.1. Effect of FIR-RGB Fusion Module

From Tables 6 and 7, it can be observed that the FIR-RGB mode shows improvements compared to the single FIR and RGB modes. In low-light conditions, FIR images tend to perform better, while in well-lit scenarios, the detection performance of RGB datasets is often superior. By fusing FIR and RGB, we can effectively leverage the advantages of different spectra in varying lighting environments, resulting in a dataset with richer features that is more conducive to further image processing and object detection. The use of the FIR-RGB mode has led to further improvements in the MR compared to detection using a single spectrum.

Table 6. MR and mAP of FIR mode, RGB mode, and FIR-RGB mode tested on KAIST.

Mode	Night (Low Illuminance)		Day (High Illuminance)	
	MR	mAP	MR	mAP
FIR	9.76	75.6	17.61	91.5
RGB	10.87	78.5	12.55	93.2
FIR-RGB	7.89	87.0	13.52	93.5

Table 7. MR and mAP of FIR mode, RGB mode, and FIR-RGB mode tested on OTCBVS.

Mode	Night (Low Illuminance)		Day (High Illuminance)	
	MR	mAP	MR	mAP
FIR	17.98	84.4	19.30	94.2
RGB	20.97	86.8	17.86	95.2
FIR-RGB	19.53	89.4	17.60	97.8

Figure 4 illustrates the detection results on three datasets. The FIR-RGB mode exhibits higher detection rates and confidence compared to the FIR and RGB modes. In the KAIST dataset, all three modes detect pedestrians. In the OTCBVS dataset, the FIR-RGB mode achieves the highest confidence score. The FIR-RGB mode optimizes detection rates over the FIR and RGB modes. The fused data can leverage the advantages of individual modalities, ultimately leading to more consistent detection results. Although multispectral image fusion may potentially reduce visual perception quality, it provides additional feature information for machine learning and object detection, thereby enhancing detection performance.

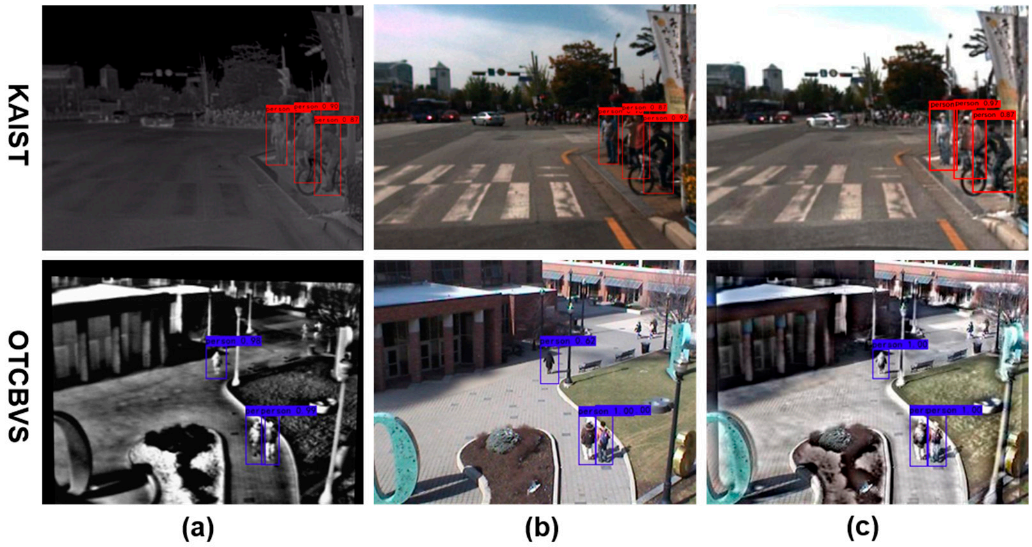

Figure 4. Detection on KAIST and OTCBVS with YOLOv5: (**a**) FIR mode; (**b**) RGB mode; (**c**) FIR-RGB mode.

4.5.2. Effect of Prior-Attention Module

Tables 8 and 9 present the comparative results of the ablation experiment of the Prior-Attention module. For low illuminance, FIR-RGB+ Prior-Attention boosts the average mAP by 7.79%, and for high illuminance, it enhances it by 3.61%. On KAIST images, FIR-RGB+ Prior-Attention excels with an 11.6% mAP improvement under low illuminance and 6.4% under high illuminance. However, on OTCBVS images, it achieves a substantial 25.55% reduction in mAP.

Table 8. Ablation study on KAIST.

Mode	Night (Low Illuminance)		Day (High Illuminance)	
	MR	mAP	MR	mAP
FIR-RGB	7.89	87.00	13.52	93.50
FIR-RGB + Saliency Attention	6.62	92.10	8.73	95.53
FIR-RGB + Segmentation	6.26	97.31	8.14	97.74
FIR-RGB + Prior-Attention	6.13	97.56	7.04	98.12

Table 9. Ablation study on OTCBVS.

Mode	Night (Low Illuminance)		Day (High Illuminance)	
	MR	mAP	MR	mAP
FIR-RGB	19.53	89.42	17.60	97.80
FIR-RGB + Saliency Attention	16.73	92.61	12.73	95.70
FIR-RGB + Segmentation	8.00	97.86	8.00	98.62
FIR-RGB + Prior-Attention	5.00	98.32	5.00	99.16

In Figure 5, it can be observed that on the KAIST and OTCBVS image datasets, the detection performance significantly improves when adopting the FIR-RGB + Prior-Attention mode compared to using the FIR-RGB + Saliency Attention mode and the FIR-RGB + Segmentation mode. The Prior-Based Saliency Attention mode effectively leverages superpixel segmentation prior features to enhance the pedestrian extraction capability of saliency attention, effectively improving detection outcomes. This method can detect pedestrians of varying scales with high confidence. Compared to other modes, the FIR-RGB + Prior-Attention mode significantly boosts the confidence of detections. For instance, on KAIST images, the confidence level of a cyclist in the image increased from 0.80 to 0.98. By incorporating superpixel segmentation prior into saliency attention, the detector identified two pedestrians that were previously undetectable. On OTCBVS images, by adding superpixel segmentation prior to saliency attention, the detector identified four smaller pedestrians and also detected pedestrians with occlusions. Although the saliency attention mode excels at highlighting pedestrian features, it may miss smaller pedestrians. By incorporating the SLIC superpixel segmentation prior, the model's performance in extracting attention for small objects can be enhanced, improving the detection accuracy for multi-scale pedestrians. Employing a saliency attention extraction network with superpixel segmentation prior can more efficiently detect pedestrians in various environments.

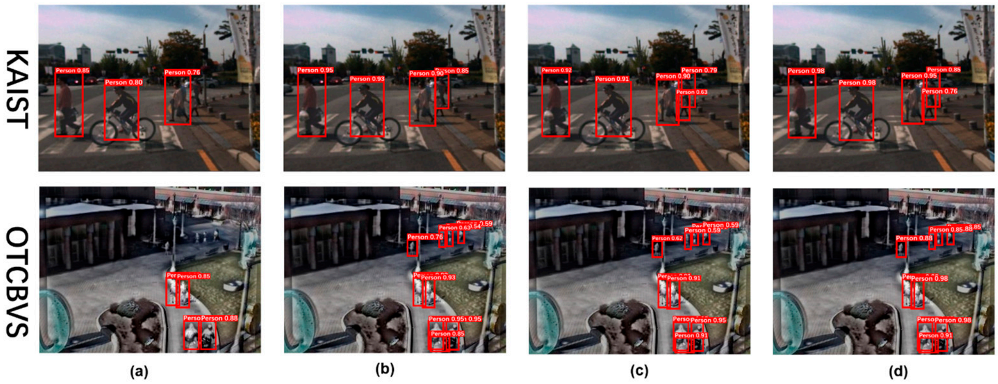

Figure 5. Detection on KAIST and OTCBVS with YOLOv5: (**a**) FIR-RGB mode; (**b**) FIR-RGB + Saliency Attention mode; (**c**) FIR-RGB + Segmentation mode; (**d**) FIR-RGB + Prior-Attention mode.

5. Conclusions

In summary, this study introduces Prior-AttentionNet, an advanced end-to-end detection model designed to address the complex task of multi-scale pedestrian detection under varying illumination conditions. Prior-AttentionNet utilizes Prior-Attention mechanisms and image fusion techniques to enhance detection accuracy. By extracting attention feature maps, it effectively highlights pedestrians in FIR-RGB fused images. Additionally, through the use of a Prior-Attention module for image fusion processing, it successfully reduces background noise from pedestrians that remain in the background even after saliency map masking. However, the method proposed in this paper has certain requirements for the quality of FIR images, and further optimization is needed to enhance its generalization performance. In future work, we will employ an adaptive fusion method for visible and infrared images to further improve the performance of pedestrian detection models under different data quality conditions.

Author Contributions: Conceptualization, J.G. and Y.T.; methodology, J.G. and Y.T.; software, J.G.; validation, J.G.; formal analysis, J.G.; investigation, J.G.; resources, J.G.; data curation, J.G. and Z.H.; writing—original draft preparation, J.G.; writing—review and editing, J.G. and Y.T.; visualization, J.G. and Z.H.; supervision, Y.T.; project administration, Y.T.; funding acquisition, Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the open fund project of the Key lab of digital signal and image Processing of Guangdong province (2022GDDSIPL-03), and the open fund project of Key Laboratory of System Control and Information Processing (Scip202105).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

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

References

1. Chen, G.; Qin, H. Class-discriminative focal loss for extreme imbalanced multiclass object detection towards autonomous driving. *Vis. Comput.* **2022**, *38*, 1051–1063. [CrossRef]
2. Bochkovskiy, A.; Wang, C.Y.; Liao HY, M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
3. Shit, S.; Das, D.K.; Ray, D.N.; Roy, B. An encoder-decoder based CNN architecture using end to end dehaze and detection network for proper image visualization and detection. *Comput. Animat. Virtual Worlds* **2023**, *34*, e2147. [CrossRef]
4. Bavirisetti, D.P.; Xiao, G.; Liu, G. Multi-sensor image fusion based on fourth order partial differential equations. In Proceedings of the 2017 20th International Conference on Information Fusion (Fusion), Xi'an, China, 10–13 July 2017; IEEE: Piscataway, NJ, USA, 2017.
5. Dogra, A.; Goyal, B.; Agrawal, S. From multi-scale decomposition to non-multi-scale decomposition methods: A comprehensive survey of image fusion techniques and its applications. *IEEE Access* **2017**, *5*, 16040–16067. [CrossRef]
6. Burt, P.J.; Adelson, E.H. The laplacian pyramid as a compact image code. *IEEE Trans. Commun.* **1983**, *31*, 532–540. [CrossRef]
7. Toet, A. Image fusion by a ratio of low-pass pyramid. *Pattern Recognit. Lett.* **1989**, *9*, 245–253. [CrossRef]
8. Li, H.; Manjunath, B.S.; Mitra, S.K. Multisensor image fusion using the wavelet transform. *Graph. Models Image Process* **1995**, *57*, 235–245. [CrossRef]
9. Nencini, F.; Garzelli, A.; Baronti, S.; Alparone, L. Remote sensing image fusion using the curvelet transform. *Inf. Fusion* **2007**, *8*, 143–156. [CrossRef]
10. Zhang, Q.; Guo, B.L. Multifocus image fusion using the non sub sampled contourlet transform. *Signal Process.* **2009**, *89*, 1334–1346. [CrossRef]
11. Yin, W.; He, K.; Xu, D.; Yue, Y.; Luo, Y. Adaptive low light visual enhancement and high-significant target detection for infrared and visible image fusion. *Vis. Comput.* **2023**, *39*, 6723–6742. [CrossRef]
12. Huang, Z.; Hui, B.; Sun, S.; Ma, Y. Infrared image super-resolution method based on dual-branch deep neural network. *Vis. Comput.* **2023**, *40*, 1673–1684.
13. Zhou, K.; Chen, L.; Cao, X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 787–803.
14. Guan, D.; Cao, Y.; Yang, J.; Cao, Y.; Yang, M.Y. Fusion of multispectral data through illuminance-aware deep neural networks for pedestrian detection. *Inf. Fusion* **2019**, *50*, 148–157. [CrossRef]
15. 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.
16. Ran, Y.; Leykin, A.; Hammoud, R. Thermal-visible video fusion for moving target tracking and pedestrian motion analysis and classification. In *Augmented Vision Perception in Infrared: Algorithms and Applied Systems*; Springer: London, UK, 2007; pp. 349–369.
17. Achanta, R.; Estrada, F.; Wils, P.; Süsstrunk, S. Salient region detection and segmentation. In Proceedings of the Computer Vision Systems: 6th International Conference, ICVS 2008, Santorini, Greece, 12–15 May 2008; Proceedings 6; Springer: Berlin/Heidelberg, Germany, 2008.
18. 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] [PubMed]
19. St-Charles, P.L.; Bilodeau, G.A.; Bergevin, R. SuBSENSE: A universal change detection method with local adaptive sensitivity. *IEEE Trans. Image Process.* **2014**, *24*, 359–373. [CrossRef] [PubMed]
20. Lim, L.A.; Keles, H.Y. Foreground segmentation using a triplet convolutional neural network for multiscale feature encoding. *Pattern Recognit. Lett.* **2018**, *112*, 256–262. [CrossRef]
21. Kim, W.; Kim, Y. Background subtraction using illuminance invariant structural complexity. *IEEE Signal Process. Lett.* **2016**, *23*, 634–638. [CrossRef]
22. Kim, W. Moving object detection using edges of residuals under varying illuminances. *Multimed. Syst.* **2019**, *25*, 155–163. [CrossRef]
23. Gautam, A.; Singh, S. Neural style transfer combined with EfficientDet for thermal surveillance. *Vis. Comput.* **2022**, *38*, 4111–4127. [CrossRef]
24. Chen, G.Q.; Duan, J.; Cai, H.; Liu, G.W. *Electronics, Communications and Networks IV*, 1st ed.; CRC Press: London, UK, 2015.
25. Su, J.; Zhang, G.; Wang, K. Compressed fusion of infrared and visible images combining robust principal component analysis and non-subsampled contour transform. *Laser Optoelectron. Prog.* **2020**, *57*, 041005.
26. Wagner, J.; Fischer, V.; Herman, M.; Behnke, S. Multispectral pedestrian detection using deep fusion convolutional neural networks. In Proceedings of the 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belgium, 27–29 April 2016.
27. Ding, L.; Wang, Y.; Laganière, R.; Huang, D.; Luo, X.; Zhang, H. A robust and fast multispectral pedestrian detection deep network. *Knowl.-Based Syst.* **2021**, *227*, 106990. [CrossRef]
28. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. *arXiv* **2016**, arXiv:1611.02644.

29. Zhang, L.; Lin, L.; Liang, X.; He, K. Is faster R-CNN doing well for pedestrian detection? In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14; Springer International Publishing: New York, NY, USA, 2016.
30. Konig, D.; Adam, M.; Jarvers, C.; Layher, G.; Neumann, H.; Teutsch, M. Fully convolutional region proposal networks for multispectral person detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017.
31. Li, C.; Song, D.; Tong, R.; Tang, M. Multispectral pedestrian detection via simultaneous detection and segmentation. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018; pp. 225.1–225.12.
32. 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]
33. 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.
34. 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]
35. Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; Liu, Z. Weakly aligned cross-modal learning for multispectral pedestrian detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5127–5137.
36. Park, K.; Kim, S.; Sohn, K. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognit.* **2018**, *80*, 143–155. [CrossRef]
37. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.W.; Heng, P.A. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [CrossRef] [PubMed]
38. Choi, Y.; Kim, N.; Hwang, S.; Park, K.; Yoon, J.S.; An, K.; Kweon, I.S. KAIST multi-spectral day/night data set for autonomous and assisted driving. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 934–948. [CrossRef]
39. Davis, J.W.; Sharma, V. Background-subtraction using contour-based fusion of thermal and visible imagery. *Comput. Vis. Image Underst.* **2007**, *106*, 162–182. [CrossRef]
40. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 743–761. [CrossRef]
41. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* **2024**, *145*, 109913. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.