electronics

MDPI

Article

Pedestrian Detection Method Based on Two-Stage Fusion of Visible Light Image and Thermal Infrared Image

Yugui Zhang [1], Bo Zhai [2], Gang Wang [3] and Jianchu Lin [4,*]

[1] Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China; zhangyugui@semi.ac.cn
[2] Beijing Institute of Astronautical Systems Engineering, China Academy of Launch Vehicle Technology, No. 1 Nandahongmen Road, Fengtai District, Beijing 100076, China; herom1985@163.com
[3] School of Computing and Data Engineering, NingboTech University, Ningbo 315100, China; wanggangnit@nit.zju.edu.cn
[4] Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huai'an 223003, China
* Correspondence: jianchulin@gmail.com

Abstract: Pedestrian detection has important research value and practical significance. It has been used in intelligent monitoring, intelligent transportation, intelligent therapy, and automatic driving. However, in the pixel-level fusion and the feature-level fusion of visible light images and thermal infrared images under shadows during the daytime or under low illumination at night in actual surveillance, missed and false pedestrian detection always occurs. To solve this problem, an algorithm for the pedestrian detection based on the two-stage fusion of visible light images and thermal infrared images is proposed. In this algorithm, in view of the difference and complementarity of visible light images and thermal infrared images, these two types of images are subjected to pixel-level fusion and feature-level fusion according to the varying daytime conditions. In the pixel-level fusion stage, the thermal infrared image, after being brightness enhanced, is fused with the visible image. The obtained pixel-level fusion image contains the information critical for accurate pedestrian detection. In the feature-level fusion stage, in the daytime, the previous pixel-level fusion image is fused with the visible light image; meanwhile, under low illumination at night, the previous pixel-level fusion image is fused with the thermal infrared image. According to the experimental results, the proposed algorithm accurately detects pedestrian under shadows during the daytime and low illumination at night, thereby improving the accuracy of the pedestrian detection and reducing the missed rate and false rate in the detection of pedestrians.

Keywords: pedestrian detection; visible light images; thermal infrared images; pixel-level fusion; feature-level fusion

check for updates

Citation: Zhang, Y.; Zhai, B.; Wang, G.; Lin, J. Pedestrian Detection Method Based on Two-Stage Fusion of Visible Light Image and Thermal Infrared Image. *Electronics* **2023**, *12*, 3171. https://doi.org/10.3390/electronics12143171

Academic Editor: Silvia Liberata Ullo

Received: 1 June 2023
Revised: 6 July 2023
Accepted: 17 July 2023
Published: 21 July 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In actual outdoor surveillance videos, when under low illumination at night and insufficient illumination in the daytime, the use of only visible light would easily result in missed detection and false detection of pedestrian objects, which cannot satisfy monitoring requirements throughout the day and thus cannot guarantee the security of people's lives and property [1–4]. Meanwhile, in terms of traffic safety and autonomous driving, video surveillance is also playing an increasingly important role in low illumination at night and insufficient illumination in the daytime [5,6]. With the advancement of technology, a variety of sensor devices have emerged, such as speed dome cameras, box cameras and high-definition sensors for use in the daytime, and thermal infrared, mid-wave infrared, and long-wave infrared sensors for use at night and in bad weather [1–3,7,8]. In order to satisfy all-day monitoring needs, it is necessary for different types of sensors to work together. Thus, it is a very challenging task to effectively utilize the data acquired by different types of sensors so as to improve the intelligent analysis level of surveillance video.

There are various spectral band data in nature. For each spectral band dataset, the information that it carries is different, as human eyes can only perceive part of the spectral band. In addition, owing to the limitation of imaging technology, there is currently no sensor device that can acquire all spectral band data [1–3,7]. For example, visible light imaging devices can be used to obtain rich visual information of objects such as texture, color, and shape from visible light spectrum bands that human eyes can perceive, whereas thermal infrared imaging devices can be used to obtain visual information of objects such as brightness, shape and contour from thermal infrared spectrum bands that human eyes cannot perceive. In particular, thermal infrared images are obtained according to the thermal radiation properties of objects, from which the object information such as contour and shape can be extracted under low illumination at night and shadows during the daytime [1,7–13]. To this end, this paper focuses on how to utilize the differences and complementarities between visible light images and thermal infrared images to obtain effective context information of pedestrians [14–22], thereby enhancing feature expression characteristics of pedestrians under low illumination at night and shadows during the daytime, and improving the detection accuracy of pedestrians.

To better illustrate the physical properties of visible light images and thermal infrared images, a visible light image and the corresponding thermal infrared image on the KAIST dataset [13] are shown in Figure 1.

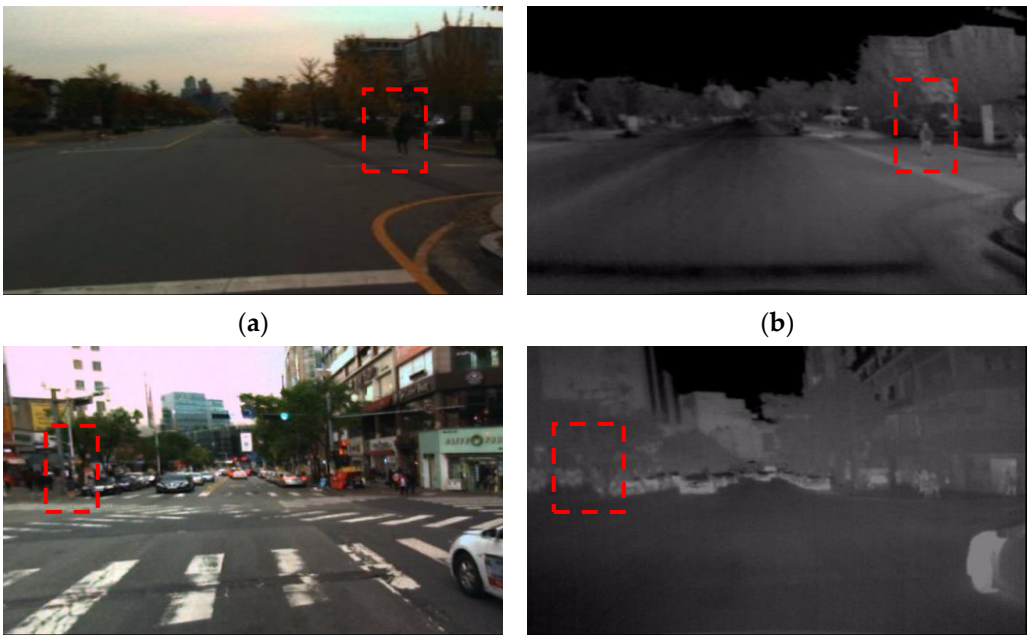


Figure 1. Visible light image and the corresponding thermal infrared image on the KAIST dataset: (**a**) visible light image, (**b**) corresponding thermal infrared image; (**c**) visible light image, (**d**) corresponding thermal infrared image.

From Figure 1, it can be seen that in the visible light image (a), the pedestrian marked by a red dotted box has no significant features and thus is difficult to distinguish even by human eyes, while in the corresponding thermal infrared image (b), the pedestrian information such as brightness, contour and shape can be better provided. In another visible light image (c), information such as color, texture, shape and contour is not clearly provided for the pedestrian marked by the red dotted box, while in the corresponding thermal infrared image (d), the pedestrian feature information is not significant, and thus it is different to distinguish the pedestrian even with human eyes. Therefore, it is necessary to study how to effectively utilize the physical properties of visible light images and thermal infrared images and fully exploit the complementarity between these two images so as to

improve pedestrian detection effects. In recent years, researchers have made many efforts in pixel-level fusion [10,23], feature-level fusion [9,13] and decision-level fusion [12] of visible light images and thermal infrared images. However, owing to the complexity and diversity of outdoor scenes and the changeable weather, it is difficult to meet complex and changeable application requirements by using existing methods.

For example, the pixel-level fusion image obtained from the visible light image and the thermal infrared image would lose color and part of the texture and background information of the scene as compared with the visible light image, and would lose part of the brightness information as compared with the thermal infrared image; therefore, it is difficult to effectively improve the detection accuracy of pedestrians only by using the pixel-level fusion image, especially because of the fact that the missed and false detection would easily occur. In some cases, given the influence caused by factors such as illumination, temperature, and weather, the pedestrian information in the visible light image, such as color, texture and shape could not be provided, and the pedestrian information in the thermal infrared image such as shape, brightness, and contour is not clearly obtained. In view of this, the feature-level fusion of the visible light image and the thermal infrared image could not significantly enhance the feature expression of the pedestrians. On the contrary, it will introduce interference information, which would easily result in missed detection and false detection of pedestrians. When the visible light image and the thermal infrared image are fused at the decision level, pedestrian detection is required to be performed respectively in the visible light image and the corresponding thermal infrared image, and then a mapping relationship and a master–slave relationship between the visible light detection results and the thermal infrared detection results is established. However, in actual monitoring, the weather changes at any time, which makes the pedestrian information in the visible light image and in the thermal infrared image not equivalent. Therefore, the mapping relationship and the master–slave relationship there between cannot be established to meet initial conditions for the decision-level fusion, and thus missed detections and false detections of the pedestrian object would easily occur. With this in mind, this paper studies and analyzes the physical properties of visible light images and thermal infrared images, and uses the complementarity between the two images to enhance the visually significant features and feature expression properties of pedestrians, so as to obtain the context information that is required for pedestrian detection and thereby improving pedestrian detection accuracy.

In order to improve the detection accuracy of pedestrians under low illumination at night and in the shadows during the daytime, this paper studies pedestrian detection based on the two-stage fusion of visible light images and thermal infrared images. This paper is distinguished from the existing references [11,13,22] mainly in that it makes full use of the complementarity and difference between visible light images and thermal infrared images to perform the two-stage fusion, including pixel-level fusion and feature-level fusion, according to the varying daytime conditions, so as to obtain effective pedestrian context information. The organization of this paper is as follows: Section 1 includes the introduction and existing problems; Section 2 introduces the existing pedestrian detection methods using pixel-level fusion, feature-level fusion and decision-level fusion of the visible light image and thermal infrared image; Section 3 introduces the proposed method; Section 4 introduces the experimental results; and Section 5 summarizes the content of this paper.

2. Related Work

In the past, thermal infrared imaging devices were mainly used for military applications. As the cost of the thermal infrared device falls, it is becoming more and more widely used in civilian applications, such as drone rescue and health monitoring [18,19]. The existing pedestrian detection methods based on the fusion of visible light images and thermal infrared images are mainly classified into the following three categories: pixel-level fusion, feature-level fusion and decision-level fusion [9].

2.1. Pixel-Level Fusion of Visible Light Images and Thermal Infrared Images

The pixel-level fusion of visible light images and thermal infrared images can effectively fuse complementary information in these two images, thereby enhancing the significance of objects in the images under shadows during the daytime or under low illumination at night, and improving the detection accuracy of objects. For instance, Davis et al. [10] present a background-subtraction technique fusing contours from thermal and visible imagery for persistent object detection in urban settings. Background-subtraction in the thermal domain is used to identify the initial regions of interest. Color and intensity information are used within these areas to obtain the corresponding regions of interest in the visible domain. Within each region, input and background gradient information are combined to form a Contour Saliency Map. The binary contour fragments, obtained from corresponding Contour Saliency Maps, are then fused into a single image. An A*path-constrained search along watershed boundaries of the regions of interest is used to complete and close any broken segments in the fused contour image. Lastly, to facilitate subsequent pedestrian detection, the contour image is flood-filled to produce silhouettes. Choi et al. [23] propose a human detection method with thermal and visible images using a joint bilateral filter. To fuse the best features of both types of images in order to achieve superior performance in human detection, a pixel-level image fusion is used, and to carry out this image fusion process, the joint bilateral filter is used for fusion of the edge information in the visible image and the white region in the thermal image. This image fusion process effectively removes the regions of human shadow. Results from different pairs of fused images that are taken sequentially are subtracted to detect the human effectively.

2.2. Feature-Level Fusion of Visible Light Images and Thermal Infrared Images

With the widespread application of pedestrian detection, pedestrian detection datasets have developed rapidly in the past decades. However, most of the existing pedestrian detection datasets are directed to visible light images, while complete object contour, brightness and part of texture information under low illumination at night and under shadows during the daytime can only be obtained from thermal infrared images. In view of this, Hwang [13] proposes a multispectral pedestrian dataset that provides well-aligned color-thermal image pairs, captured by beam splitter-based special hardware. The color-thermal dataset is as large as previous color-based datasets and provides dense annotations, including temporal correspondences. With this dataset, the authors introduce multispectral aggregated channel features (ACFs) [9], which is an extension of the standard ACFs to simultaneously handle color-thermal image pairs. This method achieves another breakthrough in the pedestrian detection task by using the feature-level fusion of the visible light image and the thermal infrared image, thereby improving pedestrian detection performance under shadows during the daytime and under low illumination at night.

2.3. Decision-Level Fusion of Visible Light Images and Thermal Infrared Images

Existing visible light images in video surveillance have deficiencies for detection performance under low illumination at night and under shadows during the daytime. In order to improve the accuracy and intelligence of outdoor surveillance video throughout the whole day, intelligent surveillance systems that can automatically adapt to both day and night are becoming more and more popular. Torresan [22] relies on the use of pairs of video (visible spectrum) and thermal infrared (TIR) cameras located around premises of interest. To automate the system, a dedicated image processing approach is required. These image sequences (video and TIR) are synchronized, geometrically corrected and temperature calibrated. The next step is to develop a segmentation strategy to extract the regions of interest (ROIs) corresponding to pedestrians in the images. Finally, a mapping relationship and a master–slave relationship are established for pedestrian detection results from visible light images and thermal infrared images, and the mapping relationship and the master–slave relationship are changed when the subsequent detection results change. The basis for the decision-level fusion of visible light images and thermal infrared images

is that there are corresponding detection results in both visible light detection results and thermal infrared detection results for the establishment of the mapping relationship and the master—slave relationship. However, owing to the influences of the varying daytime, the detection results cannot comply with this correspondence; thus, the current decision-level fusion method cannot adequately handle the recovery of the master–slave relationship to only re-establish the mapping scheme, which reduces the performance of object detection.

3. The Proposed Method

The proposed method in the paper, specifically the visible light images and thermal infrared images, are subjected to pixel-level fusion and feature-level fusion according to the varying daytime conditions. In the pixel-level fusion stage, the thermal infrared image, after being brightness enhanced, is fused with the visible image. The obtained pixel-level fusion image contains the information critical for accurate pedestrian detection, such as pedestrian context, shape, contour, and background in the scene. In the feature-level fusion stage, in the daytime, the previous pixel-level fusion image is fused with the visible light image so as to make full use of the color, texture and background information in the scene of the visible light image; meanwhile, under low illumination at night, the previous pixel-level fusion image is fused with the thermal infrared image, so as to make full use of the significant information of thermal infrared images, such as pedestrian brightness, shapes and contours.

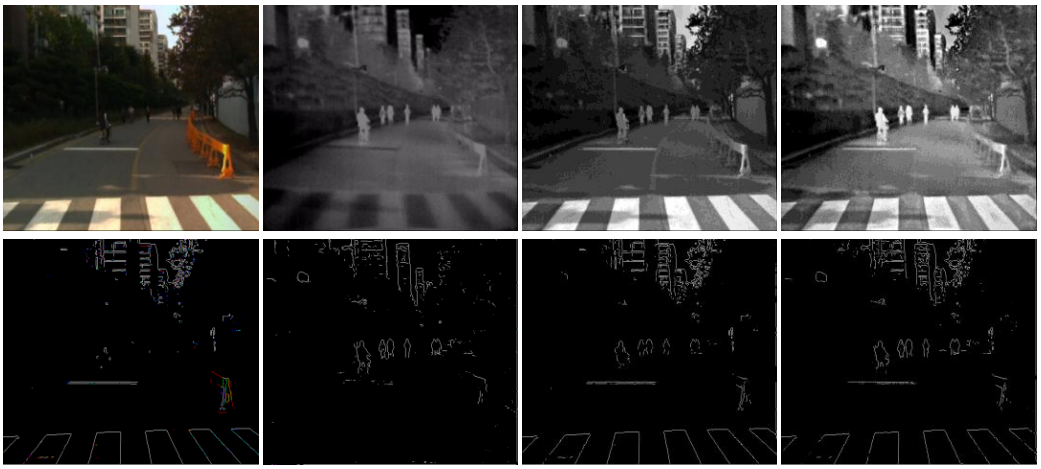
3.1. Stage I: Pixel-Level Fusion of Visual Light Images and Thermal Infrared Images

Under shadows in the daytime, it is difficult to detect pedestrians in the visible light images, while corresponding thermal infrared images have clear image information such as brightness and shape of pedestrians. Under low illumination at night, visible light images have certain information such as color, texture and background of the scene, but do not have complete edges, textures and shapes of pedestrians, while corresponding thermal infrared images have information such as brightness, shape and contour of pedestrians. Moreover, owing to the influence of imaging sensing devices, Gaussian noise or salt–pepper noise are prone to occur in visible light images [24–28]. In view of all this, many studies have been done on the fusion, especially the pixel-level fusion, of visible light images and thermal infrared images by utilizing the complementarity between the two types of images. In the literature [29], visible light images and thermal infrared images were separately visually enhanced before pixel-level fusion, resulting in overly strong contrast in the enhanced visible light images; thus, it is difficult to effectively extract pedestrian features from the fusion results. To solve this problem, this paper proposes that only the brightness of thermal infrared images is enhanced before the pixel-level fusion of visible light images and thermal infrared images. By doing so, pedestrian features can be effectively extracted from the fusion result, which conforms to human visual perception, as specifically shown in the following Formula (1):

$$norm = (p(x,y) - \min(p))/(\max(p) - \min(p)) * 255 \quad (1)$$

where $p(x,y)$ represents a pixel value at the current position (x,y), $\min(p)$ represents the minimum value of pixels in the current thermal infrared image, and $\max(p)$ represents the maximum value of pixels in the current thermal infrared image. The visible light image and the brightness-enhanced thermal infrared image are fused at the pixel level through guided filtering technology to significantly enhance the visual appearance information of pedestrians. The strategies and methods of selecting parameters are explained in the literature [29–37].

Figure 2 depicts, under shadows in the daytime, the visible light image (a), the corresponding thermal infrared image (b), the pixel-level fusion image of the visible light image and the thermal infrared image (c), and the pixel-level fusion image of the visible light image and the brightness-enhanced thermal infrared image as proposed in this paper (d), as well as the corresponding edge features of each of these images.



(**a**) (**b**) (**c**) (**d**)

Figure 2. Different types of images and corresponding edge features under shadows in the daytime: (**a**) visible light image and corresponding edge features; (**b**) thermal infrared image and corresponding edge features; (**c**) the non-brightness-enhanced fusion image and corresponding edge features; (**d**) the brightness-enhanced fusion image and corresponding edge features.

From Figure 2, it can be seen that under shadows in the daytime, as compared with the visible light image (a), the thermal infrared image (b) and the pixel-level fusion image of thermal infrared image and visible light image (c), the pixel-level fusion image of the brightness-enhanced thermal infrared image and visible light image as proposed in this paper (d) has more significant pedestrian information and corresponding bottom-layer edge features. The results show that the pixel-level fusion image as proposed in this paper has clearer pedestrian information such as context, partial texture, shape and background of the scene, which are essential for achieving accurate pedestrian detection. By making full use of the above information, it is possible for pedestrian detectors to better distinguish pedestrians and backgrounds, thereby improving the accuracy of pedestrian detection.

The effectiveness of the pixel-level fusion method proposed in this paper is further demonstrated below by using the image quality metrics, including information entropy, average gradient and edge strength [27,38]. The results are shown in Table 1.

Table 1. The image quality of different images during the daytime.

Image \ Metrics	Information Entropy	Average Gradient	Edge Strength
Visible light image	6.984	2.723	28.452
Thermal infrared image	6.645	2.028	21.317
Pixel-level fusion image of non-brightness-enhanced thermal infrared image and visible light image	7.198	3.611	37.648
Pixel-level fusion image of brightness-enhanced thermal infrared image and visible light image	7.625	5.035	52.818

It can be seen from Table 1 that the pixel-level fusion of the brightness-enhanced thermal infrared image and the visible light image results in a fusion image have higher information entropy, average gradient and edge strength, as compared with the visible light image, thermal infrared image and non-brightness-enhanced fusion image. The higher the information entropy, average gradient and edge strength, the richer the information contained in the image. In particular, more pedestrian context information indicates higher image quality and thus a clearer image. Therefore, although the pedestrian object under

shadow in the daytime does not have complete bottom-layer visual information, since the fusion image obtained by the proposed method contains more pedestrian context information and scene background information, the pedestrian and the background can be better distinguished, and thus the pedestrian detection accuracy can be improved.

Figure 3 depicts, under low illumination at night, the visible light image (a), corresponding thermal infrared image (b), pixel-level fusion image of the visible light image and the corresponding thermal infrared image (c), and pixel-level fusion image of the visible light image and the brightness-enhanced thermal infrared image (d), as well as the corresponding edge features of each of these images.

(**a**) (**b**) (**c**) (**d**)

Figure 3. Different types of images and corresponding edge features under low illumination at night: (**a**) visible light image and corresponding edge features; (**b**) thermal infrared image and corresponding edge features; (**c**) non-brightness-enhanced fusion image and corresponding edge features; (**d**) brightness-enhanced fusion image and corresponding edge features.

As shown in Figure 3, the pixel-level fusion image of the visible light image and the brightness-enhanced thermal infrared image has more significant context, shape and contour information of the pedestrian and the scene background information, as well as pedestrian edge features, all of which are beneficial to improve the accuracy of pedestrian detection.

As indicated by the evaluation metrics of the image quality, including information entropy, average gradient and edge intensity, under low illumination at night, the pixel-level fusion image obtained by the proposed method has greater detail and thus is clearer. The results are shown in Table 2.

Table 2. The image quality of different images under low illumination at night.

Image \ Metrics	Information Entropy	Average Gradient	Edge Strength
Visible light image	6.692	3.556	37.292
Thermal infrared image	5.053	0.485	5.111
Pixel-level fusion image of non-brightness-enhanced thermal infrared image and visible light image	6.716	3.681	38.649
Pixel-level fusion image of brightness-enhanced thermal infrared image and visible light image	6.979	3.940	41.368

It can be seen from Table 2 that, as compared with the visible light image, thermal infrared image, and pixel-level fusion image of the visible light image and the thermal infrared image, the pixel-level fusion image of the visible light image and the brightness-enhanced thermal infrared image have higher information entropy, average gradient and edge strength. Therefore, the pixel-level fusion image of the visible light image and the brightness-enhanced thermal infrared image contains more information, such as pedestrian context, shape, contour and scene background, all of which make pedestrians and backgrounds more distinguishable, and thus is beneficial for improving pedestrian detection accuracy.

In conclusion, the pixel-level fusion image of the visible light image and the brightness-enhanced thermal infrared image has rich information, including pedestrian context, shape, brightness, contour and scene background, both under shadows in the daytime and under low illumination at night; this information contributes to the improvement in the accuracy of pedestrian detection throughout the whole day.

3.2. Stage II: Combination of Pixel-Level Fusion and Feature-Level Fusion for Pedestrian Detection

In order to meet all-day monitoring requirements of the monitoring system, under shadows in the daytime, the pixel-level fusion image obtained in the first stage is fused at a feature level with the visible light image so as to compensate for its loss of color, texture and scene background information as compared with the visible light image. While under low illumination at night, the pixel-level fusion image obtained in the first stage is fused at a feature level with the thermal infrared image so as to compensate for its loss of some brightness information of pedestrians as compared with the thermal infrared image. Therefore, on the basis of the pixel-level fusion obtained the first stage, this part is directed to the selection of visible light images or thermal infrared images for feature-level fusion according to the varying daytime conditions, so as to improve the all-day accuracy of pedestrian detection.

3.2.1. Feature-Level Pedestrian Detection Method Based on the Combination of the Pixel-Level Fusion Image and the Visible Light Image

In the daytime, the color, texture and scene background information in the visible light image are important factors for achieving accurate pedestrian detection [28,38,39]. Keeping this in mind, when visible light images and thermal infrared images are fused at a feature level by using existing feature-level fusion methods [2,9,13], the effect of increasing the ratio of visible light images to thermal infrared images on pedestrian detection performance is studied. The increase in said ratio is meant to increase the proportion of color, texture and scene background information in the visible light image during the feature-level fusion. The pedestrian detection results are shown in Table 3.

Table 3. Effect of the ratio of visible light images to thermal infrared images on pedestrian detection performance.

The Ratio of Visible Light Images to Thermal Infrared Images	AP	$AP^{0.5}$	$AP^{0.75}$	AP^{S}	AP^{M}	AP^{L}
1:1	0.742	0.957	0.874	0.689	0.745	0.791
2:1	0.742	0.957	0.872	0.687	0.744	0.794
3:1	0.739	0.956	0.871	0.685	0.741	0.789
3:2	0.737	0.957	0.863	0.686	0.738	0.795
4:1	0.739	0.957	0.871	0.686	0.740	0.794
5:1	0.742	0.965	0.871	0.681	0.743	0.791

It can be seen from Table 3 that in the daytime, increasing the ratio of visible light images to thermal infrared images essentially has no effect on the evaluation metrics of pedestrian detection, including average accuracy, indicating that the accuracy of pedestrian detection has not been effectively improved. For the specific meaning of the evaluation metrics AP, $AP^{0.5}$, $AP^{0.75}$, AP^{S}, AP^{M} and AP^{L}, please refer to the introduction of the evaluation metrics in Section 4.2. The main reason for this is that, although increasing the proportion of visible light images enhances the color and texture information of pedestrians in the visible light images, it does not enhance the context and scene background information of pedestrians under shadows and at distance. On the other hand, in the daytime, the pixel-level fusion image of the visible light image and the thermal infrared image loses color as well as part of the texture and scene background information as compared with the visible light image, whereas the color, texture and scene background information are absent in the thermal infrared image. In view of the above, we propose a feature-level fusion of the pixel-level fusion image with the visible light image so as to compensate for the missing color, part of the texture, and scene background information, thus realizing the accurate detection of pedestrians.

Based on the above analysis, in the daytime, the pixel-level fusion image of the visible light image and the thermal infrared image is fused with the visible light image at the feature level, so as to enhance the bottom-layer and middle-layer visual features required for accurate pedestrian detection, such as the context, texture and color of pedestrians and background information in the scene, as shown in Formula (2):

$$f_{res}^{(n)} = vis^{(n)} \oplus f_{visir}^{(n)} \tag{2}$$

where $vis^{(n)}$ represents the feature map obtained from the extraction of the nth frame of the visible light image, $f_{visir}^{(n)}$ represents the feature map obtained from the extraction of the nth frame of the pixel-level fused image, $\oplus$ represents pixel-by-pixel addition, and $f_{res}^{(n)}$ represents the feature-level fusion result. By using this method, the feature-level fusion of the pixel-level fusion image and the visible light image can be obtained, which improves the detection accuracy of pedestrians in the daytime. The specific flowchart is shown in Figure 4.

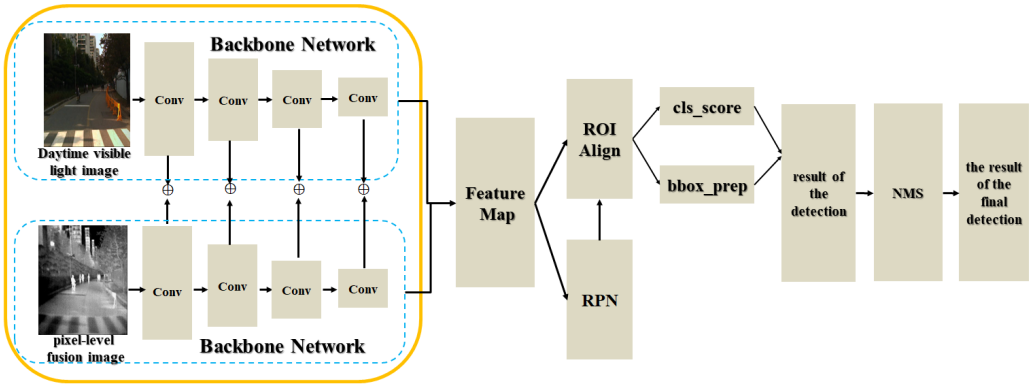


Figure 4. The pedestrian detection flowchart of feature-level fusion of the pixel-level fusion image and the visible light image.

As shown in Figure 4, in order to perform the feature-level fusion of pixel-level fusion images and visible light images, a multi-stage and currently popular Cascade R-CNN network model [40] for object detection based on the feature pyramid [41] is adopted as the pedestrian detection network model, ResNet-50 [35] is used as the backbone network, a stochastic gradient descent function is used as the optimization procedure, and the learning rate is set to 0.001, the weight decay is set to 0.0005, and the momentum is set to 0.9. The training data is expanded by image data expansion methods such as flipping, scaling, and

rotation, so that the pedestrian features are enhanced and, ultimately, the accurate detection of pedestrians in outdoor surveillance videos is achieved.

3.2.2. Feature-Level Pedestrian Detection Method Based on the Combination of the Pixel-Level Fusion Image and the Thermal Infrared Image

Under low illumination at night, visible light images contain salt–pepper noise and Gaussian noise [24,27], which would reduce the accuracy of pedestrian detection. In addition, visible light images do not have effective bottom-layer visual information for pedestrians, while thermal infrared images contain bottom-layer visual information such as brightness, contour and shape of pedestrians. In view of this, in the feature-level fusion of visible light images and thermal infrared images, the effect of increasing the ratio of thermal infrared image to visible light image on pedestrian detection performance is studied. The increase in said ratio is meant to increase the proportion of brightness, contour and shape of pedestrians in thermal infrared images during the feature-level fusion. The pedestrian detection performance results are shown in Table 4.

Table 4. Effect of the ratio of thermal infrared images to visible light images on pedestrian detection performance.

The Ratio of Thermal Infrared Images to Visible Light Images	AP	$AP^{0.5}$	$AP^{0.75}$	AP^{S}	AP^{M}	AP^{L}
1:1	0.751	0.965	0.857	0.597	0.758	0.813
1:2	0.751	0.965	0.858	0.598	0.760	0.802
1:3	0.737	0.965	0.855	0.562	0.747	0.793
2:3	0.757	0.966	0.866	0.609	0.764	0.805
1:4	0.749	0.965	0.857	0.577	0.758	0.803
1:5	0.749	0.965	0.856	0.584	0.756	0.803

It can be seen from Table 4 that increasing the ratio of the thermal infrared image to the visible light image essentially has no effect on the evaluation metrics of pedestrian detection, including average accuracy, indicating that the accuracy of pedestrian detection has not been effectively improved. The main reason is that, although increasing the proportion of thermal infrared images improves the brightness and contour information of pedestrians, it cannot improve the context, texture and scene background information, so that the accuracy of pedestrian detection under low illumination at night cannot be effectively improved. On the other hand, under low illumination at night, the pixel-level fusion image of the visible light image and the thermal infrared image loses some brightness information for pedestrians and edge contour information as compared with the thermal infrared image [42], while visible light images cannot compensate for said information loss, since they contain no effective bottom-layer visual features of pedestrians. Therefore, we propose a feature-level fusion of the pixel-level fusion image with the thermal infrared image to compensate for the bottom-layer visual information, such as partial brightness and edge contours of pedestrians, thereby improving the detection accuracy of pedestrians.

Based on the above analysis, under low illumination at night, the pixel-level fusion image and the thermal infrared image are fused at a feature level so that the pedestrian features can be enhanced, as shown in Formula (3):

$$f_{res}^{(n)} = ir^{(n)} \oplus f_{visir}^{(n)} \tag{3}$$

where $ir^{(n)}$ represents the feature map extracted from the nth frame of thermal infrared image, $f_{visir}^{(n)}$ represents the feature map extracted from the nth frame of the pixel-level fusion image, and $f_{res}^{(n)}$ represents the result of feature-level fusion. By using this formula, the feature-level fusion of the pixel-level fusion image and the thermal infrared image

can be performed, which improves the accuracy of pedestrian detection. The detection flowchart is shown in Figure 5:

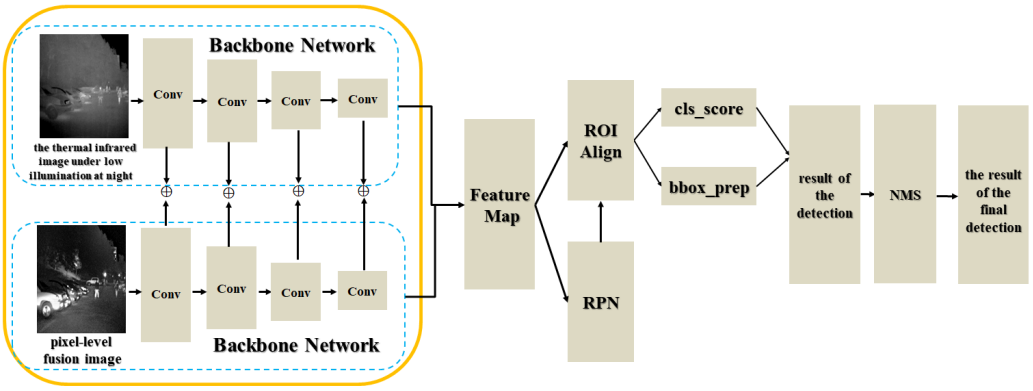


Figure 5. The pedestrian detection flowchart of feature-level fusion of the pixel-level fusion image and the thermal infrared image.

In Figure 5, during the feature-level fusion of the pixel-level fusion image with the thermal infrared image, the specific settings of the parameters, including the network model, optimization function, learning rate, weight decay and momentum, are analogous to the above model parameter settings in Figure 4. By using the feature-level fusion of pixel-level fusion images with thermal infrared images, the pedestrian features can be enhanced, and thus the accuracy of pedestrian detection under low illumination at night can be improved.

4. Experimental Results

4.1. Introduction of Datasets and Experiments

The effectiveness of the proposed method is evaluated using the publicly available KAIST dataset [13], in which visible light images and corresponding thermal infrared images are contained in pairs. The images in this dataset were acquired in urban areas with pedestrians by sensor devices mounted on the roof of cars at a speed of 30–55 km/h. The dataset contains visible light image and thermal infrared image data in various scenes under different daytime conditions, including shadows during the daytime and low illumination at night. The above two types of images have a resolution of 640 × 512, and the pedestrians in the images have a height ranging from 45 to 115 pixels. In this dataset, a total of 95,328 visible light images and corresponding thermal infrared images are manually annotated, resulting in 103,128 annotated boxes with 1182 pedestrian objects. In this dataset, the visible light image under illumination during the daytime has complete bottom-layer visual features such as color, contour and shape of the pedestrians. However, under shadows and regarding small objects at a distance, the pedestrians and backgrounds in the visible light image are not easily distinguishable, whereas the corresponding thermal infrared image has complete bottom-layer visual information, including pedestrian brightness, contour and shape. Moreover, under low illumination at night, the shape and contour of pedestrians in the visible light image cannot be distinguished, whereas the corresponding thermal infrared images exhibits significant bottom-layer visual information, such as shape, contour and brightness of pedestrians. The specific image data in this dataset is shown in Figure 6.

(**a**) (**b**) (**c**) (**d**)

Figure 6. Visible light image and corresponding thermal infrared image from the KAIST dataset. (**a**) The image in the daytime; (**b**) the image under shadows in the daytime; (**c**) the image under low illumination at night; (**d**) the image under low illumination at night.

As shown in Figure 6, the first and second columns are the visible light image and corresponding thermal infrared image under shadows during the daytime, and the third and fourth columns are the visible light image and corresponding thermal infrared images under low illumination at night. Specifically, under shadows during the daytime, the visible light image contains little effective pedestrian information, rendering the pedestrians not easily detectable, while the corresponding thermal infrared image can provide the bottom-layer visual information, such as brightness, shape and contour, required for pedestrian detection. Under low illumination at night, the visible light image has large noise interference, rendering the pedestrians not easily detectable, while the corresponding thermal infrared image has significant information, such as brightness, shape and contour, for pedestrian detection. Therefore, considering the difference and complementarity between visible light images and thermal infrared images, these two images are subsequently fused at a pixel level and then at a feature level according to the varying daytime conditions, so that the accuracy of pedestrian detection can be improved and thus accuracy can be achieved throughout the day.

4.2. Evaluation Metrics

(1) Evaluation Metrics of the Image Quality

In order to effectively evaluate the image quality, the following evaluation metrics were used [27,43]: Entropy (E), Average Gradient (AG) and Edge Intensity (EI). Specifically, the entropy represents the amount of information contained in the image. The greater the entropy, the more information the image contains. The Entropy (E) is obtained by Formula (4):

$$E = -\sum_{i=0}^{L-1} p_i \log(p_i) \tag{4}$$

where L represents the gray level of the image, and P_i represents the probability of the pixel value i in the image.

The average gradient reflects the clearness of the image. The larger the average gradient, the clearer the image. In addition, the average gradient can also be used to measure the spatial resolution of the image, and the average gradient increases with the increase of the spatial resolution. The specific calculation is shown in Formula (5):

$$AG = \frac{1}{M \times N} \sum_{x=2}^{M} \sum_{y=2}^{N} \sqrt{\left(\Delta I_x^2 + \Delta I_y^2\right)/2} \tag{5}$$

where I represents the image, ΔI_x^2 and ΔI_y^2 are the gradients in the horizontal and vertical directions, respectively, and $M \times N$ is the size of the image.

The edge intensity can reflect the quality and clearness of the image. The higher the edge intensity, the higher the quality and clearness of the image. The edge strength of the image is measured using the Sobel operator, as shown in Formulas (6) and (7):

$$S_x = I * h_x,\ S_y = I * h_y \tag{6}$$

$$EI = \sqrt{S_x^2 + S_y^2} \tag{7}$$

where I represents the image, and $h_x = \begin{pmatrix} 1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}$ and $h_y = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}$ represent the gradient operators in the horizontal and vertical directions, respectively.

(2) Evaluation Metrics of the Pedestrian Detection

We use the Average Precision (AP) as defined in the coco dataset [44] to evaluate our algorithm.

AP@[0.5:0.95] corresponds to the AP for IOU from 0.5 to 0.95 with a step size of 0.05.
AP^{IoU} = 0.50 is AP at IoU = 0.50. AP^{IoU} = 0.75 is AP at IoU = 0.75.
AP^{Small} is AP for small objects: area < 32^2.
AP^{Medium} is AP for medium objects: 32^2 < area < 96^2.
AP^{Large} is AP for large objects: area > 96^2.

During training and testing, the input images are scaled to 512 pixels in height and 640 pixels in width. NMS (Non-Maximum Suppression) is used for the final detection results.

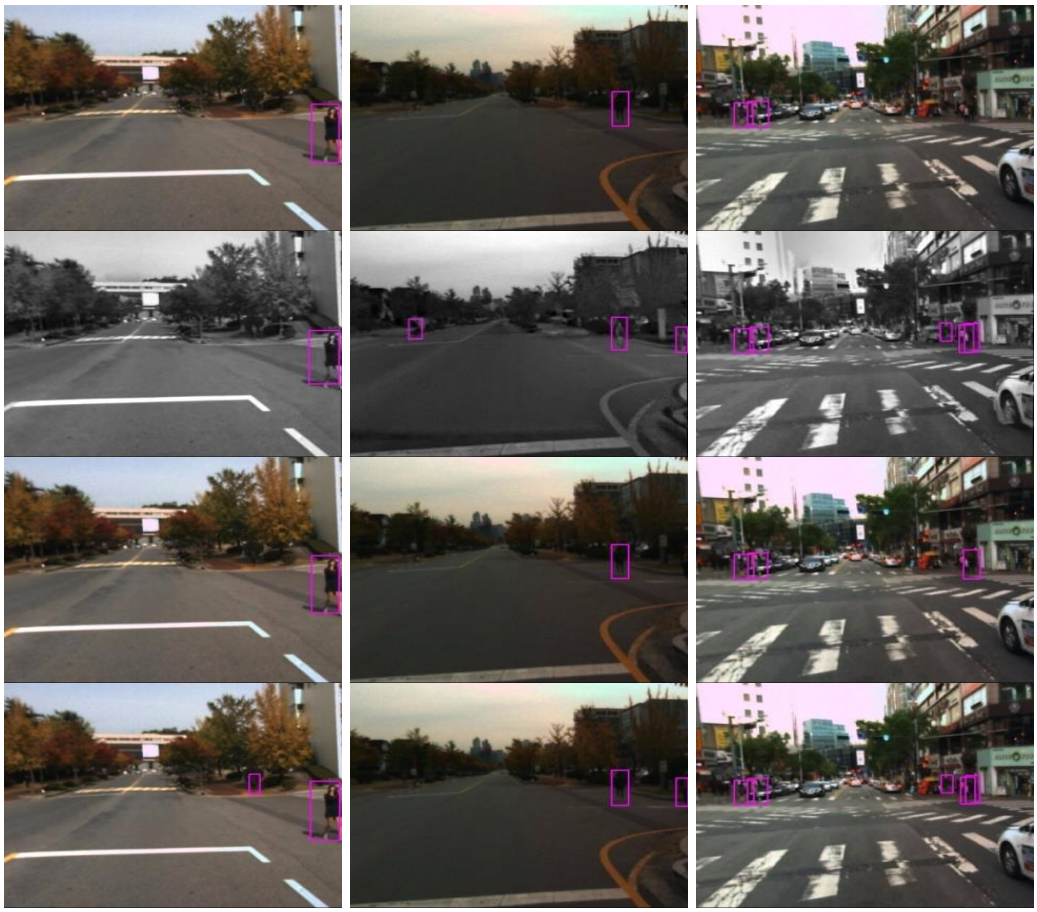
4.3. Pedestrian Detection Results Based on the Combination of the Pixel-Level Fusion and the Feature-Level Fusion

In view of the complementarity and difference between visible light images and thermal infrared images, the pixel-level fusion image is further fused at a feature level with a different original image according to the varying daytime conditions, such as with a visible light image under shadows during the daytime or with a thermal infrared image under low illumination at night, so that the feature-expressing ability of pedestrians can be significantly enhanced, and thus the accuracy of pedestrian detection in outdoor surveillance can be improved.

In order to verify the method proposed in this paper, we split the KAIST dataset into the training set, the verification set and the test set, with a ratio of 8:1:1. All comparison methods are trained for 12 epochs, and all comparison results take the last epoch as a comparison of the final results.

(a) Pedestrian detection results based on the feature-level fusion of the pixel-level fusion image and the visible light image

From the pixel-level fusion image of the visible light image and thermal infrared image, it can be seen from Figure 2 that under shadows in the daytime, said pixel-level fusion image has the bottom-layer visual information, including context, brightness, contour and shape, required for accurate pedestrian detection, but loses the information of color, texture and scene background contained in the original visible light image. Therefore, the pixel-level fusion image is further fused at a feature level with the original visible light image, so as to compensate the pixel-level fusion image for its loss of the color and texture information, thereby improving the accuracy of the pedestrian detection. The effectiveness of the proposed pedestrian detection method is evaluated in terms of visual effect and performance metrics. The visual effect is shown in Figure 7.



(**a**) Scene 1 (**b**) Scene 2 (**c**) Scene 3

Figure 7. Scene 1, scene 2 and scene 3 show feature-level fusion results of the pixel-level fusion image and the visible light image in the daytime in the KAIST dataset.

As can be seen from Figure 7, based on the same data partition method, the method proposed in this paper can detect more pedestrian objects as compared with other methods, and thus can effectively reduce the missed and false detection of pedestrians. Specifically, in the first line, when only the visible light image is used for pedestrian detection, there exists the missed detection of pedestrians, especially regarding pedestrians under shadows, pedestrians with few pixels at distance, and pedestrians with a similar appearance with the background. In the second line, when the pixel-level fusion image of visible light image and thermal infrared image is used for pedestrian detection, there exists the missed and false detection of pedestrians, especially regarding pedestrians under shadows and pedestrians with a similar appearance with the background. In the third line, when the feature-level fusion image of visible light image and thermal infrared image is used for pedestrian detection, there exists the missed detection of pedestrians, especially regarding pedestrians under shadows and pedestrians with few effective pixels at distance. In the fourth line, the method proposed in this paper can better detect pedestrian objects, including pedestrians under shadows, pedestrians with few effective pixels at distance, and pedestrians with a similar appearance with the background. Therefore, the proposed method can reduce the missed and false detection rate for the pedestrians and thus achieves accurate pedestrian detection.

The effectiveness of the proposed method is further evaluated using the performance evaluation metrics [44], including the average accuracy rate AP, $AP^{0.5}$ and $AP^{0.75}$, and so on, as shown in Table 5.

Table 5. Pedestrian detection performance of different detection methods in the daytime.

Evaluation Metrics	AP	$AP^{0.5}$	$AP^{0.75}$	AP^{S}	AP^{M}	AP^{L}
Detection of the visible light image	0.724	0.956	0.848	0.660	0.724	0.789
Detection of the pixel-level fusion of the visible light image and the thermal infrared image	0.704	0.955	0.821	0.634	0.706	0.774
Detection of the feature-level fusion of the visible light image and the thermal infrared image	0.742	0.956	0.874	0.689	0.745	0.791
Detection using the proposed method	0.764	0.980	0.899	0.706	0.766	0.814

As can be seen from Table 5, the proposed method in this paper achieves higher AP as compared with those based on the visible light image, the pixel-level fusion image and the feature-level fusion image. Specifically, as compared with the other three detection methods, the average accuracy of the proposed method is improved by 4.2%, 6.0% and 2.2%, respectively. Therefore, the proposed method in this paper can improve the accuracy of pedestrian detection and reduce the missed and false detection rate of pedestrians, thus realizing the accurate detection of pedestrians in outdoor surveillance videos.

(b) Pedestrian detection results based on the feature-level fusion of the pixel-level fusion image and the thermal infrared image

Under low illumination at night, the pixel-level fusion image has the bottom-layer visual information, including context, contour, shape and background of the scene, required for accurate pedestrian detection, but loses part of the brightness information contained in the original thermal infrared image. In view of this, the pixel-level fusion image is further fused at a feature level with the original thermal infrared image, so as to compensate the pixel-level fusion image for its loss of part of brightness information, thereby improving the accuracy of the pedestrian detection. The effectiveness of the proposed method is evaluated in terms of visual effect and performance index. The visual effect is shown in Figure 8.

As can be seen from Figure 8, the method proposed in this paper can detect more pedestrian objects and reduce the false detection of pedestrians as compared with other methods, and thus can effectively reduce the missed and false rate of pedestrian detection. Specifically, in the first line, when only the thermal infrared image is used for pedestrian detection, there exists the missed detection of pedestrians, especially regarding pedestrians with few pixels at distance and pedestrians that are partially occluded. In the second line, when the pixel-level fusion image of the visible light image and thermal infrared image is used for pedestrian detection, there exists the missed and false detection of pedestrians, especially regarding pedestrians with few effective pixels at distance and pedestrians under interference caused by background similarities such as street lights. In the third line, when the feature-level fusion image of visible light image and thermal infrared image is used for pedestrian detection, there exists the missed detection of pedestrians, especially regarding pedestrians with few effective pixels at distance and pedestrians that are partially occluded. In the fourth line, the method proposed in this paper can better detect pedestrian objects, including pedestrians with few effective pixels at distance, pedestrians that are partially occluded, and pedestrians under interference caused by background similarities such as street lights. Therefore, the proposed method can reduce the missed and false detection rate for the pedestrians and thus achieves accurate pedestrian detection under low illumination at night.

(**a**) Scene 4 (**b**) Scene 5 (**c**) Scene 6

Figure 8. Scene 4, scene 5 and scene 6 show feature-level fusion results of the pixel-level fusion image and the thermal infrared image under low illumination at night in the KAIST dataset.

The effectiveness of the proposed method is further evaluated using the performance evaluation metrics of the average accuracies AP, $AP^{0.5}$ and $AP^{0.75}$, and so on, as shown in Table 6.

Table 6. Pedestrian detection performance of different detection methods under low illumination at night.

Evaluation Metrics	AP	$AP^{0.5}$	$AP^{0.75}$	AP^{S}	AP^{M}	AP^{L}
Detection of the thermal infrared image	0.786	0.978	0.900	0.689	0.792	0.806
Detection of the pixel-level fusion of the visible light image and the thermal infrared image	0.744	0.965	0.855	0.546	0.757	0.778
Detection of the feature-level fusion of the visible light image and the thermal infrared image	0.737	0.966	0.857	0.567	0.746	0.800
Detection using the proposed method	0.824	0.986	0.943	0.741	0.830	0.836

As can be seen from Table 6, the proposed method in this paper achieves higher average accuracy as compared with those based on the thermal infrared image, the pixel-level fusion image and the feature-level fusion image. Specifically, as compared with the other three detection methods, the average accuracy of the proposed method is improved by 3.8%, 8.0% and 8.7%, respectively. Therefore, under low illumination at night, the

proposed method in this paper improves the accuracy of pedestrian detection and reduces the missed and false detection rate of pedestrians, thus realizing the accurate detection of pedestrians in outdoor surveillance videos.

5. Conclusions and Future

This paper proposed a method for pedestrian detection based on the pixel-level fusion of visible light image and thermal infrared image and the feature-level fusion of the two images according to the varying daytime conditions. In particular, in the pixel-level fusion stage, the thermal infrared image was firstly enhanced in terms of its brightness, and then was fused with the visible light image at the pixel level to obtain a pixel-level fusion image, which contains the information of context, contour, shape and background of the scene required for accurate pedestrian detection. In the feature-level fusion stage, under shadows during the daytime, the pixel-level fusion image was fused with the visible light image at the feature level to obtain a feature-level image, which compensates for the information loss of the pixel-level fusion image, such as color as well as part of the texture and background of the scene contained in the visible light image; meanwhile, under low illumination at night, the pixel-level fusion image was fused with the thermal infrared image at the feature level to obtain a feature-level image, which compensates the pixel-level fusion image for its loss of part of the brightness information contained in the thermal infrared image. The experimental results show that the proposed method achieves accurate pedestrian detection. The proposed method still has some aspects that require improvement. For example, the proposed method cannot automatically discriminate between the varying daytime conditions for the corresponding detection model, and it does not consider the effect of misalignment of data collected by different sensor devices on the detection accuracy. Nonetheless, the proposed method exhibits a novel method for detecting pedestrians, which can be potentially used in outdoor surveillance videos.

Author Contributions: Conceptualization, Y.Z. and J.L.; Methodology, Y.Z.; Software, Y.Z.; Validation, Y.Z., B.Z., G.W. and J.L.; Formal analysis, Y.Z.; Investigation, Y.Z. and Y.Z.; Resources, B.Z. and J.L.; Data curation, B.Z.; Writing—original draft preparation, Y.Z.; Writing—review and editing, B.Z. and G.W.; Visualization, G.W.; Supervision, J.L.; Project administration, J.L. and Y.Z.; Funding acquisition, J.L. and G.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Zhejiang Provincial Natural Science Foundation of China under Grant LQ20F020006, the Zhejiang Provincial Philosophy and Social Science Foundation of China under Grant 22NDQN291YB, the Ningbo Natural Science Foundation under Grant 2023J280, the Ningbo Key R&D Program (Digital Twin Project), the program of Entrepreneurship and Innovation Ph.D. in Jiangsu Province (JSSCBS20211175) and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (21KJB520002).

Data Availability Statement: Multispectral pedestrian dataset is available online: https://github.com/soonminhwang/rgbt-ped-detection.

Acknowledgments: This work was partially supported by the Zhejiang Provincial Natural Science Foundation of China under Grant LQ20F020006, and the Zhejiang Provincial Philosophy and Social Science Foundation of China under Grant 22NDQN291YB. The authors would like to express their heartfelt gratitude to those people who have helped with this manuscript and to the reviewers for their comments on the manuscript.

Conflicts of Interest: It should be understood that none of the authors have any financial or scientific conflicts of interest with regard to the research described in this manuscript.

References

1. Lee, J.H.; Choi, J.S.; Jeon, E.S.; Kim, Y.G.; Thanh Le, T.; Shin, K.Y.; Lee, H.C.; Park, K.R. Robust pedestrian detection by combining visible and thermal infrared cameras. *Sensors* **2015**, *15*, 10580–10615. [CrossRef] [PubMed]
2. Zhao, L.; Yang, H.; Dong, L.; Zheng, L.; Asiya, M.; Zheng, F. MMFuse: A multi-scale infrared and visible images fusion algorithm based on morphological reconstruction and membership filtering. *IET Image Process.* **2023**, *17*, 1126–1148. [CrossRef]

3. Hao, L.; Li, Q.; Pan, W.; Yao, R.; Liu, S. Ice accretion thickness prediction using flash infrared thermal imaging and BP neural networks. *IET Image Process.* **2023**, *17*, 649–659. [CrossRef]
4. Balsa-Barreiro, J.; Menendez, M.; Morales, A.J. Scale, context, and heterogeneity: The complexity of the social space. *Sci. Rep.* **2022**, *12*, 9037. [CrossRef] [PubMed]
5. Balsa-Barreiro, J.; Valero-Mora, P.M.; Menéndez, M.; Mehmood, R. Extraction of naturalistic driving patterns with geographic information systems. *Mob. Netw. Appl.* **2020**, 1–17. [CrossRef]
6. Balsa-Barreiro, J.; Valero-Mora, P.M.; Berné-Valero, J.L.; Varela-García, F.A. GIS mapping of driving behavior based on naturalistic driving data. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 226. [CrossRef]
7. Yang, L.J.; Li, M.B.; Wu, T.X.; Bao, Y.F.; Li, J.H.; Jiang, Y. Geo-information mapping improves Canny edge detection method. *IET Image Process.* **2023**, *17*, 1893–1904. [CrossRef]
8. Zhang, Y.G.; Shen, L.Q.; Hu, H.M. Extraction of foreground area of pedestrian objects under thermal infrared video surveillance. *J. Beijing Univ. Aeronaut. Astronaut.* **2020**, *46*, 1721–1729.
9. 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; pp. 509–514.
10. 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]
11. Li, B.Y.; Liu, Y.; Wang, X.G. Gradient harmonized single-stage detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 8577–8584.
12. Zhang, Y.; Shen, L.; Wang, X.; Hu, H.M. Drone Video Object Detection using Convolutional Neural Networks with Time Domain Motion Features. In Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval, Shenzhen, China, 6–8 August 2020; pp. 153–156.
13. 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, 8–10 June 2015; pp. 1037–1045.
14. Wu, B.; Nevatia, R. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In Proceedings of the IEEE International Conference on Computer Vision, Beijing, China, 17–21 October 2005.
15. Wang, S.; Cheng, J.; Liu, H.; Tang, M. PCN: Part and Context Information for Pedestrian Detection with CNNs. In Proceedings of the British Machine Vision Conference, London, UK, 4–7 September 2017.
16. Wang, C.; Ning, X.; Sun, L.; Zhang, L.; Li, W.; Bai, X. Learning Discriminative Features by Covering Local Geometric Space for Point Cloud Analysis. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
17. Ning, X.; Tian, W.; Yu, Z.; Li, W.; Bai, X.; Wang, Y. HCFNN: High-order Coverage Function Neural Network for Image Classification. *Pattern Recognit.* **2022**, *131*, 108873. [CrossRef]
18. Stark, B.; Smith, B.; Chen, Y.Q. Survey of thermal infrared remote sensing for Unmanned Aerial Systems. In Proceedings of the 2014 International Conference on Unmanned Aircraft Systems (ICUAS), Orlando, FL, USA, 27–30 May 2014; pp. 1294–1299.
19. Sobrino, J.A.; Del Frate, F.; Drusch, M.; Jimenez-Munoz, J.C.; Manunta, P.; Regan, A. Review of thermal infrared applications and requirements for future high-resolution sensors. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 2963–2972. [CrossRef]
20. Parikh, D.; Zitnick, C.L.; Chen, T. Exploring tiny images: The roles of appearance and contextual information for machine and human object recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1978–1991. [CrossRef]
21. Lu, Y.; Li, W.; Ning, X.; Dong, X.; Zhang, L.; Sun, L.; Cheng, C. Blind image quality assessment based on the multiscale and dual-domains features fusion. *Concurr. Comput. Pract. Exp.* **2021**, *2021*, e6177. [CrossRef]
22. Torresan, H.; Turgeon, B.; Ibarra-Castanedo, C.; Hebert, P.; Maldague, X.P. Advanced surveillance systems: Combining video and thermal imagery for pedestrian detection. In Proceedings of the International Society for Optics and Photonics, Orlando, FL, USA, 12 April 2004; pp. 506–515.
23. Choi, E.J.; Park, D.J. Human detection using image fusion of thermal and visible image with new joint bilateral filter. In Proceedings of the International Conference on Computer Sciences and Convergence Information Technology, Seoul, Republic of Korea, 30 November–2 December 2010; pp. 882–885.
24. Zhou, Z.; Wang, B.; Li, S.; Dong, M. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters. *Inf. Fusion* **2016**, *30*, 15–26. [CrossRef]
25. Ding, Y.Y.; Xiao, J.; Yu, J.Y. Importance filtering for image retargeting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 89–96.
26. Li, S.T.; Kang, X.D.; Hu, J.W. Image fusion with guided filtering. *IEEE Trans. Image Process.* **2013**, *22*, 2864–2875. [PubMed]
27. Zhou, Z.; Dong, M.; Xie, X.; Gao, Z. Fusion of infrared and visible images for night-vision context enhancement. *Appl. Opt.* **2016**, *55*, 6480–6490. [CrossRef]
28. Zhang, L.; Li, W.; Yu, L.; Sun, L.; Dong, X.; Ning, X. GmFace: An explicit function for face image representation. *Displays* **2021**, *68*, 102022. [CrossRef]
29. Li, H.; Wu, X.J.; Kittler, J. MDLatLRR: A novel decomposition method for infrared and visible image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4733–4746. [CrossRef]
30. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]

31. He, K.M.; Sun, J. Fast Guided Filter. *arXiv* **2015**, arXiv:1505.00996.
32. He, K.; Rhemann, C.; Rother, C.; Tang, X.; Sun, J. A global sampling method for alpha matting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 2049–2056.
33. Wilson, T.A.; Rogers, S.K.; Kabrisky, M. Perceptual-based image fusion for hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **1997**, *35*, 1007–1017. [CrossRef]
34. Chen, H.; Varshney, P.K. A human perception inspired quality metric for image fusion based on regional information. *Inf. Fusion* **2007**, *8*, 193–207. [CrossRef]
35. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; p. 31.
36. He, K.M.; Sun, J.; Tang, X. Guided image filtering. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 1–14.
37. Zhang, L.; Sun, L.; Yu, L.; Dong, X.; Chen, J.; Cai, W.; Wang, C.; Ning, X. ARFace: Attention-aware and regularization for face recognition with reinforcement learning. *IEEE Trans. Biom. Behav. Identity Sci.* **2021**, *4*, 30–42. [CrossRef]
38. Wang, X.Y.; Han, T.X.; Yan, S.C. An HOG-LBP human detector with partial occlusion handling. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 32–39.
39. Khan, F.S.; Anwer, R.M.; Van De Weijer, J.; Bagdanov, A.D.; Vanrell, M.; Lopez, A.M. Color attributes for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3306–3313.
40. Li, S.; Sun, L.; Ning, X.; Shi, Y.; Dong, X. Head pose classification based on line portrait. In Proceedings of the 2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), Shenzhen, China, 9–11 May 2019; pp. 186–189.
41. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 2117–2125.
42. Li, D.W.; Xu, L.H.; Goodman, E.D. Illumination-robust foreground detection in a video surveillance system. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 1637–1650. [CrossRef]
43. Rajalingam, B.; Priya, R. Hybrid multimodality medical image fusion technique for feature enhancement in medical diagnosis. *Int. J. Eng. Sci. Invent.* **2018**, *2*, 52–60.
44. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.

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.