Journal of the Optical Society of America A
OPTICS, IMAGE SCIENCE, AND VISION

Multispectral image fusion based pedestrian detection using a multilayer fused deconvolutional single-shot detector

YUNFAN CHEN AND HYUNCHUL SHIN*

Division of Electrical Engineering, Hanyang University, Ansan 15588, South Korea
**Corresponding author: shin@hanyang.ac.kr*

Received 20 December 2019; revised 19 March 2020; accepted 22 March 2020; posted 23 March 2020 (Doc. ID 386410); published 23 April 2020

Recent research has demonstrated that effective fusion of multispectral images (visible and thermal images) enables robust pedestrian detection under various illumination conditions (e.g., daytime and nighttime). However, there are some open problems such as poor performance in small-sized pedestrian detection and high computational cost of multispectral information fusion. This paper proposes a multilayer fused deconvolutional single-shot detector that contains a two-stream convolutional module (TCM) and a multilayer fused deconvolutional module (MFDM). The TCM is used to extract convolutional features from multispectral input images. Then fusion blocks are incorporated into the MFDM to combine high-level features with rich semantic information and low-level features with detailed information to generate features with strong a representational power for small pedestrian instances. In addition, we fuse multispectral information at multiple deconvolutional layers in the MFDM via fusion blocks. This multilayer fusion strategy adaptively makes the most use of visible and thermal information. In addition, using fusion blocks for multilayer fusion can reduce the extra computational cost and redundant parameters. Empirical experiments show that the proposed approach achieves an 81.82% average precision (AP) on a new small-sized multispectral pedestrian dataset. The proposed method achieves the best performance on two well-known public multispectral datasets. On the KAIST multispectral pedestrian benchmark, for example, our method achieves a 97.36% AP and a 20 fps detection speed, which outperforms the state-of-the-art published method by 6.82% in AP and is three times faster in its detection speed.

https://doi.org/10.1364/JOSAA.386410

1. INTRODUCTION

Accurate pedestrian detection has attracted attention from researchers in the computer vision and image processing fields. The objective of pedestrian detection is to accurately locate the position of pedestrians from images captured in various real-world surveillance situations. Pedestrian detection provides important functions to boost many humancentric applications, such as intelligent robots, video monitoring, and intelligent transportation systems [1–3].

Although many studies in this area have been conducted in the past decade [4–20], most only consider detecting pedestrians at daytime using visible cameras. However, a visible camera relies on the lighting conditions of the surrounding environment, since it is ineffective under circumstances with poor illumination (i.e., nighttime). For safety and better driving, it is important to also achieve robust, reliable pedestrian detection at nighttime. To overcome the aforementioned limitations, multispectral detectors employing a fusion of thermal and visible images have been developed [21–29]. The pedestrians can be enhanced by thermal images from a visual spectrum background in an environment with poor illumination, which provides complementary information about the regions of interest. This approach facilitates the building of more robust pedestrian detectors in a variety of lighting conditions.

Most existing multispectral pedestrian detectors are built using a two-stage approach, like Faster R-CNN [30] and VGG16 [31]. Faster R-CNN is adopted as the main structure for anchor boxes and proposal-driven mechanisms. VGG16 [31] is used to extract features. Although satisfactory performance has been achieved on reasonable scale pedestrians using existing methods, the accuracy decreases significantly when applied to small-sized pedestrian detection because it is challenging to use anchor boxes to generate positive samples for small-sized pedestrians. In addition, there is another factor that makes it difficult for small-sized pedestrian detection. The spatial resolution of the feature map gradually decreases as the number of convolution layers increases. On the one hand, the low-level feature layers contain high spatial resolution but

lack semantic information, which is not conducive for small instance detection. On the other hand, the high-level feature layers have rich semantic information but low spatial resolution, which also result in detectors ignoring small-sized pedestrians. Furthermore, existing fusion strategies for integrating the multispectral information vary between early fusion, halfway fusion, or late fusion, depending on the fusion position in region proposal network (RPN) and the combination of the three types of fusion methods above. These fusion strategies are all based on two-stage approaches resulting in high computational costs and redundant parameters due to complex architectures for fusing visible and thermal subnetworks.

To solve the problems mentioned above, we developed what we believe is a novel multilayer fused deconvolutional single-shot detector (MFDSSD) for effective multispectral pedestrian detection, which consists of a two-stream convolutional module (TCM) and a multilayer fused deconvolutional module (MFDM). Different from two-stage detectors, one-stage detectors such as a single-shot detector (SSD) [32] and you only look once (YOLO) [33] eliminate the procedures of region proposal generation and feature resampling. The one-stage detectors encapsulate all operations in a single network, which significantly outperforms the two-stage detectors in detection speed. Considering the requirement for accurate real-time detection of practical applications, we propose a one-stage detector to detect pedestrians. Our method takes a pair of aligned visible and thermal images as the inputs to the TCM and adopts MFDM to enhance feature representation for small-sized pedestrians and to effectively fuse the multispectral information for pedestrian detection in various illumination conditions. We believe that the MFDSSD achieves noticeable improvement, accurately detecting small-sized pedestrians even when the input images are low resolution. It is also worth mentioning that our approach can process 20 frames per second (fps) on a single NVIDIA GeForce Titan X GPU, which almost meets the real-time requirement for autonomous driving applications. The contributions of this work are listed below.

1. We propose what we believe is a novel one-stage fusion network to fuse visible and thermal images for multispectral pedestrian detection, which can accurately detect pedestrians in real-time. Existing fusion networks for multispectral pedestrian detection are two-stage methods that suffer from the problem of high computational time and are inapplicable to real applications. The proposed network is composed of a two-stream convolutional module (TCM) and a multilayer fused deconvolutional module (MFDM). The TCM is used to extract convolutional features of input visible and thermal images. The MFDM is added after the TCM to improve the detection performance on small-sized pedestrians and fuse the multispectral information.
2. A multilayer fused deconvolutional module (MFDM) is proposed to effectively integrate the rich semantic features in high-level and high-resolution detailed features in low-level, which can enhance the feature maps with more detailed and semantically strong information for small-sized pedestrians.
3. We design a new fusion block that is incorporated into multiple deconvolutional layers in the MFDM. The proposed fusion block adaptively fuses the complementary characteristics of visible and thermal images without increasing the number of parameters and computation time. Most existing fusion methods adopted addition or concatenation operations, which are not adaptive and likely to lose important information.
4. We generated a new challenging multispectral pedestrian dataset with small-sized pedestrian instances to demonstrate the robustness of our method. The experimental results reveal that the average precision (AP) on the new multispectral pedestrian dataset is 81.82%. In addition, our method achieves the best detection performance on the well-known KAIST multispectral pedestrian detection benchmark dataset [21] and UTokyo multispectral dataset [39]. With the KAIST multispectral pedestrian benchmark dataset [21], the performance of our method in AP exceeds that of the state-of-the-art published method [29] by 6.82%. Furthermore, our method's detection speed is three times faster. This shows that our method significantly outperforms both in performance and speed.

The remaining part of this paper has four sections. Section 2 briefly introduces the previous related works. Section 3 explains our proposed MFDSSD in detail. Section 4 presents the experimental results and analysis. Finally, Section 5 summarizes our work and describes future work.

2. RELATED WORK

A. Pedestrian Detection Using Visible Images

Over the past decades, a large number of approaches have been proposed to improve the performance of pedestrian detection from visible images. Triggs *et al.* [4] designed a histogram of oriented gradient (HOG) features to describe pedestrians and applied a support vector machine (SVM) for classification. Dollar *et al.* [5] extended the HOG features to integral channel features (ICF) by infusing LUV color channels. On the basis of ICF, aggregated channel features (ACFs) [6] were developed by adding gradient magnitude channel features, in which the computational time can be reduced by reducing the number of channels. With the rapid development of CNN-based methods, the performance of pedestrian detection has been improved to a new stage. Yang *et al.* [7] introduced the convolutional channel features combined with a forest classifier for training a pedestrian detector. Tian *et al.* [8] proposed using multiple parts of a human body to train detectors, which was helpful to overcome the heavy occlusion problem. In [10], the different complexities of the features were considered, and a cascade learning method based on handcrafted and CNN features was presented to optimize detection accuracy. In [11] an RPN was adopted to produce pedestrian candidates, which was then classified by the boosted forest (BF). In [12], two subnetworks designed for large-scale pedestrians and small-scale pedestrians were trained simultaneously, which aims at alleviating the problem that detection accuracy decreases due to the large variance of pedestrian scales. In [9,13–15], semantic and detection tasks were jointly learned to enhance the feature discriminability of

pedestrians. To accurately detect dense pedestrians with occlusion, Wang *et al.* [16] and Zhang *et al.* [17] formulated two new loss functions for regression. Although the above-mentioned two-stage methods achieve satisfying accuracy, the computation time is high. On the contrary, one-stage detectors perform better in speed and show competitive accuracy. Liu *et al.* [18] designed an asymptotic localization fitting module to refine the anchor boxes in several steps that gradually improve detection results. Lin *et al.* [19] incorporated fine-grained features into CNN and proposed an attention strategy to identify pedestrians. Most recently, CSP [20] introduced an anchor-free algorithm through convolution to look for the central points and the central scale to detect pedestrians.

B. Pedestrian Detection Using Multispectral Images

Existing research works have proved that multispectral pedestrian detectors trained by visible and thermal images are more robust than detectors trained by visible images alone. In [21], ACF + T + THOG features, which is a combination of ACF from visible images, intensity channel features T from thermal images, and THOG features from thermal images, were designed to train the AdaBoost classifier for multispectral pedestrian detection. Wagner *et al.* [22] first applied the CNN to detect multispectral pedestrians. They presented two fusion strategies–early-fusion and late-fusion–that were decided by the fusion position in CNNs. FRCNN Halfway Fusion [23] was then proposed that integrated two-stream CNNs on the middle-level of FRCNN, which achieved better results. Based on FRCNN Halfway Fusion, Fusion RPN + BDT [24] used a boosted decision tree (BDT) for classification instead of the original downstream classifier in FRCNN. Chen *et al.* [25] proposed a multilayer fusion RPN in which a summation fusion strategy is introduced. Guan *et al.* [26] and Li *et al.* [27] presented adaptive weighting mechanisms to fuse the multispectral images more efficiently. In [28], a semantic segmentation task was infused into a multispectral fusion network to assist the pedestrian detection task. MSDS–RCNN [29] further improved the detection performance by adding an additional subnetwork to handle hard negatives. Although the above-mentioned studies have facilitated the development of multispectral pedestrian detection, the detection performance on small-sized pedestrians is poor, and the detection speed is slow. Therefore, it is necessary to develop more effective techniques to detect small-scale pedestrians by boosting the performance of multispectral pedestrian detection.

In this research, a new MFDSSD is developed, in which the deconvolutional module is infused to enhance the feature representation of small-sized pedestrians. It can overcome the problem of the shrinking resolution of feature maps and predict accurate locations of pedestrians at various scales, since the information related to small objects is retained, even in deep layers. In addition, multispectral information is effectively fused in multiple layers through fusion blocks. It can support pedestrian detection in various lighting conditions. As a result, we believe the proposed MFDSSD achieves satisfactory detection performance on small-sized pedestrians and outperforms well-known existing approaches.

3. PROPOSED MFDSSD MODEL

This section describes our method for multispectral pedestrian detection in detail. Figure 1 presents an overview of our MFDSSD framework, which consists of a two-stream convolutional module (TCM) and a multilayer fused deconvolutional module (MFDM). The input visible and thermal images are first processed by the TCM to generate a series of progressively smaller convolutional layers. Then, the Conv7-V and Conv7-T layers generated by TCM are integrated via a summation operation to produce the fused Conv7 layer. Based on the fused Conv7 layer, the MFDM produces a sequence of deconvolutional layers with a gradually increasing resolution that is combined with corresponding feature maps from the TCM through fusion blocks. Finally, the generated feature maps in MFDM are fed to the prediction stage for pedestrian classification and bounding box regression.

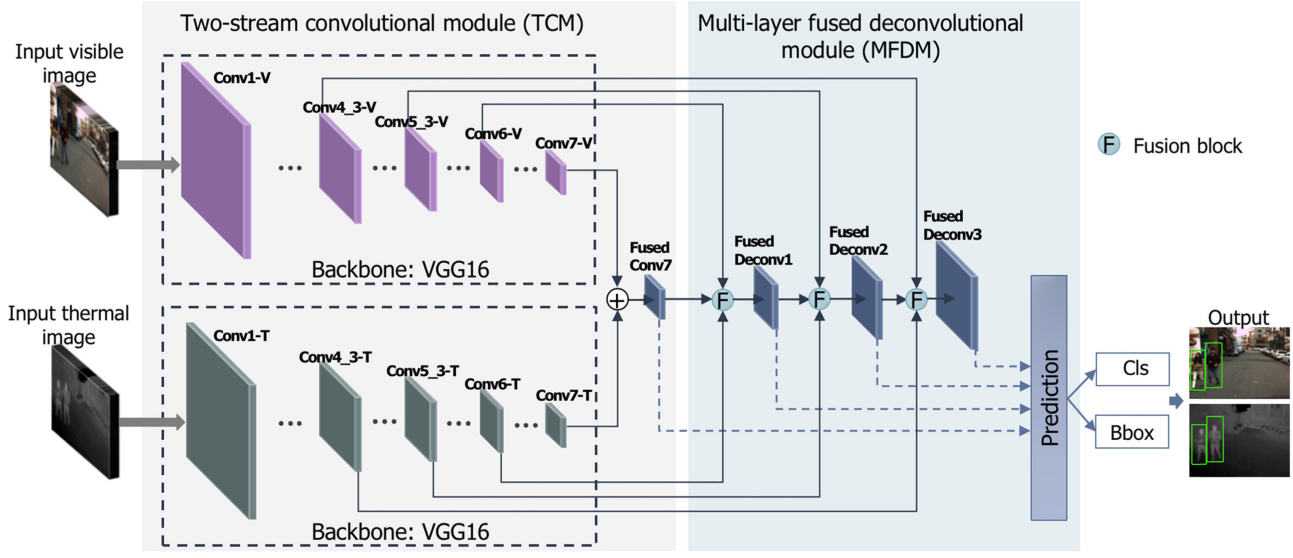

Fig. 1. Architecture of the proposed MFDSSD model.

A. Two-Stream Convolutional Module

The TCM aims to extract a series of progressively smaller convolution layers from the input visible and thermal images. As shown in Fig. 1, the TCM is built on top of two of the same backbone networks (a feedforward convolutional network) to extract convolutional features on a pair of pixel-aligned thermal and visible images. We choose VGG16 [31] as the backbone network because it performs well at classification and has a fast processing speed. However, a standard VGG16 cannot widely cover the scale range of pedestrians because its limited depth results in weak semantic information in feature maps. Despite these feature layers containing high spatial resolutions, it is still difficult to detect small pedestrian instances accurately. Therefore, we add two groups of additional convolutional layers (Conv6-V/Conv7-V and Conv6-T/Conv7-T) to the end of the truncated two-stream VGG-16 to increase semantic information at a high-level, which is conducive to detect multiscaled pedestrians, as proven in [34]. Thus, each feature extraction stream contains seven convolution and pooling layers (Conv1-V to Conv7-V from the visible stream and Conv1-T to Conv7-T from the thermal stream).

B. Multilayer Fused Deconvolutional Module

The MFDM has two functions: (1) to generate features with strong representational power for small pedestrian instances by integrating high-level features with rich semantic information and the low-level features with detailed information; and (2) to effectively fuse visible and infrared information from the TCM without increasing the computational time too much.

In the TCM, the low-level feature maps contain high spatial resolutions but lack semantic information. On the contrary, the high-level feature maps have rich semantic information but low spatial resolutions. Therefore, it is difficult to apply the feature layers of TCM to detect small objects. Fu *et al.* [35] introduced a deconvolutional network to overcome the shortcomings of the traditional downsampled convolutional network, which equalizes the represent ability of each layer and make the network more informative. Therefore, to enhance features for small pedestrian detection, we propose to add extra deconvolution layers to combine the complementary properties from the high-level and low-level features. We build the deconvolutional module at the end of the TCM module. As shown in Fig. 1, the MFDM starts from the fused Conv7 layer, which is generated via summation of the Conv7-V and Conv7-T layers. From that, a set of upsampled deconvolutional layers with gradually increasing resolution are generated. While five deconvolutional layers are used in [35], there are four layers used in our MFDM to save computation time.

C. Fusion Block

To bring rich context information as well as to effectively fuse multispectral information, a new fusion block is proposed to combine the feature layers in the MFDM and the corresponding feature layers in the TCM. There are three fusion blocks at different depths in Fig. 1. We take the first fusion block as an example here. Figure 2 shows an illustration for the fusion block. We apply (3 × 3 Conv, ReLU, and 3 × 3 Conv) on Conv6-V and Conv6-T separately and denote the outputs as F_V and F_T. The concatenation of the Conv6-V and Conv6-T is passed to a 1 × 1 Conv to generate the joint feature output F_C. F_D denotes the output of deconvolution on the Fused Conv7. We then perform an element-wise summation on F_V, F_T, F_C, and F_D. Finally, a block (ReLU, 3 × 3 convolution, ReLU) is applied to process the summed layer, which can help extract discernable features, as suggested by Zhang *et al.* [34]. The operations of the fusion block are summarized in

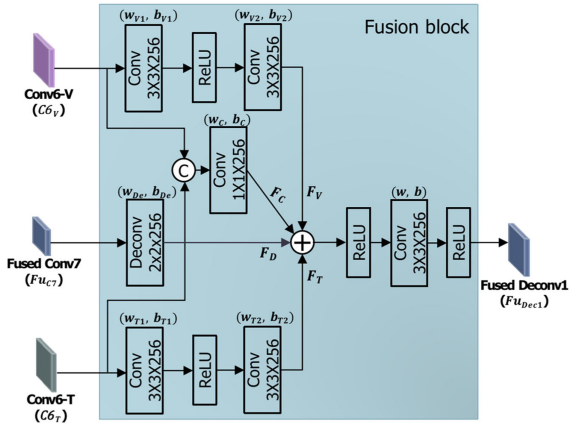

Fig. 2. Fusion block. Conv6-V ($C6_V$), Conv6-T ($C6_T$), and Fused Conv7 (Fu_C7) denote input visible, thermal, and fused features, respectively. Fused Deconv1 ($Fu_{\mathrm{Dec}}1$) denotes the fused output from the fusion block. © denotes concatenation, and ⊕ denotes element-wise summation.

$$F_V = (\mathrm{ReLU}\,(C6_V * w_{V1} + b_{V1})) * w_{V2} + b_{V2}$$

$$F_T = (\mathrm{ReLU}\,(C6_T * w_{T1} + b_{T1})) * w_{T2} + b_{T2}$$

$$F_C = (C6_V \odot C6_T) * w_C + b_C$$

$$F_D = Fu_{C7} * w_{De} + b_{De}$$

$$Fu_{\mathrm{Dec}}1 = \mathrm{ReLU}\,(F_V \oplus F_T \oplus F_C \oplus F_D) * w + b) \quad (1)$$

where
⊙ concatenation;
⊕: element-wise summation;
$C6_V$, $C6_T$, F_V, F_T, F_C, F_D, Fu_{C7}, Fu_{Dec1}: feature maps;
w_{V1}, w_{V2}, w_V, w_{T1}, w_{T2}, w_C, w_{De}, w: kernel weights; and
b_{V1}, b_{V2}, b_V, b_{T1}, b_{T2}, b_C, b_{De}, b: kernel bias.

D. Training

1. Matching and Hard Negative Mining

During training, we match the anchor box A to the ground truth box B using the Jaccard overlap, which is defined by

$$J(A, B) = \frac{\mathrm{area}(A \cap B)}{\mathrm{area}(A \cup B)}. \quad (2)$$

First, the anchor box with the highest $J(A, B)$ is matched to each ground truth. Then the rest of anchor boxes will be matched to any ground truth box if $J(A, B)$ exceeds 0.5. It is a conductive matching strategy to handle multiple predicted

bounding boxes with high scores while detecting overlapped pedestrians.

There are a majority of nonmatching anchor boxes after matching. The nonmatching anchor boxes that have the highest loss value will be selected as the negative samples to make the ratio between the positive and negative samples1:3.

2. Loss Function

The overall loss function of our method is shown in

$$L(\{x_i\}, \{t_i\}, \{p_i\}, \{b_i\}) = L_1(\{x_i\}, \{t_i\}) + L_2(\{p_i\}, \{b_i\}), \tag{3}$$

which includes two parts (i.e., the loss of ATM and the loss of MFDM), which are denoted by L_1 and L_2, respectively. Each part is a weighted summation of the two branches; one is the confidence loss (conf) of SoftMax classifier, and the other is the localization loss (loc) of the bounding box regression, as shown in

$$L_1(\{x_i\}, \{t_i\}) = \frac{1}{N_1}((L_{\text{conf1}}(x_i, x_i^*)) + \lambda L_{\text{loc1}}(t_i, t_i^*)), \tag{4}$$

$$L_2(\{p_i\}, \{b_i\}) = \frac{1}{N_2}((L_{\text{conf2}}(p_i, p_i^*)) + \lambda L_{\text{loc2}}(b_i, b_i^*)), \tag{5}$$

where x_i^*/p_i^* and t_i^*/b_i^* denote the ground truth label and location of an anchor i in a mini batch, respectively; x_i/p_i expresses the probability value that the anchor i is a pedestrian; t_i/b_i indicates the predicted location value of the anchor iI; and N_1 and N_2 note the number of positive anchors in the ATM and MFDM, respectively. Notably, if $N_1/N_2 = 0$, the loss is $L_1/L_2 = 0$ and λ is set to 1 through cross validation. The confidence loss $L_{\text{conf1}}/L_{\text{conf2}}$ is the cross-entropy loss over pedestrian class and nonpedestrian class. The localization loss is formulated by smooth L1 loss as in Fast R-CNN [36], and the smooth L1 loss is adopted as our regression loss L_{loc}, where smooth_{L1} is defined as

$$L_{\text{loc}}(t_i, t_i^*) = \sum_{j\in\{x,y,w,h\}} \text{smooth}_{L_1}(t_j, t_j^*). \tag{6}$$

3. Optimization

The VGG-16 [31] pre-trained on the ILSVRC dataset [37] is adopted as the backbone network in the proposed MFDSSD. The parameters of the additional layers of the MFDSSD are initialized by the "Xavier" approach [38]. We set the batch size to 5 for training. Then, we use stochastic gradient descent (SGD) with a momentum 0.9 and a weight decay 5×10^{-4} to fine-tune the entire network. We adopt a multiple learning rate training strategy to avoid gradient explosion. Specifically, 1×10^4 iterations are first running with a learning rate of 5×10^{-5}. Then, the next 7×10^4 iterations are running with a learning rate of 1×10^{-3}. After finishing 7×10^4 iterations, the learning rate is reduced by a factor of 10 after every 2×10^4 iterations. The total number of learning iterations is 1.2×10^5. The decreasing trend of the SoftMax loss through iteration times is shown in Fig. 3. We used the KAIST dataset as an example here.

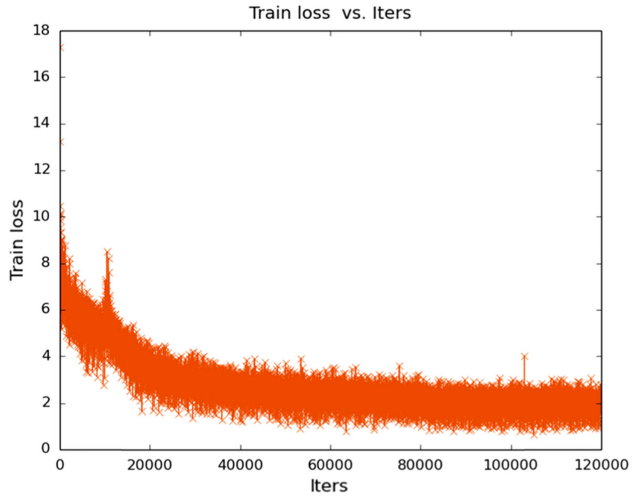

Fig. 3. Loss reduction with the number of iterations.

4. EXPERIMENTAL RESULTS

A. Datasets and Processing Platform

1. Proposed Multispectral Pedestrian Dataset

To validate the performance of the proposed approach, we built a challenging dataset for small-sized multispectral pedestrian detection, named Hanyang and Huins (HH). The multispectral pedestrian dataset contains pedestrian images with heights of 50 pixels or below, as shown in Fig. 4. The HH dataset consists of pixel-level aligned VI and IR images with ground truth labels. The pictures were taken using RGB and FIR dual cameras mounted on drones. In total, we collected 7247 pairs of images (each with a size of 720×480 pixels), including pedestrians with various sizes and postures, varying moving speeds, and partial or full occlusions. We divided the whole dataset into training and testing parts. The training part has 6247 pairs of images, and the testing part contains 1000 pairs of images. The ground truth contains bounding box coordinates and labels.

2. KAIST Multispectral Pedestrian Dataset

The KAIST dataset [21] was captured by a color camera and a thermal camera mounted on a moving vehicle during the day

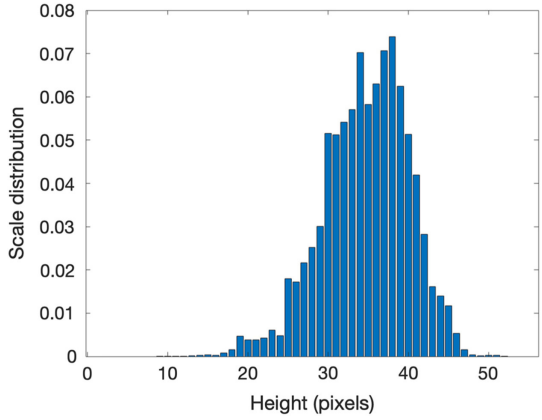

Fig. 4. Scale distribution of pedestrian heights from the HH dataset.

and night, which contains well-aligned visible and thermal images with the same size of 640 × 512 pixels. In particular, the dataset includes pedestrians with various scales, a variety of activities, and partial or heavy occlusions. In addition, it contains images with adverse lighting conditions, such as overexposure, shadows, dark night, dawn, or dusk. The training data contains 25086 pairs of visible and thermal images with two-frame skips. The total amount of test data is 2252 pairs of visible and thermal images with 20-frame skips, in which 1455 pairs were collected during the daytime, while 797 others were collected during the nighttime.

3. UTokyo Multispectral Object Detection Dataset

The UTokyo dataset [39] was captured at 1 fps using visible, far infrared, mid-infrared, and near-infrared cameras. It consists of 3740 group daytime images and 3772 group nighttime images. The dataset contains 1446 groups of aligned color-thermal images with five labeled classes (bike, car, car_stop, color_cone, person) and with the same size of 320 × 256 pixels. In our experiment, the 1446 groups of aligned color-thermal images are used for comparison of pedestrian (person) detection.

B. Evaluation Metrics

1. Precision and Recall

The precision-recall curve is widely used to evaluate the performance of object detectors. The overlap ratio between the predicted bounding boxes and ground truth boxes is measured to classify the detection results into three categories: true positive (TP), false positive (FP), and false negative (FN). The number of properly predicted pedestrians is denoted as TP. In general, the predicted result is judged as TP if the overlap ratio between the predicted bounding box and the ground truth is more than 0.5. The number of missing pedestrians is expressed by FN, and the number of nonpedestrian regions detected as pedestrians denoted as FP. Precision is defined as $TP/(TP + FP)$, and recall is defined as $TP/(TP + FN)$. The AP is computed via averaging several precision values at equally spaced recall levels by changing the threshold of the confidence scores. In this work, we had AP values at 100, and evenly spaced recall levels between 0 and 1 to obtain the AP.

2. Log-Average Miss Rate (MR)

We also use the log-average miss rate (MR) versus a false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$ to evaluate the detector performance [40]. A minimum overlap ratio of 0.5 is adopted to match the detected bounding box with the ground truth bounding box.

3. Multiple Object Detection Accuracy and Multiple Object Detection Precision

Multiple object detection accuracy (MODA) assesses the accuracy aspect of detector performance, which accounts for missed detections and false positives, and multiple object detection precision (MODP) assesses the localization precision of the detector performance [41]. We report MODA and MODP for radius $r = 0.5$ m, as suggested by Chavdarova *et al.* [42].

C. Detection Performance on the HH Dataset

This section evaluates the performance of our approach on the new HH dataset. Because the HH dataset is our proposed new benchmark, we only use our method for experiment. Fig. 5 shows the precision-recall curve and the MR-FPPI curve of our proposed MFDSSD, which reveals that our approach performs well when detecting small pedestrians, with an AP of 81.82% and an MR of 37.50%. The results make sense because the HH dataset contains plenty of far-scale pedestrians (those below 50 pixels in bounding-box height). Furthermore, some images contain occluded pedestrian instances.

Figure 6 provides a visualization of the detection results of the proposed method on the HH test set. We marked the range of pedestrian heights, which are from 25 to 42 pixels. The pedestrians with various scales are successfully detected by the proposed method, which demonstrates that the proposed method detects small-sized pedestrian instances well.

D. Detection Performance on the KAIST Dataset

In this section, the performance of MFDSSD is compared with a set of six well-known approaches, including ACF + T + THOG [21], FRCNN Halfway Fusion [23], Fusion RPN + BDT [24], MLF-CNN [25], IAF-RCNN [27], and MSDS-RCNN [29]. Comparison results are examined using the AP and log-average MR for pedestrian instances of three cases: (a) reasonable case (i.e., more than 55 pixels in

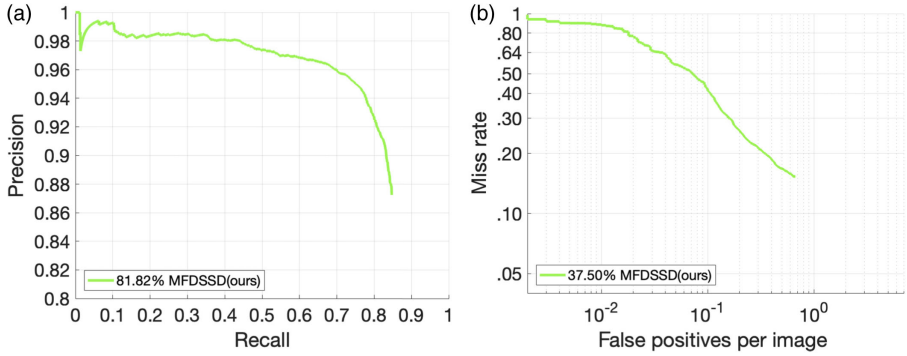

Fig. 5. Detection result on the HH test set. (a) Precision versus recall; (b) miss rate versus FPPI.

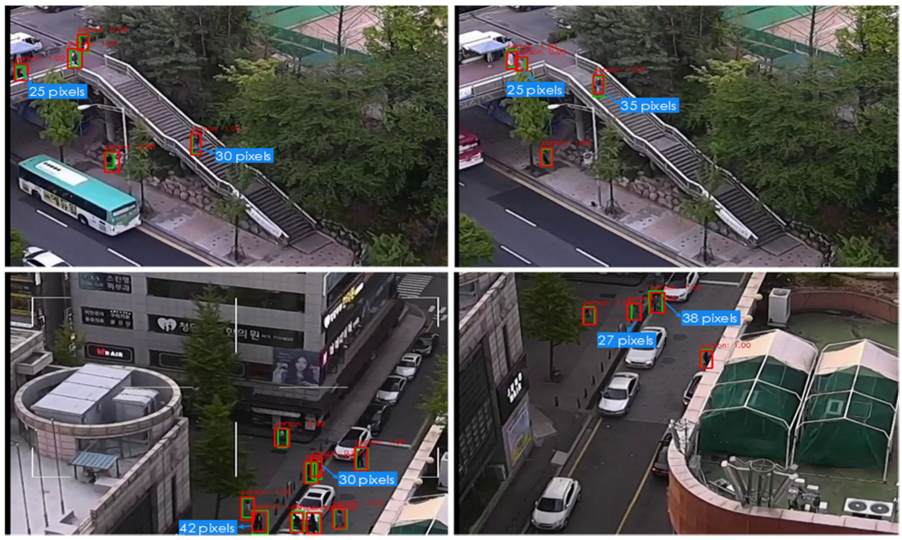

Fig. 6. Select detection results on the HH test set. The green bounding boxes denote the ground truth, and the red bounding boxes show the detection results.

height), (b) far-scale case (i.e., less than 55 pixels), and (c) overall case, which is a combination of both. Figures 7(a) and 7(d), respectively, present the precision-recall curves and MR-FPPI curves of a reasonable case. Our method clearly shows better performance than all other approaches and achieves the highest AP of 97.39%, which significantly outperforms the two recent well-known results, MSDS-RCNN [29] by 6.76%, and IAF-RCNN [27] by 10.17%. Our method achieves the lowest MR of 7.09%, which is smaller by 4.35% than that of the state-of-the-art method, MSDS-RCNN [29]. Furthermore, for the far-scale case, our approach showed significant improvement compared to the top competitive approach. Our method achieves the highest AP of 64.44% and the lowest MR of 43.35%, where the next best method (MSDS-RCNN [29]) has an AP of 46.22% (18.22% worse) and an MR of 57.85% (14.50% worse), as displayed in Figs. 7(b) and 7(e). Figs. 7(c) and 7(f) show the overall performance, so we can again observe a similar tendency like the results of the far-scale case: Our method evidently outperforms other methods. Since the majority of the KAIST dataset are far-scale instances, the overall detection accuracy can be improved by improving the small pedestrian detection.

We also evaluated detectors on two subsets in terms of different lighting conditions of the input images. The test sets were

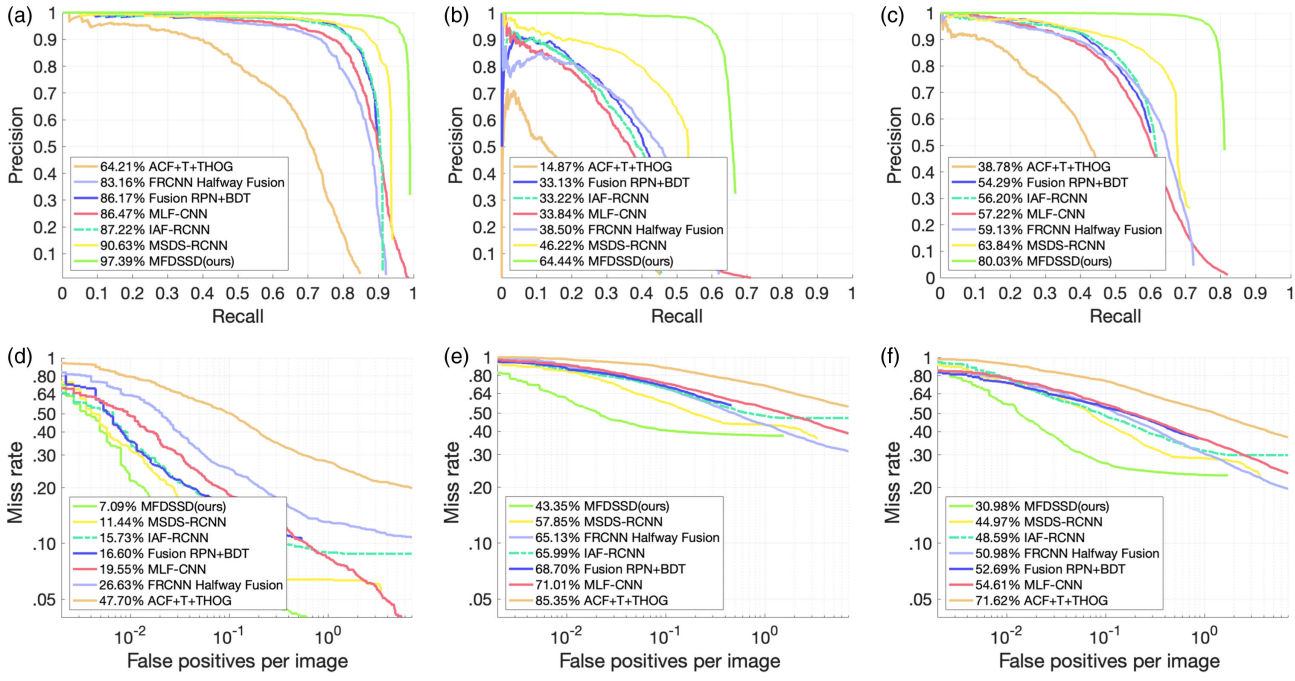

Fig. 7. Comparison of detection results (precision versus recall and miss rate versus FPPI) on the KAIST test set, in terms of different scales. (a) and (d) Reasonable scale (pedestrian height >= 55 pixels); (b)and (e) Far scale (pedestrian height < 55 pixels); and (c) and (f) Overall.

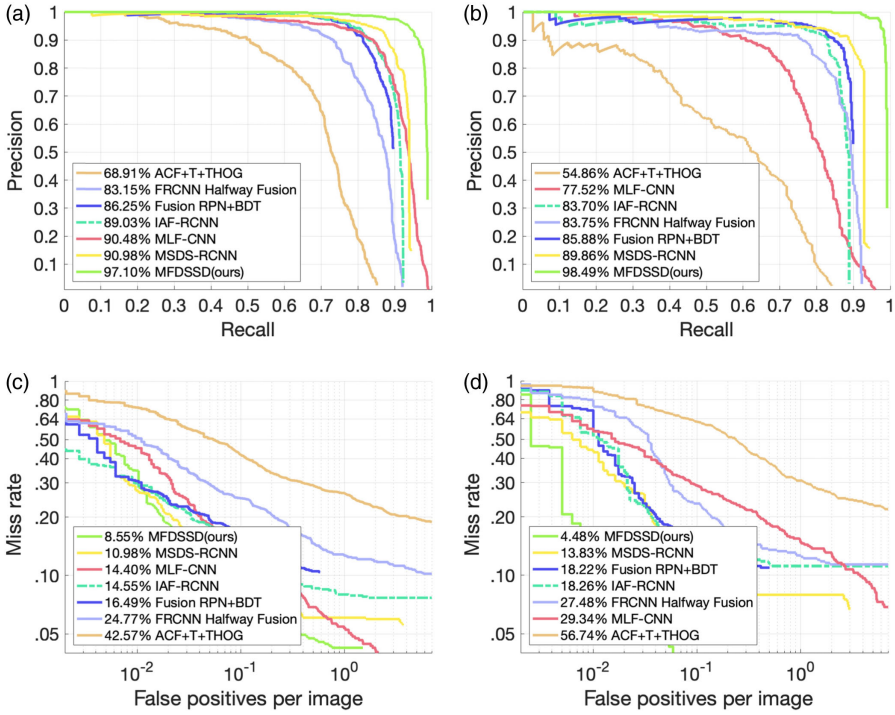

Fig. 8. Comparison of detection results (precision versus recall and miss rate versus FPPI) on the KAIST test set, in terms of daytime and nighttime. (a) and (c) Daytime. (b) and (d) Nighttime.

Table 1. Comparison of Detection Results on the KAIST Test Set Using the MODA, MODP, Precision, and Recall Metrics

Method	MODA (%)	MODP (%)	Precision (%)	Recall (%)
ACF + T + THOG [21]	23.15	49.24	64.21	52.30
FRCNN Halfway Fusion [23]	58.51	57.05	83.16	73.37
Fusion RPN + BDT [24]	70.01	60.82	86.17	83.40
MLF-CNN [25]	67.86	64.45	86.47	80.45
IAF-RCNN [27]	71.92	69.81	87.22	84.27
MSDS-RCNN [29]	79.40	70.22	90.63	88.56
MFDSSD (ours)	90.42	71.84	97.39	92.91

divided into daytime and nighttime. As shown in Figs. 8(a) and 8(c), for reasonable daytime, when compared with the state-of-the-art approach, MSDS-RCNN [29], our method outperforms by 6.12% in AP and 2.43% in MR. A similar tendency can be observed for the reasonable nighttime subset: the MSDS-RCNN [29] shows AP of 89.86% and MR of 13.83%, while our method shows significantly enhanced AP of 98.49% and reduced MR of 4.48%, as illustrated in Figs. 8(b) and 8(d).

The perfromance comparison of our approach with other approaches in terms of the MODA and MODP is presented in Table 1. As shown, our approach achieves significantly better performance and significantly improves the results of the MSDS-RCNN [29] method, from 79.40% to 90.42% in MODA and from 70.22% to 71.84% in MODP.

Fig. 9 presents comparisons of the detection results on four example images including challenging scenes at day and night, which contains pedestrians in low-visibility and at various scales. The visual comparisons evidently show that our method outperforms other state-of-the-art methods. The MSDS-RCNN [29], MLF-CNN [25], and FRCNN Halfway Fusion [23] methods produce many misses when the pedestrian size is too small or the illumination conditions are poor, while the baseline method ACF + T + THOG [21] generates many misses as well as false alarms. However, our method successfully detects pedestrians with various scales at day and night in all four examples. In conclusion, our approach works effectively to detect pedestrians at various scales and is robust under various illumination conditions.

E. Detection Performance on the UTokyo Dataset

We further verify the performance of our proposed method by using the UTokyo multispectral dataset. Fig. 10 displays the precision-recall curves and MR-FPPI curves of the overall case. Our method outperforms again the state-of-the-art method MSDS-RCNN [14] by 3.75% in AP and by 3.59% in MR.

The perfromance comparison of our approach with other approaches in terms of the MODA and MODP is presented in Table 2, which shows that our approach consistently achieves the best performance.

Figure 11 compares the detection results. Other methods produce considerably more false alarms and missing instances than our method. Because the UTokyo multispectral dataset involves the presence of many small-sized pedestrians in dark environments, this result demonstrate that our method works effectively to detect the small-sized instances by fusing the information of visible and thermal images.

Fig. 9. Visual comparison of our detection results in the KAIST test set with other approaches. The five rows show detection results of MFDSSD (ours) (see Visualization 1), MSDS-RCNN [29], MLF-RCNN [25], FRCNN Halfway Fusion [23], and ACF + T + THOG [21], respectively. The green bounding boxes denote the ground truth. The red bounding boxes show the detection results.

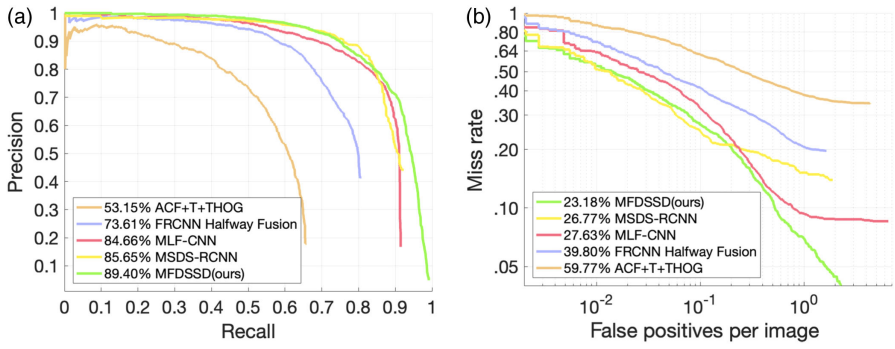

Fig. 10. Comparison of detection results on the UTokyo test set. (a) Precision versus recall curves; (b) miss rate versus FPPI curves.

F. Comprehensive Comparison in Terms of Detection Accuracy and Detection Speed

The computation time of the proposed method on 1000 HH test images is given in Fig. 12(a), which shows an average computation time of 0.05 s/f (20_fps). We also compare the computation times between our approach and other approaches on 2252 KAIST test images and 1446 UTokyo multispectral test images, as displayed in Figs. 12(b) and 12(c). All methods are tested on the same machine.

Table 3 gives a comprehensive comparison of the detection accuracy and the average computation time of 2252 KAIST test images. It is obvious that our method is faster than all other

Table 2. Comparison of Detection Results on the UTokyo Test Set Using the MODA, MODP, Precision, and Recall Metrics

Method	MODA (%)	MODP (%)	Precision (%)	Recall (%)
ACF + T + THOG [21]	4.77	46.91	53.15	40.23
FRCNN Halfway Fusion [23]	38.62	56.73	73.61	60.20
MLF-CNN [25]	59.26	64.55	84.66	72.37
MSDS-RCNN [29]	60.96	69.15	85.65	73.23
MFDSSD (ours)	67.71	70.94	89.40	76.82

methods, three times faster than the MSDS-RCNN method, and two times faster than MLF-CNN method. Our method outperforms all other methods in terms of detection accuracy under all evaluation conditions. Table 3 also compares the number of parameters of each method. It is clear that our method has far fewer parameters than other methods. For example, the number of parameters of our method is less than one tenth of that of the MSDS-RCNN. In summary, the proposed MFDSSD achieves the most accurate detection accuracy and is the fastest in detection speed. Furthermore, our method has the least number of parameters.

G. Ablation Studies

To verify the effectiveness of our proposed MFDSSD, we conduct ablation studies using the KAIST dataset. For the first simulation, we remove the multilayer fused deconvolutional module (MFDM) to keep only the two-stream convolutional module (TCM) for an experiment to see the performance. Then, the MFDM model without the fusion blocks is added to make an improved version. Specifically, we integrate high-level and low-level features by a simple addition operation. This simulation is to evaluate how the integration of high-level

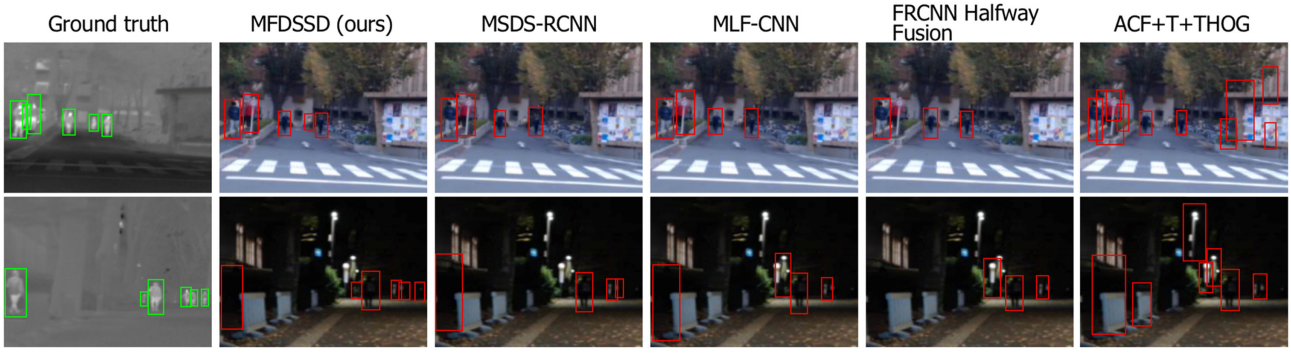

Fig. 11. Visual comparison of our detection results in the UTokyo test set with other approaches. The green bounding boxes denote the ground truth, illustrated in thermal images. The red bounding boxes show the detection results, displayed in visible images.

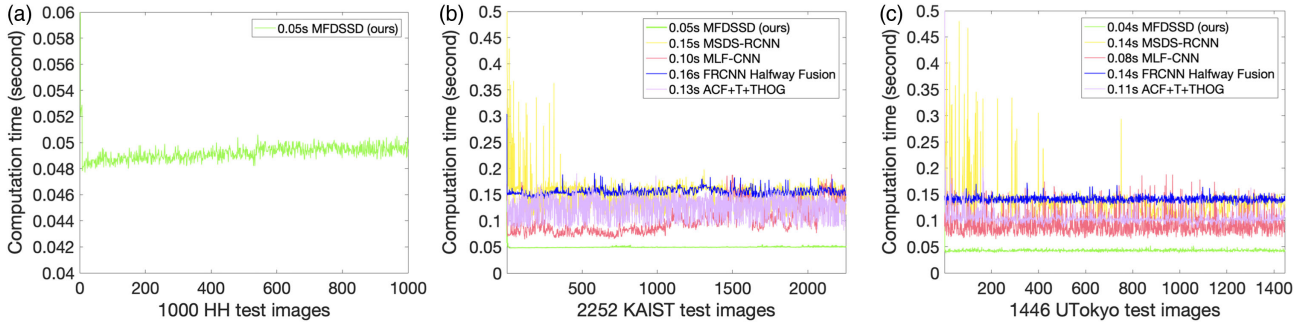

Fig. 12. Comparison of computation times. (a) HH test set; (b) KAIST test set; and (c) UTokyo test set.

Table 3. Comprehensive Comparison on KAIST Test Set

Methods	AP (%) in Terms of Different Scales			AP (%) in Terms of Different Lighting Conditions		Average Computation Time (s/f)	Number of Parameters (Mb)
	Reasonable Scale	Far Scale	Overall	Daytime	Nighttime		
ACF + T + THOG [21]	64.21	14.87	38.78	68.91	54.86	0.13	—
FRCNN Halfway Fusion [23]	83.16	38.50	59.13	83.15	83.75	0.16	579.4
MLF-CNN [25]	86.47	33.84	57.22	90.48	77.52	0.10	312.8
MSDS-RCNN [29]	90.63	46.22	63.84	90.98	89.86	0.15	3481.6
MFDSSD (ours)	**97.39**	**64.44**	**80.03**	**97.10**	**98.49**	**0.05**	**249.3**

Table 4. Results of the Ablation Experiments on KAIST Reasonable Test Set

Models	AP (%)	MR (%)
TCM	80.42	26.82
TCM + MFDM without Fusion blocks	89.54	14.25
TCM + MFDM + 1 Fusion block	92.35	11.70
TCM + MFDM + 2 Fusion blocks	95.08	9.53
TCM + MFDM + 3 Fusion blocks (MFDSSD)	97.39	7.09

and low-level features contribute to detection performance. Finally, we gradually infuse fusion blocks into the MFDM to replace simple addition to evaluate the effect of fusion blocks. Table 4 shows the average precision and log-average MR for each simulation.

1. *Evaluation of the Integration of High-Level and Low-Level Features*

As shown in Table 4, by adding the MFDM without fusion blocks to TCM, the detection performance significantly improved by 9.12% in AP and 12.57% in log-average MR. This comparison demonstrates that the integration of low-level and high-level features is effective to boost the detection performance. The TCM consists of two branches of feedforward convolution networks with a series of progressively smaller convolutional layers. It is hard for TCM to classify small-sized pedestrians. To overcome this issue, our method introduces the multilayer deconvolutional module to integrate low-level layers with high-resolution detailed features and high-level layers with rich semantic features, which is helpful to boost the pedestrian detection rate.

2. *Evaluation of the Fusion Blocks*

In Table 4 shows a comparison of the performance for the use of a diverse number of fusion blocks. One can observe that the performance improves with an increasing number of fusion blocks. The best performance is achieved when all three fusion blocks are added to MFDM. This proves that the proposed fusion block applied in multiple layers is effective to fuse information from VI and IR features and significantly improve detection performance. The addition operation to integrate the visible and infrared information is widely used by existing multispectral pedestrian detectors. However, the addition operation makes VI and IR features contribute equally, ignoring that the VI and IR images have different effects on pedestrian detection. The proposed fusion block automatically learns the weights to effectively integrate VI and IR features.

5. CONCLUSIONS AND FUTURE WORKS

We developed an effective approach that we call MFDSSD to detect pedestrians during the day and night in multispectral images. The MFDSSD consists of a TCM and a multilayer fused deconvolutional module MFDM. The TCM extracts the multispectral features from the input visible and thermal images, while the MFDM fuses the multispectral information and strengthens the feature representativity for small-sized pedestrian instances. A novel fusion block is proposed to be incorporated into MFDM, which is effective to integrate combine the low-level features with high spatial resolution and high-level features with rich semantic information to improve the detection accuracy of small-sized pedestrians and effectively fuse the multispectral information without increasing the computational cost. Experimental results on the new HH multispectral pedestrian dataset, KAIST multispectral pedestrian dataset, and UTokyo multispectral object detection dataset fully verify that our method achieves state-of-the-art detection accuracy with a fast computation speed, which is valuable in the area of pedestrian detection. In future work, we plan to further explore how to more effectively fuse multispectral information and extend our detector to a tracker by using multiple frames of video instead of a single image frame.

Funding. Ministry of Trade, Industry and Energy (10080619).

Disclosures. The authors declare no conflicts of interest.

REFERENCES

1. B. Li, Y. An, D. Cappelleri, J. Xu, and S. Zhang, "High-accuracy, high-speed 3D structured light imaging techniques and potential applications to intelligent robotics," Int. J. Intell. Robot. Appl. **1**, 86–103 (2017).
2. S. Kim, S. Kwak, and B. Ko, "Fast pedestrian detection in surveillance video based on soft target training of shallow random forest," IEEE Access **7**, 12415–12426 (2019).
3. L. Guo, P. Ge, M. Zhang, L. Li, and Y. Zhao, "Pedestrian detection for intelligent transportation systems combining AdaBoost algorithm and support vector machine," Exp. Syst. Appl. **39**, 4274–4286 (2012).
4. B. Triggs and N. Dalal, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2005), pp. 886–893.
5. P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. BMVC* (2009), pp. 91.1–91.11.
6. P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," IEEE Trans. Pattern Anal. Mach. Intell. **36**, 1532–1545 (2014).
7. B. Yang, J. Yan, Z. Lei, and S. Li, "Convolutional channel features," in *Proc. IEEE Int. Conf. Computer Vision*, Santiago, Chile (2015), pp. 82–90.
8. Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Computer Vision*, Santiago, Chile (2015), pp. 1904–1912.
9. Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA (2015), pp. 5079–5087.
10. Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proc. IEEE Int. Conf. Computer Vision*, Santiago, Chile (2015), pp. 3361–3369.
11. L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection," in *Proc. European Conf. Computer Vision*, Amsterdam, The Netherlands (2016), pp. 443–457.
12. J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," IEEE Trans. Multimedia **20**, 985–996 (2017).
13. G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *Proc. IEEE Int. Conf. Computer Vision* (2017).
14. X. Du, M. El-Khamy, J. Lee, and L. S. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *IEEE Winter Conference on Applications of Computer Vision (WACV)* (2017).

15. J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2017).
16. X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2018).
17. S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Li, "Occlusion-aware R-CNN: detecting pedestrians in a crowd," in *Proc. European Conf. Computer Vision* (2018).
18. W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. European Conf. Computer Vision* (2018).
19. C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. European Conf. Computer Vision* (2018).
20. W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: a new perspective for pedestrian detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2019).
21. S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: benchmark dataset and baseline," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA (2015), pp. 1037–1045.
22. J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *Proc. European Symp. Artificial Neural Networks*, Bruges, Belgium (2016), pp. 509–514.
23. J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proc. British Machine Vision Conf.*, York, UK (2016), pp. 1–13.
24. D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. IEEE Workshop on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017, pp. 243–250.
25. Y. Chen, X. Han, and H. Shin, "Multi-layer fusion techniques using a CNN for multispectral pedestrian detection," IET Comput. Vis. **12**, 1179–1187 (2018).
26. D. Guan, Y. Cao, J. Liang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," Inform. Fusion. **50**, 148–157 (2019).
27. C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," Pattern Recogn. **85**, 161–171 (2019).
28. D. Guan, Y. Cao, J. Yang, Y. Cao, and C. L. Tisse, "Exploiting fusion architectures for multispectral pedestrian detection and segmentation," Appl. Opt. **57**, D108–D116 (2018).
29. C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," arXiv:1808.04818 (2018).
30. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems* (2015).
31. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proc. Int. Conf. Learning Representations*, San Diego, CA, May 2015, pp. 1–14.
32. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Read, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. European Conf. Computer Vision* (Springer, 2016).
33. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2016).
34. S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2018).
35. C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," arXiv:1701.06659 (2017).
36. R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Computer Vision*, Santiago, Chile, December 2015, pp. 1440–1448.
37. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. C. Berg, "Imagenet large scale visual recognition challenge," Int. J. Comput. Vis. **115**, 211–252 (2015).
38. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. of the Thirteenth International Conference on Artificial Intelligence and statistics* (2010), pp. 249–256.
39. K. Takumi, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada, "Multispectral object detection for autonomous vehicles," *Proc. Thematic Workshops of ACM Multimedia*, Mountain View, CA, USA, October 2017, pp. 35–43.
40. P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," IEEE Trans. Pattern Anal. Mach. Intell. **34**, 743–761 (2012).
41. R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol," IEEE Trans. Pattern Anal. Mach. Intell. **31**, 319–336 (2009).
42. T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret, "WILDTRACK: A multi-camera HD dataset for dense unscripted pedestrian detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2018), pp. 5030–5039.