S & M 2330

Nighttime Pedestrian Detection Based on Thermal Imaging and Convolutional Neural Networks

Yung-Yao Chen,[1*] Guan-Yi Li,[2] Sin-Ye Jhong,[3]
Ping-Han Chen,[2] Chiung-Cheng Tsai,[2] and Po-Han Chen[2]

[1]Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology,
No. 43, Sec. 4, Keelung Rd., Taipei City 106, Taiwan (R.O.C.)
[2]Graduate Institute ofAutomation Technology, National Taipei University of Technology,
No. 1, Sec. 3, Zhongxiao E. Rd., Taipei City 106, Taiwan (R.O.C.)
[3]Department of Engineering Science, National Cheng Kung University,
No. 1, University Rd., Tainan City 701, Taiwan (R.O.C.)

(Received February 24, 2020; accepted June 3, 2020)

Keywords: nighttime pedestrian detection, convolutional neural network, thermal imaging

Pedestrian detection is a high-profile topic in computer vision, in part because it has great relevance to autonomous driving and intelligent surveillance applications. However, most pedestrian detection algorithms perform stably only during the daytime with sufficient illumination. At night, there is still room for improvement and many challenges exist. These challenges include occlusion caused by objects or crowds, and the problem of image background segmentation caused by environments with varying illumination. In this paper, we propose a nighttime thermal image pedestrian detection system, which can be viewed as an extension of the Faster region-based convolutional neural network (R-CNN) method. The proposed system can be used for static surveillance scenarios. First, a part model branch is proposed to realize the learning of partial pedestrian block features. Second, a segmentation branch is incorporated to strengthen the positioning of the pedestrian foreground. Finally, the branches are integrated through the fused loss function to enable joint training and optimization of the detection model. To evaluate the performance of the proposed model, we tested the system with several nighttime surveillance scenes. The experimental results show that the proposed method can effectively deal with the occlusion problem under challenging illumination environments and achieve performance levels superior to those of some state-of-the-art deep-learning pedestrian detection methods.

1. Introduction

With the rise of autonomous driving systems, intelligent surveillance, and smart robots, image processing topics involving people have attracted the attention of academia and industry.(1–4) Because human beings are one of the most important objects in any consideration, pedestrian detection is an important challenge in computer vision. Owing to the wide application of

[*] Corresponding author: e-mail: yungyaochen@mail.ntust.edu.tw

https://doi.org/10.18494/SAM.2020.2838

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

convolutional neural networks (CNNs), especially Faster region-based CNN (R-CNN)[5] and its variants,[6] the pedestrian detection performance has markedly improved over the past few years. However, most current pedestrian detectors are designed for daytime applications, i.e., under high illumination. In other words, they may fail at night or on foggy days because the quality of the captured color images degrades significantly under low illumination. The main reasons are the lack of stable natural light sources in these environments and the poor visibility of visible cameras at night.

Compared with daylight conditions, luminance under nighttime conditions is weaker, and many accidents occur at night due to poor visibility.[7] Therefore, night pedestrian detection systems are very important for related fields. In recent years, night pedestrian detection methods have been focused on by researchers. By using other types of sensors, such as near-infrared cameras, time-of-flight cameras, and long-wave infrared thermal cameras, the problem of the poor visibility of optical cameras at night (or under low lighting conditions) can be solved. Among these new types of imaging sensors, thermal cameras are the most widely used, because they can maintain high visibility of pedestrian characteristics at night even when there is insufficient light, as shown in Fig. 1.

However, nighttime pedestrian detection systems based on thermal images still have many challenges that must be addressed. First, because thermal images are generated by far-infrared sensing of the temperature of an object, in some bad weather, such as rain and snow, it is difficult to distinguish foreground objects from the background. Some techniques have been proposed to address this challenge. Among them, the pixel-level segmentation network is considered to be able to effectively deal with foreground and background segmentation problems. Brazil *et al.*[8] improved Faster R-CNN in a semantic segmentation network, where the feature fusion method was used to increase the segmentation feature information and improve the foreground and background segmentation performance of the system. However, although this type of method improves the pixel boundary segmentation between different classes of objects, it is difficult to segment objects of the same class, e.g., the problem of crowd

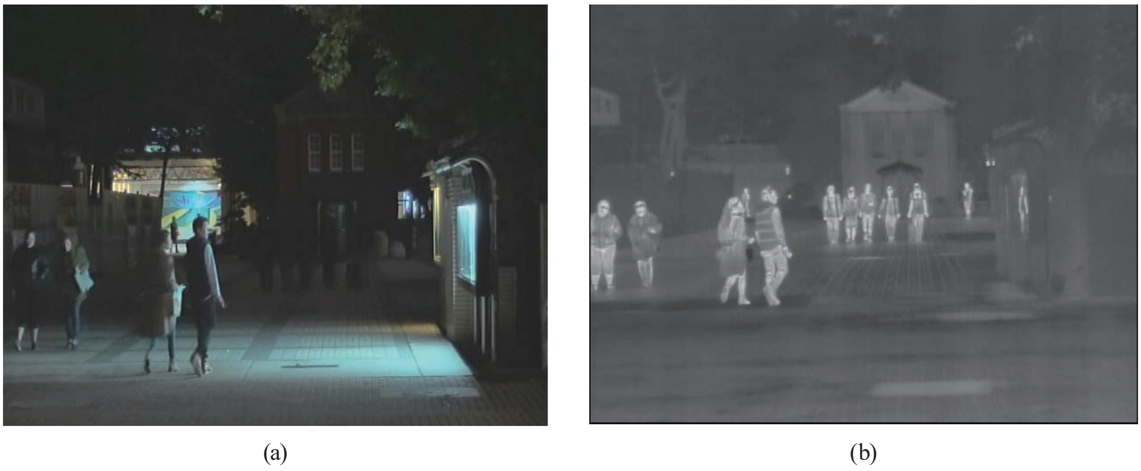

(a) (b)

Fig. 1. (Color online) Comparison between images from (a) a visible camera and (b) a thermal camera.

occlusion. In contrast, He *et al.*(9) proposed the Mask R-CNN method, which has a relatively stable performance in distinguishing similar classes of objects from each other. The Mask R-CNN method improves the original Faster R-CNN network by establishing a Mask branch, fusing it with multitask loss functions, and thus achieving better performance in detection tasks. Inspired by the Mask R-CNN method, in this study, we propose a semantic segmentation branch, which uses a newly defined fusion loss function for optimization. In addition to solving the problem of background and foreground segmentation, we also aim to improve the segmentation of objects from the same class and the performance of pedestrian detection systems at night.

Another challenge in pedestrian detection is occlusion. In many scenes, this problem is inevitable. In this work, the occlusion problem can be divided into two categories: (1) object occlusion: occlusion by other objects, such as a car, rock, or umbrella; (2) crowded occlusion: occlusion by a group of people. Either of these occlusion categories will lead to detection failure. Because some parts of a body are missed when occlusion occurs, the discriminability of a pedestrian detector will degrade significantly. In this case, global appearance features that are trained from the entire pedestrian are insufficient to detect a partially occluded body precisely. A technique called the deformable part model (DPM)(10) and its related methods have been proposed to deal with this problem. The DPM has an inherent advantage in handling occlusion because the detection process is separated into the detection of individual body parts by the histogram of oriented gradients (HOG) method, and therefore, the occluded parts can still be handled individually at the decision stage. Recently, Tian *et al.*(11) proposed the DeepParts method, which constructs a part pool consisting of 45 prototypes, and changed the original HOG feature extraction method to the CNN in the training stage. In this paper, we combine the part model with Faster R-CNN to solve the problem of occlusion through convolutional layers of different scales and custom block layers based on the concept of the part model.

To overcome the above-mentioned limitations of some related works, we present a new thermal pedestrian detection method. First, we integrate a part-model branch in the classification network stage, which divides the pedestrian features into nine custom layers based on the Faster R-CNN. Second, multiscale convolutional layers are used to obtain pedestrian features at different scales, and the Dropout layer is used to simulate random occlusion features as high-dimensional data to obtain the loss corresponding to the result of classification and the use of a bounding box. In addition, we propose a semantic segmentation branch, which is trained synchronously during the model training process. Finally, all the classification results, the bounding box results, and the segmentation losses are used to synchronize the optimization through the loss fusion procedure. In summary, the contributions of this study are as follows.

(1) We built a new static nighttime thermal pedestrian dataset, that is, we constructed a thermal pedestrian dataset by using a fixed thermal camera. Currently, most popular thermal databases, such as the Flir(12) and KAIST(13) datasets, are dynamic thermal pedestrian datasets. In the future, we plan to make this dataset publicly available to researchers.

(2) We proposed a thermal-based R-CNN model for nighttime pedestrian detection and simultaneously solved the problems of occlusion and background segmentation, which occur in nighttime pedestrian detection.

(3) We proposed a new loss fusion function that enables the models to be trained jointly, making the training process more efficient. From the experimental results, the proposed thermal R-CNN method outperforms state-of-the-art methods.

2. Proposed Thermal R-CNN for Nighttime Pedestrian Detection

As shown in Fig. 2, the proposed nighttime pedestrian detection method is based on the structure of Faster R-CNN, and the backbone network used is VGG-16. The reason why VGG-16 is used is because it has good network scalability, which enables the proposed branch network to be integrated into our architecture. First, through the region proposal network (RPN), the preliminary classification and bounding box regression results (of foreground objects and the background) are obtained. To achieve better detection accuracy, the original RoIPooling step is changed to the RoIAlign step for resizing. Then, we proposed the following extended branch network, which consists of three parts: part model head, segmentation head, and loss fusion.

2.1 Part model head

In view of the occlusion problem in pedestrian detection, we use a part model architecture, where the aspect ratio and block model are redefined, and the region discard block (RDB) is proposed to strengthen the network's generalization to occluded pedestrians. As shown in Fig. 3, first, we perform RoIAlign on the pedestrian feature map of the area obtained by the backbone network and RPN, and resize it to 3 × 6 × 512. Compared with the size used in the original Faster R-CNN (i.e., 7 × 7 × 512), the new size with a different aspect ratio more closely fits the properties of pedestrians and also helps the model to learn finer details. The features are then analyzed by two parts, a full body branch and a region decomposition branch.

Full body branch: This branch inherits the original Faster R-CNN, which flattens the original feature map of the whole pedestrian shape, and learns the depth features of the whole pedestrian

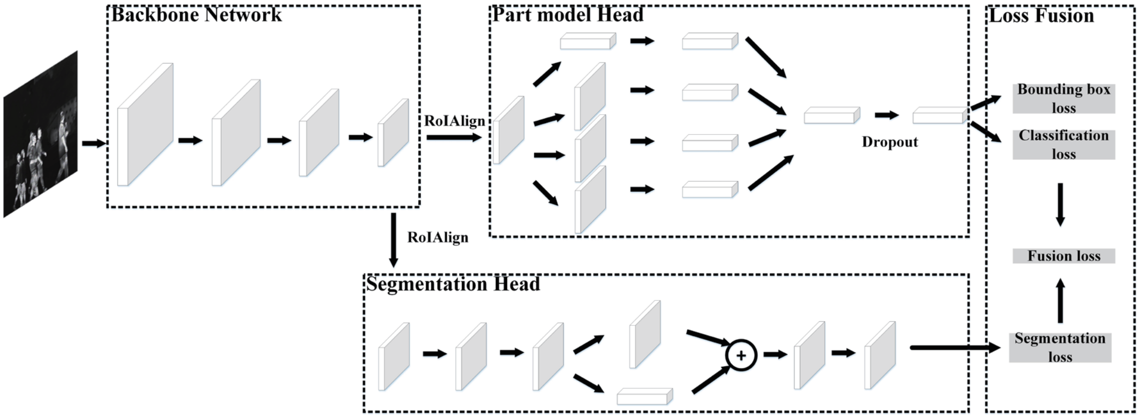


Fig. 2. Architecture of the proposed thermal R-CNN.

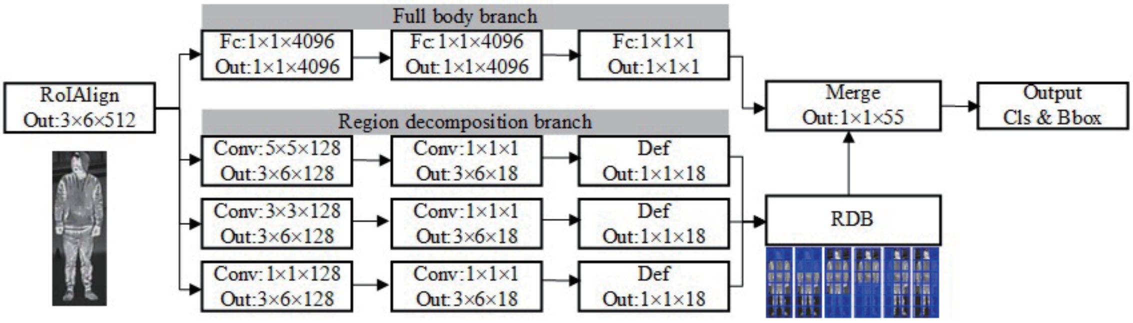


Fig. 3. (Color online) Detailed structure of the part model head.

in two fully connected (FC) layers of 4096 dimensions. The output 1 × 1 × 1 feature can be considered as the pedestrian information for the whole body.

Region decomposition branch: In this branch, we convolve the pedestrian feature maps with convolutional layers of different scales, including 5 × 5, 3 × 3, and 1 × 1. Among them, we add a padding procedure to the 5 × 5 convolution layer to maintain an aspect ratio consistent with the output feature map. The obtained features are divided into 54 parts of pedestrian features representing individual scales by using different deformation layers. Finally, the RDB is used to simulate different occlusion cases; the RDB is a block composed of the Merge layer and the Dropout layer. Inspired by the CutOut(14) method, the Dropout scheme is utilized during the training process, which randomly discards 54 neurons representing pedestrian features previously arranged with the Merge layer. Compared with training with occlusion and non-occlusion on the input layer, the proposed method can not only achieve the purpose of learning occlusion features without adding pedestrian data, but also avoids the confusion of the model caused by the direct training of two pedestrian features (occlusion and non-occlusion). Finally, we use the Merge layer to fuse the features obtained from the two branches, and obtain the scores from the full body branch and region decomposition branch for classification.

2.2 Segmentation head

At night, thermal-based pedestrian detection often encounters the problem of segmenting individual foreground objects in crowded scenes, which will degrade the performance of the RPN. To address this problem, we propose adding a segmentation branch network (Fig. 4) to incorporate the segmentation information when selecting the bounding box and distinguishing between foreground and background. The main path consists of four consecutive convolutional layers and one deconvolutional layer. Each convolutional layer is composed of 128 3 × 3 filters, and the deconvolutional layer up-samples features with a factor of 2. In addition, to increase the accuracy of the bounding box, we add FC layers of the main path, and to enhance the distinguishability between foreground and background, we design a short path dominated by the FC layer architecture after the third convolution layer. The previous convolution layer has a

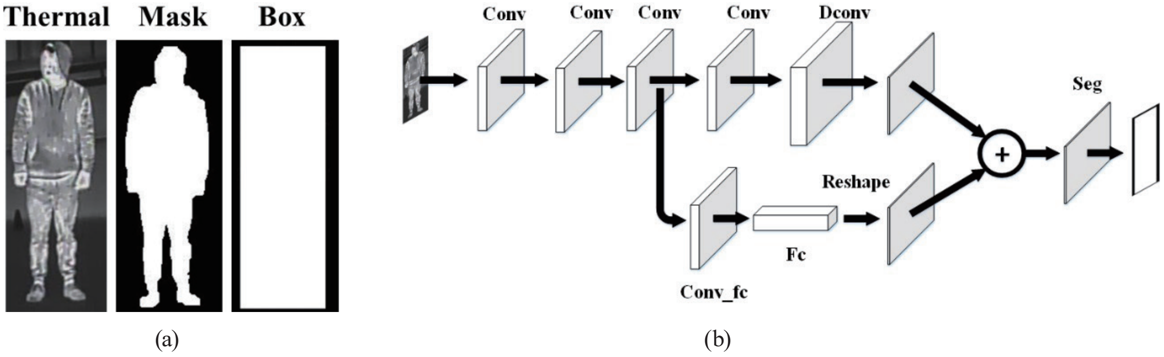


Fig. 4. Detailed structure of the segmentation head. (a) Visualization of the similarity between pixelwise segmentation masks (from Mask R-CNN) and weak box-based masks. (b) Architecture of the segmentation network.

size of 3 × 3 × 64, which is mainly to reduce the dimension of features to reduce the calculation time of subsequent FC layers. The mask size used in this study is 28 × 28. Finally, we connect the two path features in series and obtain weak box-based masks. Note that we use weak box-based masks, as shown in Fig. 4(a). In contrast to the pixelwise segmentation masks of Mask R-CNN, we do not need pixel-level annotation data, which makes the model easier to train, and the lightweight design makes the network easier to implement.

2.3 Loss fusion

To jointly train the models, the proposed thermal Faster RCNN is trained by minimizing the following joint loss function with five terms:

$$L = \lambda_1 L_{cls}^{rpn} + \lambda_2 L_{bbox}^{rpn} + \lambda_3 L_{cls}^{part} + \lambda_4 L_{bbox}^{part} + \lambda_5 L_{mask}^{seg} . \tag{1}$$

Considering the ratio rates, we set λ_1, λ_2, λ_3, and λ_4 to 1 and λ_5 to 2. In addition, L_{cls}^{rpn} and L_{bbox}^{rpn} represent the classification loss and the bounding box regression loss in the RPN, respectively. L_{cls}^{part} and L_{bbox}^{part} represent the classification loss and the bounding box regression loss in the part model head, respectively. By letting $GT_{i,j}$ and $P_{i,j}$ respectively represent the ground-truth and the predicted weak box-based masks, the image-level pixel loss can be expressed as

$$L_{mask}^{seg} = \frac{1}{S}\sum_{i,j} -\log[P_{i,j}^{*} P_{i,j} + (GT_{i,j} - P_{i,j}^{*})(GT_{i,j} - P_{i,j})] , \tag{2}$$

where S is the size of the feature map. Finally, we can perform simultaneous training through the fusion loss function. Note that we only use segmentation branch networks for training, and the branch network is used to optimize the overall detection performance.

3. Experiments

3.1 Nighttime thermal dataset

In this section, in order to effectively evaluate the feasibility and effectiveness of the proposed method, we constructed a complete nighttime pedestrian detection system, which included software implementations of original algorithms and hardware with a distinctive architecture. The hardware equipment includes a thermal camera, a computer, a frame grabber, and a camera tripod. The thermal camera used in this study had a length of 25 mm and its effective detection distance was approximately 10 to 60 m. The thermal camera was firmly mounted on a camera tripod so that we could simulate the actual situations of different surveillance scenarios for data collection and performance testing.

Although many public pedestrian databases exist, most of them are based on optical cameras and mobile platforms. As presented in Table 1, we collected a database using our thermal camera; the database has five different scenes with a resolution of 640 × 480; the dataset includes approximately 5000 images and depicts 8000 pedestrians. In addition, we defined challenging categories for different scenes as follows:

Scale: We used this data category to test whether the model retained a stable performance in an environment with considerable complexity of scale. We mined the dataset for statistics and used them to define the aspect ratio as 0.4 and to divide the data into three intervals based on length: near (10–30 m and ~100 pixels), medium (30–50 m and 50–100 pixels), and far (~50 m and ~50 pixels). Any video sequence that contained a pedestrian in all three intervals was considered to be in a scale-challenging category.

Clutter: For thermal imaging, different light sources and heat sources may degrade the quality of the captured images, which may cause the captured images to have muted foreground and background differences. If differences are muted, the algorithm has difficulty distinguishing between foreground and background objects. Therefore, we analyzed the background complexity of each scene and defined this type of challenge category.

Object occlusion and crowded occlusion: T occlusion is a crucial challenge in pedestrian detection. With reference to the Caltech Pedestrian Dataset,[15] which is one of the most widely

Table 1
Features of each video sequence.

Location	Resolution	Frames	Objects	Scale	Clutter	Object occlusion	Crowd occlusion
Video 1	640 × 480	640	471				
Video 2	640 × 480	770	845	✓			✓
Video 3	640 × 480	986	2611		✓	✓	
Video 4	640 × 480	470	259	✓	✓		✓
Video 5	640 × 480	2025	3574	✓	✓	✓	

used datasets for visible pedestrian detection, we marked each pedestrian with two boxes. As shown in Fig. 5(a), the red box indicates the visible region (BB-vis) and the green box indicates the complete shape of a pedestrian (BB-full). Given BB-vis and BB-full, the pedestrian occlusion rate can be calculated. When the occlusion rate is over 80%, the appearance of the pedestrian is almost lost, which indicates a completely occluded pedestrian. Therefore, in our dataset, we only considered pedestrians who had occlusion rates between 1 and 80%. In addition, we defined BB-full as the ground truth (GT), which is convenient and objective when comparing with other methods. In this paper, two types of occlusion are defined: object occlusion (i.e., occluded by an object such as an umbrella) and crowd occlusion (i.e., occluded by other people). Overall, the occlusion set includes 300 crowd occlusion cases and approximately 200 object occlusion cases. To refine the distribution of the actual occlusion, we quantized the bounding box to a 3 × 6 matrix. Figure 5(b) presents the statistics of the six types of occlusion that appeared most frequently in the dataset.

3.2 Evaluation of system performance

The extended model used in the experiment was based on the Faster R-CNN framework, which was implemented with Tensorflow. The network was initialized with the Glorot uniform initializer and was trained by the stochastic gradient descent (SGD) solver for 80000 iterations with a learning rate of 0.0001 and a batch size of 128. In addition, we did not use the data augmentation technique in the experiments. All experiments were performed on a machine with a single 2080Ti GPU and an Intel Core i7 8700 4.6 GHz CPU.

To confirm the feasibility of the proposed method, we adopted various metrics to evaluate the output of the pedestrian detection methods; the set of metrics included the number of true positives (TP), the number of false positives (FP), and the number of false negatives (FN). Using these quantities, the precision, recall, and F1-measure were calculated. As listed in Table 2, we analyzed the network architecture, including Modified Faster R-CNN (base), Base + Part

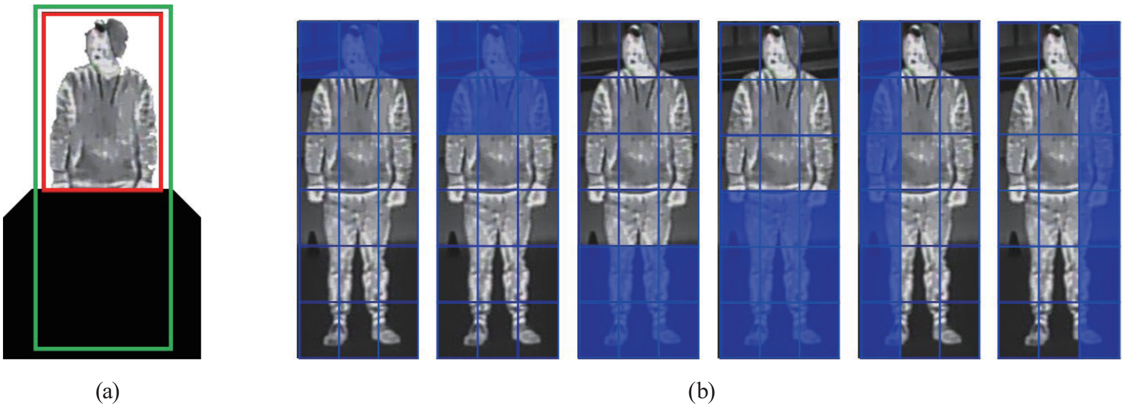


Fig. 5. (Color online) Definition of occlusion. (a) BB-full (green rectangle) and BB-vis (red rectangle). (b) Top six pedestrian occlusion types.

Table 2
False alarm and missed rate of detection.

Method	TP	FP	FN	Precision	Recall	F1-score
Modified Faster R-CNN	1069	752	918	0.587	0.538	0.561
Base + Part branch	1135	588	852	0.658	0.571	0.611
Base + Segmentation branch	1037	450	950	0.697	0.521	0.597
Thermal R-CNN (proposed)	1317	334	670	0.797	0.662	0.724

Table 3
Comparison of detection results of different methods.

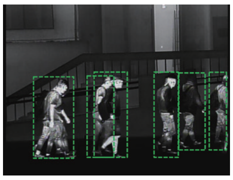
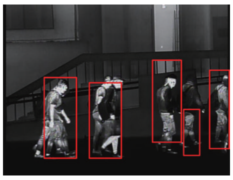
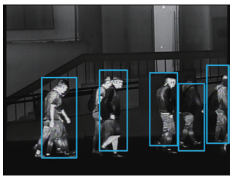
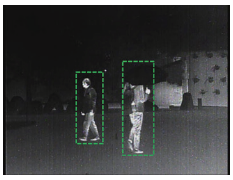
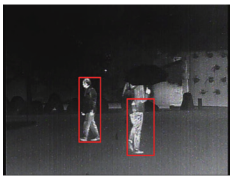
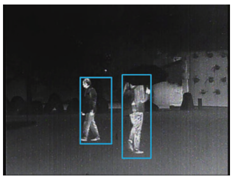
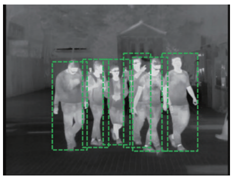
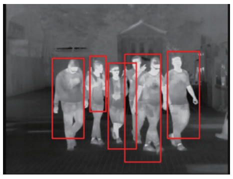
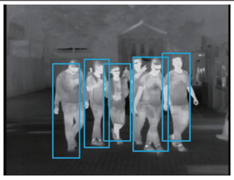
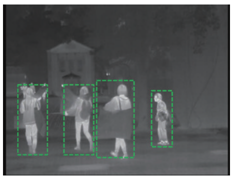
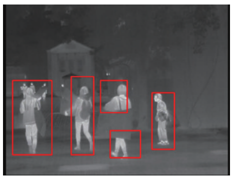
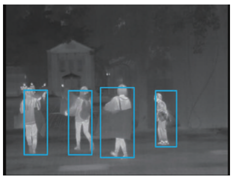
Method	TP	FP	FN	Precision	Recall	F1-score
Faster R-CNN	992	740	995	0.572	0.499	0.533
Mask R-CNN	1050	720	937	0.593	0.528	0.558
YOLOv3	1250	954	734	0.567	0.630	0.661
Thermal R-CNN (proposed)	1317	334	670	0.797	0.662	0.724

branch, Base + Segmentation branch, and Thermal R-CNN (proposed). From the experimental results, it can be observed that the expansion of Faster R-CNN on the basis of whether it is part branch or segmentation improves the performance. Base + Part branch has higher recall than Base + Segmentation. The main reason is that part branch is designed to manage multiple levels of scale and multiple pedestrian part features. It has high sensitivity to detailed and regional features, so it enables the system to distinguish numerous possibilities easily, but this ease tends to limit the precision because it is also relatively easy to misjudge images. However, Base + Segmentation uses the semantic segmentation feature to enhance the information of the complete pedestrian bounding box, so that the bounding boxes output by the network are more accurate and realistic, resulting in higher precision.

3.3 Comparison with state-of-the-art methods

Using the same benchmarks, we compared the proposed method with other published methods on the metrics mentioned in the previous section. As shown in Table 3, the methods used for comparison were Faster R-CNN, Mask R-CNN, and YOLOv3.(16) For the experiment, we used all the original parameters and code for data training and testing, where the number of training samples was 3261 and the number of test samples was 1630. The results revealed that our method achieved the highest performance. Note that the performance difference between Faster R-CNN and Mask R-CNN was small, mainly because the dataset used in this study did not contain complete semantic segmentation in labeling. That is, the label of semantic segmentation we used only included the segmentation features of the bounding box; therefore, it may degrade the superior performance of mask R-CNN segmentation. Table 4 shows several examples of detection results, which demonstrate the effectiveness of the proposed method.

Table 4
(Color online) Examples of false positives removed by thermal R-CNN fusion model.

Example	GT	Modified Faster R-CNN	Proposed
A			
B			
C			
D			

4. Conclusions

In this study, we provided two contributions to nighttime pedestrian detection. First, in the later stage of network classification, we added a part structure and proposed an RDB to allow the network to randomly learn about pedestrians. Masking was implemented to mitigate the decline in detection rate caused by occlusion. Second, we designed a segmentation branch based on weak box-based masks. This branch optimizes the network only through the training phase, allowing it to more accurately classify and locate pedestrians. This branch cannot be trained without pixel-level polygon data set annotations, but because it was used only in the training phase, it did not increase the performance burden during actual testing. Experimental results prove that the proposed thermal R-CNN delivered stable performance under occlusion and different illumination conditions. It is also superior to the state-of-the-art pedestrian detection methods on the collected static night pedestrian detection dataset. In the future, we will develop more datasets for indoor and outdoor surveillance scenarios.

References

1 C. Hsia: IEEE Sens. J. **18** (2018) 790. https://doi.org/10.1109/JSEN.2017.2772799
2 C. Hsia: J. Imaging Sci. Technol. **62** (2018) 30402. https://doi.org/10.2352/J.ImagingSci.Technol.2018.62.3.030402

3 C. Hsia, J. Guo, and C. Wu: Multimedia Tools Appl. **76** (2017) 25179. https://doi.org/10.1007/s11042-016-4296-z

4 C. Hsia: J. Internet Technol. **15** (2014) 1083. https://doi.org/10.6138/JIT.2014.15.7.01

5 S. Ren, K. He, R. Girshick, and J. Sun: IEEE Trans. Pattern Anal. Mach. Intell. **39** (2017) 1137. https://doi.org/10.1109/TPAMI.2016.2577031

6 J. Hosang, M. Omran, R. Benenson, and B. Schiele: Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2015) 4073–4082. https://doi.org/10.1109/CVPR.2015.7299034

7 J. Li, F. Zhang, L. Lisong Wei, T. Yang, and Z. Lu: Sensors **17** (2017) 2354. https://doi.org/10.3390/s17102354

8 G. Brazil, X. Yin, and X. Liu: Proc. 2017 IEEE Int. Conf. Computer Vision (IEEE, 2017) 4960–4969. https://doi.org/10.1109/ICCV.2017.530

9 K. He, G. Gkioxari, P. Dollár, and R. Girshick: Proc. 2017 IEEE Int. Conf. Computer Vision (IEEE, 2017) 2980–2988. https://doi.org/10.1109/ICCV.2017.322

10 R. Girshick, F. Iandola, T. Darrell, and J. Malik: Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2015) 437–446. https://doi.org/10.1109/CVPR.2015.7298641

11 Y. Tian, P. Luo, X. Wang, and X. Tang: Proc. 2015 IEEE Int. Conf. Computer Vision (IEEE, 2015) 1904–1912. https://doi.org/10.1109/ICCV.2015.221

12 FLIR Thermal Datasets: https://www.flir.com/oem/adas/dataset/ (accessed February 20).

13 S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon: Proc. 2015 IEEE Int. Conf. Computer Vision (IEEE, 2015) 1037–11045. https://doi.org/10.1109/CVPR.2015.7298706

14 T. DeVries and G. W. Taylor: arXiv:1708.04552 (2017).

15 P. Dollar, C. Wojek, B. Schiele, and P. Perona: IEEE Trans. Pattern Anal. Mach. Intell. **34** (2012) 743. https://doi.org/10.1109/TPAMI.2011.155

16 J. Redmon and A. Farhadi: arXiv: arXiv:1804.02767 (2018).