Information Fusion 50 (2019) 148–157

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full Length Article

Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection

Dayan Guan[a,b], Yanpeng Cao[a,b,*], Jiangxin Yang[a,b], Yanlong Cao[a,b], Michael Ying Yang[c]

[a] State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou, China
[b] Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, School of Mechanical Engineering, Zhejiang University, Hangzhou, China
[c] Scene Understanding Group, University of Twente, Hengelosestraat 99, Enschede, 7514 AE, The Netherlands

ARTICLE INFO

Keywords:
Multispectral fusion
Pedestrian detection
Deep neural networks
Illumination-aware
Semantic segmentation

ABSTRACT

Multispectral pedestrian detection has received extensive attention in recent years as a promising solution to facilitate robust human target detection for around-the-clock applications (*e.g.,* security surveillance and autonomous driving). In this paper, we demonstrate illumination information encoded in multispectral images can be utilized to boost the performance of pedestrian detection significantly. A novel illumination-aware weighting mechanism is present to depict illumination condition of a scene accurately. Such illumination information is incorporated into two-stream deep convolutional neural networks to learn multispectral human-related features under different illumination conditions (daytime and nighttime). Moreover, we utilized illumination information together with multispectral data to generate more accurate semantic segmentation which is used to supervise the training of pedestrian detector. Putting all of the pieces together, we present an effective framework for multispectral pedestrian detection based on multi-task learning of illumination-aware pedestrian detection and semantic segmentation. Our proposed method is trained end-to-end using a well-designed multi-task loss function and outperforms state-of-the-art approaches on KAIST multispectral pedestrian dataset.

1. Introduction

Pedestrian detection becomes a very important research topic within the area of computer vision in the past decades [4,5,9,11,12,31,44]. Given images captured in various real-world surveillance situations, the pedestrian detector should generate accurate bounding boxes to locate individual pedestrian targets. It provides an essential functionality to facilitate a board range of human-centric applications, such as video monitoring [1,26,39] and autonomous driving [25,40,42].

Although significant improvements have been accomplished during recent years, developing a robust pedestrian detector remains a challenging task. It is noticed that most existing pedestrian detectors are trained using visible information alone thus their performances are sensitive to changes of illumination, weather, and occlusions. To overcome the aforementioned limitations, many research works have been focused on the development of multispectral pedestrian detection solutions to enable accurate and robust human detection for around-the-clock application [14,17,22,23,30,37]. The underlying intuition is that multispectral images (*e.g.,* visible and thermal) contain complementary information of the targets, thus the effective fusion of such data can lead to more accurate and stable detections.

In this work, we present a framework for learning multispectral human-related characteristics under various illumination conditions (daytime and nighttime) through the proposed illumination-aware deep neural networks. We observed that multispectral pedestrian samples present different features under both day and night illumination conditions as illustrated in Fig. 1. Therefore, using multiple built-in subnetworks, each of which specializes in capturing illumination-specific visual patterns, provides an effective solution to handle substantial intra-class variances caused by illumination changes. Illumination information can be robustly estimated based on multispectral data and is further infused into multiple illumination-aware sub-networks to learn multispectral semantic feature maps for simultaneous pedestrian detection and semantic segmentation under different illumination conditions. Given a pair of multispectral images captured during the daytime, our proposed illumination-aware weighting mechanism adaptively assigns a high weight for day-illumination sub-networks (pedestrian detection and semantic segmentation) to learn human-related characteristics in the daytime. In comparison, multispectral images of a nighttime scene are utilized to train night-illumination sub-networks. We provide an illustration of how this illumination-aware weighting mechanism works in Fig. 2. Detections are generated by fusing the outputs of multiple illumination-aware sub-networks and remain robust to large variance of scene illumination. The contributions of this work are as follows.

Firstly, we demonstrate that illumination condition of a scene can be robustly determined through an architecture of fully connected

* Corresponding author.
E-mail address: caoyp@zju.edu.cn (Y. Cao).

https://doi.org/10.1016/j.inffus.2018.11.017
Received 1 January 2018; Received in revised form 12 November 2018; Accepted 28 November 2018
Available online 29 November 2018

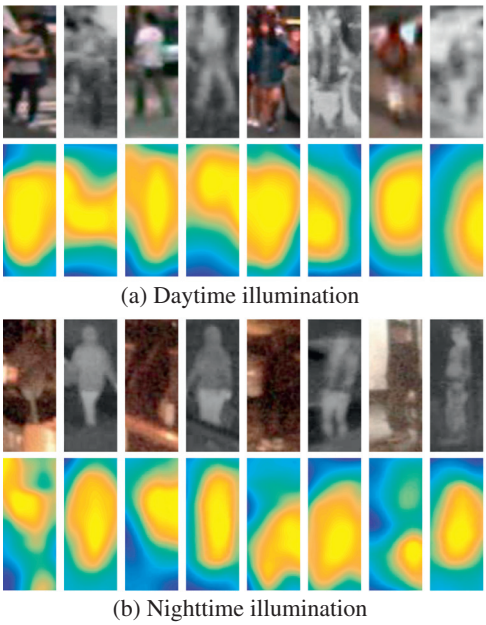


Fig. 1. Characteristics of multispectral pedestrian instances captured in (a) daytime and (b) nighttime scenes. The first rows in (a) and (b) show the multispectral picture of pedestrian instances. The second rows in (a) and (b) show the feature map visualizations of the corresponding pedestrian instances. The feature maps of visible and thermal images are generated using the deep neural region proposal networks (RPN) [41] well-trained in their corresponding channels. It is observed that multispectral pedestrian samples present different characteristics under day and night illumination conditions.

neural networks by considering multispectral semantic features and the estimated illumination weights provide useful information to boost the performance of pedestrian detection.

Secondly, we incorporate an illumination-aware mechanism into two-stream deep convolutional neural networks to learn multispectral human-related features under different illumination conditions (daytime and nighttime). To the best of our knowledge, we are the first to utilize illumination information for training multispectral pedestrian detector.

Thirdly, we present a complete framework for multispectral pedestrian detection based on joint learning of illumination-aware pedestrian detection and semantic segmentation which is trained end-to-end using a well-designed multi-task loss, achieving higher accuracy and faster runtime in comparison with the current state-of-the-art multispectral detectors [17,19,20].

2. Related work

In this section, we review a number of pedestrian detectors using visible, thermal and multispectral images, which are relevant to our research work.

Visible and thermal pedestrian detection: Many successful pedestrian detection solutions using visible images have been reported in the literature. Integrate Channel Features (ICF) pedestrian detector presented by Piotr et al. is based on feature pyramids and boosted classifiers [7]. Its performance has been further improved through multiple techniques including ACF [8], LDCF [29], and Checkerboards [43] etc. Recently, object detection models based on deep neural networks [13,15,34] have been used to improve the accuracy of pedestrian detection. Li et al. [24] proposed a unified deep network framework in which scale-aware sub-networks are combined to depict unique pedestrian features at different scales. Cai et al. presented a unified architecture of multi-scale deep neural networks to combine complementary scale-specific detectors. Such architecture provides a number of receptive fields to identify objects of different scales. Zhang et al. [41] made use of high-resolution convolutional feature maps for classification and presented a powerful detector for pedestrian detection based on region proposal networks (RPN) and boosted trees. Mao et al. [28] proposed a multi-task training framework, utilizing the information of given features to improve detection performance without extra inputs in inference. Brazil et al. [3] developed a segmentation infusion scheme to boost pedestrian detection accuracy with the joint supervision on target detection and semantic segmentation. Experimental results verified that the weakly annotated boxes provide sufficient information to achieve considerable performance gains. Davis et al. presented a template-based approach to localize pedestrians in thermal images captured in varying scenes. Potential persons are initially located using a generalized template and further verified through an AdaBoosted ensemble classifier [6]. Recently, multidimensional templates based on local steering kernel (LSK) descriptors were proposed by Biswas et al. for detecting pedestrians in low resolution and noisy infrared images [2]. However, strong solar radiation will cause background clutters and false detections in the daytime thermal images.

Multispectral pedestrian detection: Multispectral sensors (e.g., visible and thermal) capture information of target objects in complementary spectral channels. As a result, pedestrian detectors trained using multi-modal data produce robust detection results. Hwang et al. [17] built up a large-size multispectral pedestrian benchmark dataset

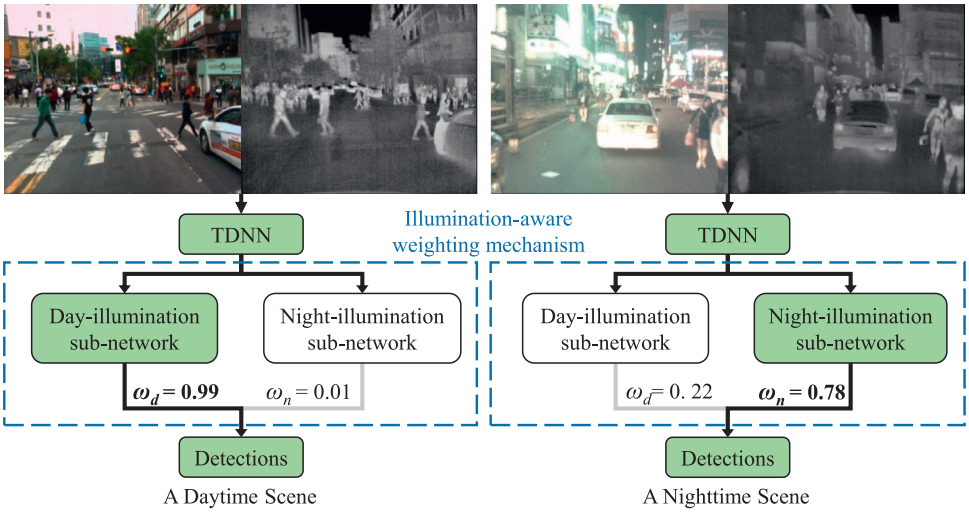


Fig. 2. Illustration of our proposed illumination-aware weighting mechanism. Given well-aligned multispectral images, two-stream deep neural networks (TDNN) generate multispectral semantic feature maps. Day-illumination sub-networks and night-illumination ones utilize the multispectral semantic feature maps for pedestrian detection and semantic segmentation under different illumination conditions. Detections are generated by fusing the outputs of day-illumination sub-networks and night-illumination ones with the computed illumination-aware weights ω_d and ω_n.

(KAIST), which contains well-aligned visible/thermal images and dense pedestrian annotations. The authors also presented a new technique to extract multispectral aggregated features (ACF + T + THOG) and applied boosted decision trees (BDT) for target classification. Wagner et al. [38] presented the first application of DNNs for multispectral pedestrian detection and evaluated the performance of two decision networks (early-fusion and late-fusion). Liu et al. [19] investigated how to utilize Faster R-CNN [34] for multispectral pedestrian detection and designed four ConvNet fusion architectures in which two-branch ConvNets are integrated at different DNNs stages. The optimal architecture is the Halfway Fusion model which merges two-branch ConvNets using the middle-level convolutional features. König et al. [20] presented an effective Fusion RPN + BDT model, in which two-stream deep neural networks are merged in the middle-level convolutional layers. Xu et al. proposed a new cross-modality transferring framework to learn the relations between color and thermal data and to improve the robustness of detectors against significant illumination changes. However, this multispectral pedestrian detector only considers visible images during the testing stage. Therefore, its performance is not comparable with multispectral detectors using both color and thermal data (e.g., Halfway Fusion model [19] and Fusion RPN + BDT [20]). Park et al. presented a three-branch DNN architecture, capable of handling multi-modal inputs [32]. A channel weighting fusion (CWF) layer is developed to improve the detection performance by considering all detection probabilities from each modality. Recently, Loveday et al. developed an orthogonal dual camera imaging system to capture parallax-free and well-aligned multispectral images [27]. It is experimentally shown that visible and infrared data fusion achieves improved overall performance of foreground object detection than using single-channel visible or infrared information.

It is worth mentioning that our approach is distinctly different from the above methods. A unified framework is proposed to learn multispectral human-related features under different illumination conditions (daytime and nighttime) through the proposed illumination-aware multispectral deep neural networks. To the best of our knowledge, this is the first research work exploring illumination information to boost the performance of multispectral pedestrian detector.

3. Our approach

3.1. Overview of the proposed model

As illustrated in Fig. 3, the architecture of illumination-aware multispectral deep neural networks consists of three integrated processing modules including illumination fully connected neural networks (IFCNN), illumination-aware two-stream deep convolutional neural networks (IATDNN), and illumination-aware multispectral semantic segmentation (IAMSS). Given aligned visible and thermal images, IFCNN computes the illumination-aware weights to determine whether it is a daytime scene or night one. Through the proposed illumination-aware mechanism, IATDNN and IAMSS make use of multi sub-networks to simultaneously generate classification scores (Cls), bounding boxes (Bbox), and segmentation masks (Seg). For instance, IATDNN employ two individual classification sub-networks (D-Cls and N-Cls) for human classification under day and night illuminations. Cls, Bbox, and Seg results of each sub-networks are integrated to obtain the final output through a gate function which is defined over the illumination condition of a scene. Our proposed method is trained end-to-end based on multi-task learning of illumination-aware pedestrian detection and semantic segmentation.

3.2. Illumination fully connected neural networks (IFCNN)

As shown in Fig. 3, a pair of visible and infrared images are passed into the first five convolutional layers and pooling ones of two-stream deep convolutional neural networks (TDNN) [20] to extract semantic features in each stream. Each stream of feature extraction layers in TDNN uses Conv1-5 from VGG-16 [36] as the backbone. Then feature maps from two channels are fused to generate the two-stream feature maps (TSFM) through a concatenate layer (Concat). TSFM is utilized as the input of IFCNN to compute illumination-aware weights ω_d and $\omega_n = (1 - \omega_d)$ which determine the illumination condition of a scene.

The IFCNN consist of a pooling layer (IA-Pool), three fully connected layers (IA-FC1, IA-FC2, IA-FC3), and the soft-max layer (Soft-max). Similar to the spatial pyramid pooling (SPP) layer which removes the fixed-size constraint of the network [16], IA-Pool resizes the features of TSFM to a fixed-length figure maps (7×7) using bilinear interpolation and generates fixed-size outputs for the fully connected layers. The number of channels in IA-FC1, IA-FC2, IA-FC3 are empirically set to 512, 64, 2 respectively. Soft-max is the final layer of IFCNN. The outputs of Soft-max are ω_d and ω_n. We define the illumination error term L_I as

$$L_I = -\hat{\omega}_d \cdot \log(\omega_d) - \hat{\omega}_n \cdot \log(\omega_n), \quad (1)$$

where ω_d and $\omega_n = (1 - \omega_d)$ are the estimated illumination weights for day and night scenes, $\hat{\omega}_d$ and $\hat{\omega}_n = (1 - \hat{\omega}_d)$ are the illumination labels. If the training images are captured under daytime illumination conditions, we set $\hat{\omega}_d = 1$, otherwise $\hat{\omega}_d = 0$.

3.3. Illumination-aware two-stream deep convolutional neural networks (IATDNN)

The architecture of IATDNN is designed based on the two-stream deep convolutional neural networks (TDNN) [20]. RPN model [41] is

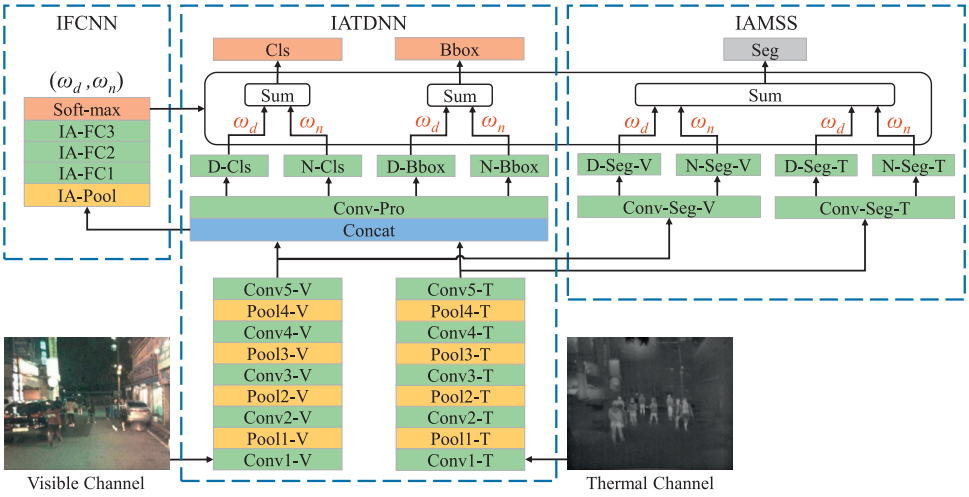


Fig. 3. The architecture of our proposed illumination-aware multispectral deep neural networks (IATDNN + IAMSS). Note that green boxes denote convolutional layers and fully-connected ones, yellow boxes denote pooling layers, blue boxes denote fusion layers, gray boxes denote segmentation layers, and orange boxes denote output layers. Should be viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

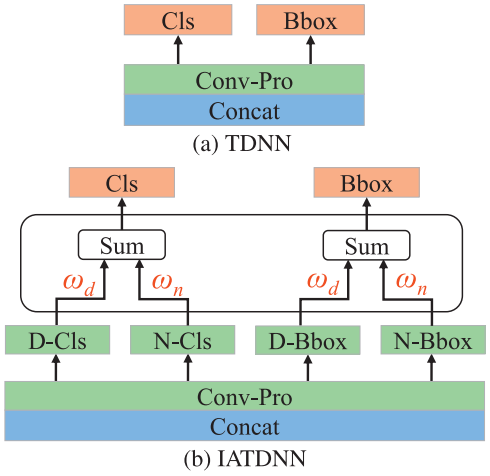


Fig. 4. The comparison of (a) TDNN and (b) IATDNN architectures. Note that ω_d and ω_n are the computed illumination-aware weights, green boxes denote convolutional layers and fully-connected ones, yellow boxes denote pooling layers, blue boxes denote fusion layers, and orange boxes denote output layers. Should be viewed in color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

adopted in IATDNN due to its superior performance for pedestrian detection. Given a single input image, RPN outputs a number of bounding boxes associated with confident scores to generate pedestrian proposals through classification and bounding box regression. As shown in Fig. 4(a), a 3×3 convolutional layer (Conv-Pro) is attached after the Concat layer with two sibling 1×1 convolutional layers (Cls and Bbox) for classification and bounding box regression respectively. TDNN model provides an effective framework to utilize TSFM for robust pedestrian detection.

We further incorporate illumination information into TDNN to generate classification and regression results for various illumination conditions. Specifically, IATDNN contains four sub-networks (D-Cls, N-Cls, D-Bbox, and N-Bbox) to produce illumination-aware detection results as shown in Fig. 4(b). D-Cls and N-Cls calculate classification scores under day and night illumination conditions while D-Bbox and N-Bbox generate bounding boxes for daytime and nighttime scenes respectively. The outputs of these sub-networks are combined using the illuminating weights calculated in IFCNN to produce final detection results. The detection loss term L_{DE} is defined as

$$L_D = \sum_{i\in S} L_{cls}(c_i^f, \hat{c}_i) + \lambda_{bb}\cdot\hat{c}_i\cdot\sum_{i\in S} L_{bbox}(b_i^f, \hat{b}_i), \tag{2}$$

where L_{DE} defines the sum of classification loss L_{cls} and regression loss L_{bbox}, λ_{bb} defines the regularization parameter between them (we set $\lambda_{bb}=5$ following previous work [41]), S defines the set of training samples in a mini-batch. A training sample is considered positive if its Intersection-over-Union (IoU) ratio with one ground truth bounding box is greater than 0.5, and otherwise negative. Here the training label $\hat{c}_i$ is set to 1 for positive samples and 0 for negative ones. For each positive sample, its bounding box is set to $\hat{b}_i$ for computing the bounding box regression loss. In Eq. (2), the classification loss term L_{cls} is defined as

$$L_{cls}(c_i^f, \hat{c}_i) = -\hat{c}_i\cdot\log(c_i^f) - (1-\hat{c}_i)\cdot\log(1-c_i^f), \tag{3}$$

and the regression loss term L_{bbox} is defined as

$$L_{bbox}(b_i^f, \hat{b}_i) = \sum smooth_{L_1}(b_{ij}^f, \hat{b}_{ij}) \tag{4}$$

where c_i^f and b_i^f are the predicted classification score and bounding box respectively, and the $L1$ loss function $smooth_{L_1}$ is defined in [13] to learn the transformation mapping between b_i^f and $\hat{b}_i^f$. In IATDNN, c_i^f is calculated as the weighted sum of day-illumination classification score c_i^d and night-illumination classification score c_i^n as

$$c_i^f = \omega_d\cdot c_i^d + \omega_n\cdot c_i^n, \tag{5}$$

and b_i^f is the illumination weighted combination of two bounding boxes b_i^d and b_i^n predicted by D-Bbox and N-Bbox sub-networks respectively as

$$b_i^f = \omega_d\cdot b_i^d + \omega_n\cdot b_i^n. \tag{6}$$

Through the above illumination weighting mechanism, the day-illumination sub-networks (classification and regression) will be given a high priority to learn human-related characteristics in daytime scenes. On the other hand, multispectral feature maps of a nighttime scene are utilized to generate reliable detection results under night-illumination conditions.

3.4. Illumination-aware multispectral semantic segmentation (IAMSS)

Recently, semantic segmentation masks have been successfully used as strong cues to improve the performance of single channel based object detection [3,15]. The simple box-based segmentation masks provide additional supervision to guide features in shared layers become more distinctive for the downstream pedestrian detector. In this paper, we incorporate the semantic segmentation scheme with two-stream deep convolutional neural networks to enable simultaneous pedestrian detection and segmentation on multispectral images.

Given information from two multispectral channels (visible and thermal), fusion at different stages (feature-stage and decision-stage) would lead to different segmentation results. Therefore, we hope to investigate what is the best fusion architecture for multispectral segmentation task. To this end, we design two multispectral semantic segmentation architectures that perform fusions at different stages, denoted as feature-stage multispectral semantic segmentation (MSS-F) and decision-stage multispectral semantic segmentation (MSS). As shown in Fig. 5(a) and (b), MSS-F firstly concatenates the feature maps from Conv5-V and Conv5-T and then applies a common Conv-Seg layer to produce segmentation masks. In comparison, MSS applies two convolutional layers (Conv-Seg-V and Conv-Seg-T) to produce different segmentation maps for individual channels and then combine two-stream outputs to generate the final segmentation masks.

Moreover, we hope to investigate whether the performance of semantic segmentation can be boosted by considering the illumination condition of the scene. Based on MSS-F and MSS architectures, we design two more illumination-aware multispectral semantic segmentation networks (IAMSS-F and IAMSS). Two segmentation sub-networks (D-Seg and N-seg) are employed to generate illumination-aware semantic segmentation results as shown in Fig. 5(c) and (d). Note that IAMSS-F contains two sub-networks and IAMSS contains four sub-networks. The outputs of these sub-networks are fused through the illumination weighting mechanism to generate the multispectral semantic segmentation using the illuminating weights predicted by IFCNN. In Section 4, we provide evaluation results of these four different multispectral segmentation architectures.

The segmentation loss term is defined as

$$L_S = \sum_{i\in C}\sum_{j\in S}[-\hat{s}_j\cdot\log(s_{ij}^f) - (1-\hat{s}_j)\cdot\log(1-s_{ij}^f)], \tag{7}$$

where s_{ij}^f defines the predicted segmentations, C defines segmentation streams (MSS-F and IAMSS-F contain only one segmentation stream while MSS and IAMSS contain two streams), and S defines the set of training samples in a mini-batch. Here the training segmentation mask $\hat{s}_i$ is set to 1 for positive samples and 0 for negative ones.

In illumination-aware multispectral semantic segmentation architectures IAMSS-F and IAMSS, s_{ij}^f is the illumination weighted combination

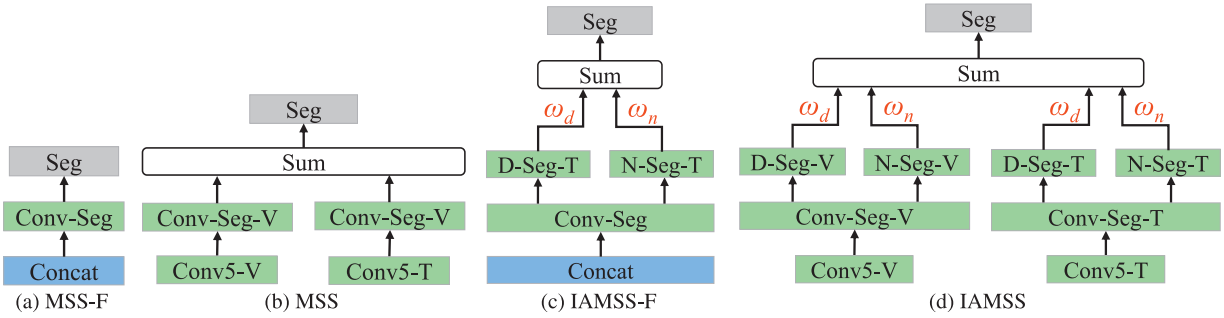


Fig. 5. The comparison of (a) MSS-F, (b) MSS, (c) IAMSS-F and (d) IAMSS architectures. Note that green boxes denote convolutional layers, blue boxes denote fusion layers, and gray boxes denote segmentation layers. Should be viewed in color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of two segmentation masks s_{ij}^d and s_{ij}^n predicted by D-Seg and N-Seg sub-networks respectively as

$$s_{ij}^f = \omega_d \cdot s_{ij}^d + \omega_n \cdot s_{ij}^n. \tag{8}$$

We integrate the loss terms defined in Eqs. (1), (2), (7) to conduct multi-task learning of illumination-aware pedestrian detection and segmentation. The multi-task loss function is defined as

$$L_{I+D+S} = L_D + \lambda_{ia} \cdot L_I + \lambda_{sm} \cdot L_S \tag{9}$$

where λ_{ia} and λ_{sm} are the trade-off coefficient of loss term L_I and L_S respectively. We set $\lambda_{ia} = 1$ and $\lambda_{sm} = 1$ according to the method presented by Brazil et al. [3]. We make use of this loss function to jointly train illumination-aware multispectral deep neural networks.

4. Experiments

4.1. Experimental setup

Datasets: The public KAIST multispectral pedestrian benchmark [17] is utilized to perform our experiments. The KAIST training dataset consists of 50,172 pairs of well-aligned multispectral images captured using visible and infrared cameras under different lighting conditions. Following the previous work [20], images are sampled every two frames in the training dataset, and 25,086 pairs of training images are obtained. The KAIST testing dataset consists of 2252 pairs of multispectral images, in which 1455 pairs were captured during the daytime. We evaluate the detection performance use the KAIST testing annotations following the reasonable setting introduced in [17]. It is noted that the CVC-14 [14] is another multispectral pedestrian benchmark containing visible-thermal image pairs. However, this multi-modal dataset was acquired using a stereo-vision system and the visible and thermal images are not properly aligned. Moreover, the annotations are individually generated in thermal and visible channels. Some pedestrian annotations are only generated in one channel but not available in another one. Therefore, we only make use of the KAIST dataset for performance evaluation in this paper.

Implementation details: We train all the multispectral pedestrian detectors using the image-centric training scheme [41]. Every mini-batch contains 1 image and 120 anchors, which are randomly selected. An anchor is considered as positive if its IoU ratio with one ground truth box is greater than 0.5, and otherwise negative. The first five convolutional layers in each stream of TDNN are initialized using parameters of the first five convolutional layers in VGG-16 [36] deep neural networks, which are pre-trained on the large-scale ImageNet dataset [35]. The fully connected layers and all the other convolutional ones are initialized with a zero-mean Gaussian distribution with standard deviation. See Table 1 for the detailed configurations of individual modules. The source code of our proposed model and the detection results will be made publicly available in the future. Deep neural networks are trained in the Caffe [18] framework with Stochastic Gradient Descent (SGD) [45] with a momentum of 0.9 and a weight decay of 0.0005 [21]. To avoid learning failures caused by exploding gradients [33], a threshold of 10 is used to clip the gradients.

Table 1
Configurations of our proposed IATDNN, IFCNN, and IAMSS modules. The input sizes of visible and thermal channels are both $960 \times 768 \times 3$. The first five convolutional layers in the visible and thermal streams (Conv1-V to Conv5-V and Conv1-T to Conv5-T) make use of Conv1-5 from VGG-16 [36] as the backbone.

Module	Layers	Output size	Operation
IATDNN	Conv5-V/T	$60 \times 48 \times 512$	
	Concat	$60 \times 48 \times 1024$	concatenation
	Conv-Pro	$60 \times 48 \times 512$	3×3 conv
	D/N-Cls	$60 \times 48 \times 18$	1×1 conv
	Cls	$60 \times 48 \times 18$	sum
	D/N-Reg	$60 \times 48 \times 36$	1×1 conv
	Reg	$60 \times 48 \times 36$	sum
IFCNN	IA-Pool	$7 \times 7 \times 512$	interpolation
	IA-FC1	512	inner product
	IA-FC2	64	inner product
	IA-FC3	2	inner product
IAMSS	Conv-Seg-V/T	$60 \times 48 \times 512$	3×3 conv
	D/N-Seg-V/T	$60 \times 48 \times 2$	1×1 conv
	D/N-Seg-V/T	$60 \times 48 \times 2$	sum

Evaluation metrics: The log-average miss rate (MR) [8] is utilized to evaluate the performance of various multispectral pedestrian detection algorithms. Following previous work [17], a detected bounding box result is counted as a true positive if the IoU with a ground truth one exceeds 50%. Unmatched detected bounding boxes are counted as false positives and unmatched ground truth ones are counted as false negatives. According to the method presented by Dollar et al. [8], any detected bounding box matched with any ignore ground truth label will not be considered as true positives and any unmatched ignore ground truth label will not be considered as false negatives. We compute the MR by averaging the missing rate calculated at nine false positives per image (FPPI) values evenly spaced in log-space from 10^{-2} to 10^{0} [17,19,20].

4.2. Evaluation on IFCNN

The illumination weighting mechanism provides an essential functionality in our proposed illumination-aware deep neural networks. We first evaluate whether IAFCNN can accurately calculate the illumination weights which provide critical information to balance outputs of illumination-aware sub-networks. We utilize the KAIST testing dataset, which contains multispectral images taken during daytime (1,455 frames) and nighttime (797 frames), to evaluate the performance

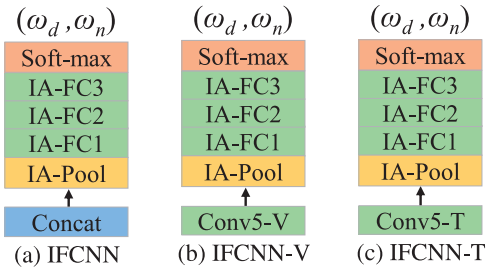


Fig. 6. The architecture of (a) IFCNN, (b) IFCNN-V and (c) IFCNN-T. Note that green boxes denote convolutional layers and fully connected ones, yellow boxes denote pooling layers, blue boxes denote fusion layers, and orange boxes denote soft-max layers. Should be viewed in color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Accuracy of illumination prediction using IFCNN-V, IFCNN-T, and IFCNN. The best results are highlighted in bold.

	Daytime	Nighttime
IFCNN-V	97.94%	97.11%
IFCNN-T	93.13%	94.48%
IFCNN	**98.35%**	**99.75%**

of IAFCNN. Given a pair of well-aligned multispectral images, IAFCNN will output a day-illumination weight ω_d. The illumination condition is correctly predicted if $\omega_d > 0.5$ for a daytime scene or $\omega_d < 0.5$ for a nighttime one. Moreover, we evaluate the performance of illumination prediction using feature maps extracted using visible images (IFCNN-V) or thermal ones (IFCNN-T) individually, to investigate which channel provides the most reliable information to determine illumination condition of a scene. The architectures of IFCNN-V, IFCNN-T, and IFCNN are shown in Fig. 6 and their prediction accuracies are compared in Table 2.

It is observed that information from the visible channel can be used to generate reliable illumination prediction for both daytime and nighttime scenes (daytime - 97.94% and nighttime - 97.11%). This is a reasonable result as a human can easily determine it is a daytime scene or a nighttime one based on visual observation. Although thermal channel cannot be individually used for illumination prediction, it provides supplementary information to the visible channel to enhance the performance of illumination prediction. Through the fusion of complementary information of visible and thermal channels, IFCNN compute more accurate illumination weights compared with IFCNN-V (using only visible images) and IFCNN-T (using only thermal images). Fig. 7 shows some cases when IFCNN fails. When the illumination condition is not good during daytime or street lights provide good illumination during nighttime, the IFCNN model will generate false prediction results. Overall, the illumination condition of a scene can be robustly determined based on IFCNN by considering multispectral semantic features.

Table 3
The calculated MRs of TDNN and IATDNN. The best results are highlighted in bold.

	All-day	Daytime	Nighttime
TDNN	32.60%	33.80%	30.53%
IATDNN	**29.62%**	**30.30%**	**26.88%**

4.3. Evaluation of IATDNN

We further evaluate whether illuminate information can be utilized to boost the performance of multispectral pedestrian detector. Specifically, we evaluate the performances of TDNN and IATDNN without using information of semantic segmentation. The illumination loss term described in Eq. (1) and detection loss term described in Eq. (2) are combined to jointly train IAFCNN and IATDNN, and use the detection loss term to train TDNN. TDNN model provides an effective framework to utilize multispectral features for robust pedestrian detection [20]. However, it does not differentiate human instances under day and night illumination conditions and uses a common Con-Prov layer to generate detection results. In comparison, IATDNN apply an illumination weighting mechanism to adaptively combine outputs of multiple illumination-aware sub-networks (D-Cls, N-Cls, D-Reg, N-Reg) to generate the final detection results.

MR is utilized as the evaluation metrics and the comparative results are shown in Table 3. Through the illumination weighting mechanism, IATDNN significantly improve detection accuracy for both daytime and nighttime scenes. It also worth mentioning that such performance gain (TDNN 32.60% MR v.s. IATDNN 29.62% MR) is achieved at the cost of small computational overhead. Based on a single Titan X GPU, TDNN model takes 0.22s to process a paired of visible and thermal images (640×512 pixels) while IATDNN model needs 0.24s. More comparative results of computational efficiency are provided in Section 4.5. The experimental results demonstrate that illumination information can be infused into multiple illumination-aware sub-networks for better learning of human-related feature maps to boost the performance of pedestrian detector.

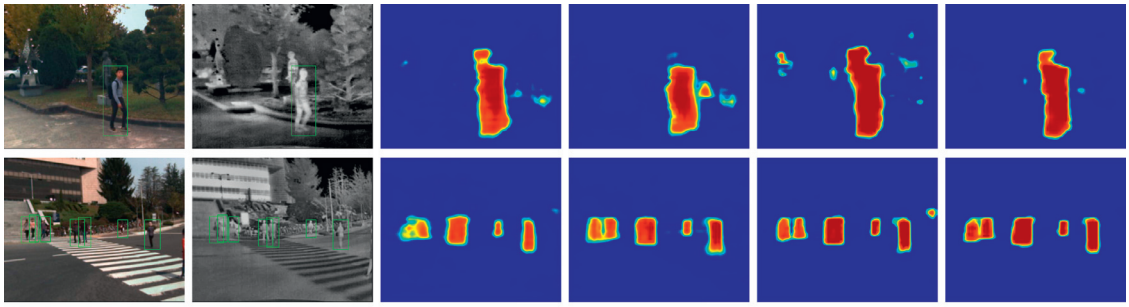
4.4. Evaluation of IAMSS

The performance gain of combining illumination-aware multispectral segmentation scheme with IATDNN is further evaluated. Here we consider four different multispectral semantic segmentation models including MSS-F (feature-stage MSS), MSS (decision-stage MSS), IAMSS-F (illumination-aware feature-stage MSS) and IAMSS (illumination-aware decision-stage MSS). Architectures of these four models are shown in Fig. 5. Multispectral semantic segmentation models output a number of box-based segmentation masks, and such weakly annotated boxes provide useful information to enable the training of more distinctive features in IATDNN. The detection performance of IATDNN, IATDNN + MSS-F, IATDNN + MSS and IATDNN + IAMSS-F and IATDNN + IAMSS are compared in Table 4.

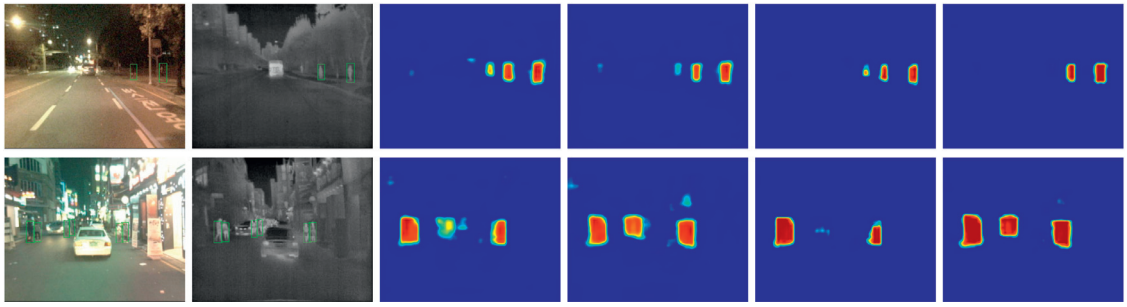
It is observed that integrating the semantic segmentation module with the illumination-aware pedestrian detection can generally achieve

(a) Daytime (b) Nighttime

Fig. 7. Samples of false IFCNN prediction results during (a) daytime and (b) nighttime. When the illumination condition is not good during daytime or street lights provide good illumination during the nighttime, the IFCNN model will generate false prediction results.



(a) Daytime



(a) Nighttime

Fig. 8. Examples of multispectral pedestrian semantic segmentation results using four different multispectral sematic segmentation models in (a) daytime and (b) nighttime scenes. The first two columns show the pictures of visible and thermal pedestrian instances respectively. The third to the sixth columns show the semantic segmentation results of MSS-F, MSS, IAMSS-F, and IAMSS respectively. It should be noted that green bounding boxes (BBs) in solid line denote positive labels, yellow BBs in dashed line denote ignore ones. Best viewed in color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Detection results (MR) of IATDNN, IATDNN + MSS-F, IATDNN + MSS, IATDNN + IAMSS-F, and IATDNN + IAMS. The best results are highlighted in bold.

	All-day	Daytime	Nighttime
IATDNN	29.62%	30.30%	26.88%
IATDNN+MSS-F	29.17%	29.92%	26.96%
IATDNN+MSS	27.21%	27.56%	25.57%
IATDNN+IAMSS-F	28.51%	28.98%	27.52%
IATDNN+IAMSS	**26.37%**	**27.29%**	**24.41%**

better detection performance using all four different multispectral semantic segmentation models (except using IATDNN + MSS-F for nighttime scenes). The underlying principle is that semantic segmentation masks provide additional supervision to facilitate the training of more sophisticated features to enable more robust detection [3]. Another important observation is that the choice of fusion scheme (feature-stage or decision-stage) will significantly affect detection performance. Overall, decision-stage multispectral semantic segmentation models (MSS and IA-MSS) perform much better the feature-stage models (MSS-F and IA-MSS-F). One possible explanation of this phenomenon is that late stage fusion (e.g., decision-stage fusion) is a more suitable strategy to combine high-level segmentation results. Finding the optimal segmentation fusion strategy to process multispectral data will be our future research. Last but not least, the performance of semantic segmentation can be further boosted by considering the illumination condition of the scene. Outputs of sub-networks are adaptively fused through the illumination weighting mechanism to generate more accurate segmentation results under various illumination conditions. Fig. 8 shows comparative semantic segmentation results using four different MSS models. It is observed that semantic segmentation masks generated by IATDNN + IAMSS more accurately cover small targets and suppress the background noise. More accurate segmentation results can provide better supervision to train more distinctive human-related feature maps.

In Fig. 9 we visualize the feature map of TDNN, IATDNN, and IATDNN + IAMSS to illustrate improvements gains achieved by different illumination-aware modules. We find that IATDNN generate more distinctive pedestrian features than TDNN by incorporating illumination information into multiple illumination-aware sub-networks for better learning of human-related feature maps. IATDNN + IAMSS can achieve further improvements through the segmentation infusion scheme in which illumination-aware semantic segmentation masks are used to supervise the training of feature maps.

4.5. Comparison with the current state-of-the-art multispectral pedestrian detection methods

Our proposed IATDNN and IATDNN + IAMSS are compared with three multispectral pedestrian detectors including ACF + T + THOG [17], Halfway Fusion [19] and Fusion RPN + BDT [20]. For performance comparison, we plot MR against FPPI (using log-log plots) by varying the threshold on detection confidence. As shown in Fig. 10, our proposed IATDNN + IAMSS achieves the best detection accuracy (26.37% MR) in all-day scenes, which is 11% lower than the second best performing solution Fusion RPN + BDT (29.68% MR). Furthermore, our proposed IATDNN, without incorporating the semantic segmentation architecture, can also achieve performance comparable to the state-of-art method. We visualize some detection results of the Fusion RPN + BDT and our proposed IATDNN and IATDNN + IAMSS in Fig. 11. It is observed that IATDNN and IATDNN + IAMSS both generate more robust detection results under varying illumination conditions, while IATDNN + IAMSS further reduce the false positives through the supervision of illumination-aware semantic segmentation. As illustrated in Fig. 11, IATDNN + IAMSS can even successfully predicted pedestrian instances which are unlabeled in the KAIST testing dataset. These correctly detected targets are considered as false positive detections. In our future work, we plan to restore these missed labels to facilitate better evaluation of multispectral pedestrian detection approaches.

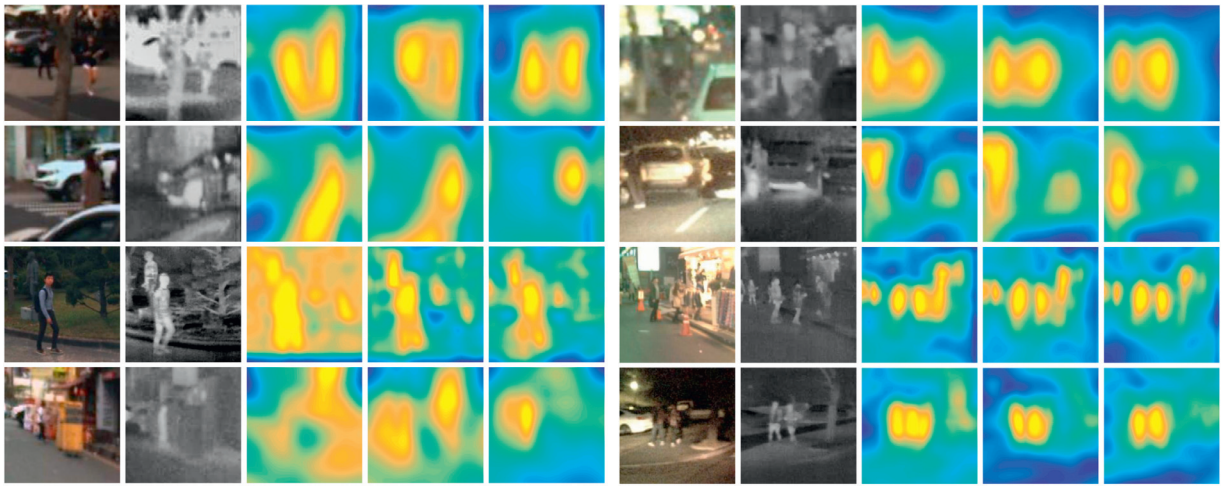


Fig. 9. Examples of multispectral pedestrian feature maps which are promoted by illumination-aware mechanism captured in (a) daytime and (b) nighttime scenes. The first two columns show the pictures of visible and thermal pedestrian instances respectively. The third to the fifth columns show the feature map visualizations generated from TDNN, IATDNN, and IATDNN + IAMSS respectively. It is noticed that the feature maps of multispectral pedestrian become more distinct by using our proposed illumination-aware modules (IATDNN and IAMSS).

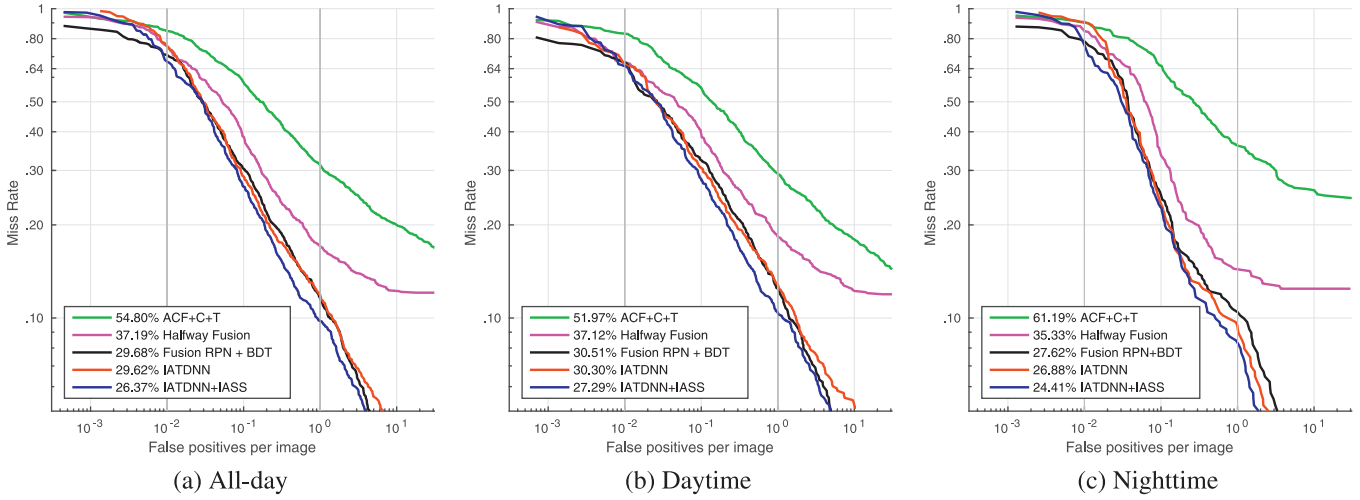


Fig. 10. Comparisons on the KAIST testing dataset under the reasonable setting during (a) all-day, (b) daytime, and (c) nighttime. It should be noted that legends indicate MR.

Table 5
Comprehensive comparison of IATDNN and IATDNN + IAMSS with the current state-of-the-art multispectral pedestrian detectors [19,20]. The computation efficiency is evaluated utilizing a single Titan X GPU. We execute each method for 100 times and compute the averaged runtime. It should be noted that DL represents deep learning and BF represents boosted forest [10].

	MR(%)	Runtime (s)	Method
Halfway Fusion [19]	37.19	0.40	DL
Fusion RPN+BDT [20]	29.68	0.80	DL+BF
TDNN	32.60	0.22	DL
IATDNN	29.62	0.24	DL
IATDNN+IAMSS	26.37	0.25	DL

We show the runtimes of IATDNN, IATDNN + IAMSS and state-of-the-art methods [19,20] in Table 5. We execute each method for 100 times and compute the averaged runtime. It is noticed that the efficiency of IATDNN + IAMSS outperform the state-of-the-art DNN-based approaches by a large margin (IATDNN + IAMSS 0.25s vs. Halfway Fusion 0.40s vs. Fusion RPN + BDT 0.80s). The architecture of Halfway Fusion includes an extra Fast R-CNN model [13] which significantly decreases the computational efficiency. Fusion RPN + BDT model utilizes boosting trees for classification, which increases the runtime by almost three times. It worth mentioning that our proposed illumination-aware networks will significantly improve detection performance while only incur a small computational overhead (TDNN 0.22s vs. IATDNN 0.24s vs. IATDNN + IAMSS 0.25s).

5. Conclusion

In our paper, we propose a novel multispectral pedestrian detector which is based on the joint learning of illumination-aware multispectral pedestrian detection and illumination-aware multispectral semantic segmentation. The illumination information encoded in multispectral images is utilized to accurately compute the illumination-aware weights through our designed illumination fully connected neural network (IFCNN). A novel illumination-aware weighting mechanism is developed to combine the day and night illumination sub-networks

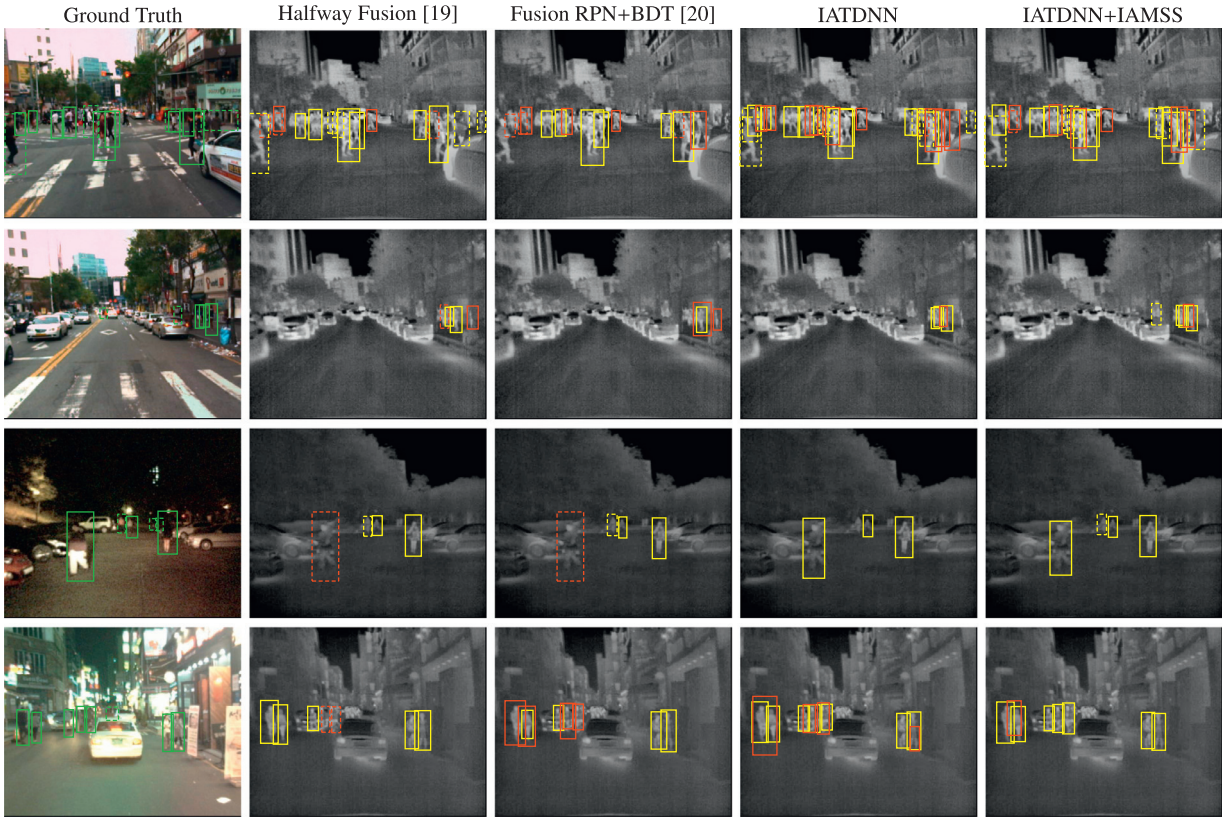


Fig. 11. Comparison with the current state-of-the-art multispectral pedestrian detectors [19,20]. The first column shows the input multispectral images with the ground truth labels in the visible channel and the other columns show the detection results of Halfway Fusion, Fusion RPN + BDT, IATDNN, and IATDNN + IAMSS in the thermal channel. It should be noted that green bounding boxes (BBs) in solid line denote positive labels, green BBs in dashed line denote ignore ones, yellow BBs in solid line denote true positives, yellow BBs in dashed line denote ignore detections, and red BBs denote false positives. We can observe that our proposed model can generate more accurate detections in comparison with the current state-of-the-art multispectral pedestrian detectors [19,20]. Some detected pedestrian instances are not even labeled by human observers. Best viewed in color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(illumination-aware pedestrian detection and illumination-aware semantic segmentation) together. Experimental results show that illumination-aware weighting mechanism provides an effective strategy to improve multispectral pedestrian detector. Moreover, we design four different multispectral segmentation infusion networks and find that the illumination-aware decision-stage multispectral semantic segmentation (IAMSS) generates the most reliable output. Experimental results on the KAIST public multispectral pedestrian benchmark illustrate that our proposed approach achieves more accurate detection results using less runtime in comparison with the current state-of-the-art multispectral detectors.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Nos. 51605428, 51575486 and U1664264). The authors would also like to thank the editors and the anonymous reviewers for their valuable suggestions.

References

[1] M. Bilal, A. Khan, M.U.K. Khan, C.M. Kyung, A low-complexity pedestrian detection framework for smart video surveillance systems, IEEE Trans. Circuits Syst. Video Technol. 27 (10) (2017) 2260–2273.
[2] S.K. Biswas, P. Milanfar, Linear support tensor machine with lsk channels: pedestrian detection in thermal infrared images, IEEE Trans. Image Process. 26 (9) (2017) 4229–4242.
[3] G. Brazil, X. Yin, X. Liu, Illuminating pedestrians via simultaneous detection & segmentation, in: IEEE International Conference on Computer Vision, IEEE, 2017.
[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 3213–3223.
[5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 1, IEEE, 2005, pp. 886–893.
[6] J.W. Davis, M.A. Keck, A two-stage template approach to person detection in thermal imagery, in: Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on, 1, IEEE, 2005, pp. 364–369.
[7] P. Dollár, Z. Tu, P. Perona, S. Belongie, Integral channel features, in: British Machine Vision Conference, 2009.
[8] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 743–761.
[9] A. Ess, B. Leibe, L. Van Gool, Depth and appearance for mobile scene analysis, in: IEEE International Conference on Computer Vision, IEEE, 2007, pp. 1–8.
[10] Y. Freund, R.E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: European conference on computational learning theory, Springer, 1995, pp. 23–37.
[11] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3354–3361.
[12] D. Geronimo, A.M. Lopez, A.D. Sappa, T. Graf, Survey of pedestrian detection for advanced driver assistance systems, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1239–1258.
[13] R. Girshick, Fast r-cnn, in: IEEE International Conference on Computer Vision, IEEE, 2015, pp. 1440–1448.
[14] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, A.M. López, Pedestrian detection at day/night time with visible and fir cameras: a comparison, Sensors 16 (6) (2016) 820.
[15] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask r-cnn, in: IEEE International Conference on Computer Vision, IEEE, 2017.
[16] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, in: European Conference on Computer Vision, Springer, 2014, pp. 346–361.

[17] S. Hwang, J. Park, N. Kim, Y. Choi, I. So Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 1037–1045.
[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 675–678.
[19] L. Jingjing, Z. Shaoting, W. Shu, M. Dimitris, Multispectral deep neural networks for pedestrian detection, in: British Machine Vision Conference, 2016, pp. 73.1–73.13.
[20] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, M. Teutsch, Fully convolutional region proposal networks for multispectral person detection, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2017, pp. 243–250.
[21] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.
[22] S.J. Krotosky, M.M. Trivedi, Person surveillance using visual and infrared imagery, IEEE Trans. Circuits Syst. Video Technol. 18 (8) (2008) 1096–1105.
[23] A. Leykin, Y. Ran, R. Hammoud, Thermal-visible video fusion for moving target tracking and pedestrian classification, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[24] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast r-cnn for pedestrian detection, IEEE Trans. Multimedia PP (99) (2017) 1–11.
[25] X. Li, L. Li, F. Flohr, J. Wang, H. Xiong, M. Bernhard, S. Pan, D.M. Gavrila, K. Li, A unified framework for concurrent pedestrian and cyclist detection, IEEE Trans. Intell. Transp. Syst. 18 (2) (2017) 269–281.
[26] X. Li, M. Ye, Y. Liu, F. Zhang, D. Liu, S. Tang, Accurate object detection using memory-based models in surveillance scenes, Pattern Recognit. 67 (2017) 73–84.
[27] M. Loveday, T.P. Breckon, On the impact of parallax free colour and infrared image co-registration to fused illumination invariant adaptive background modelling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1186–1195.
[28] J. Mao, T. Xiao, Y. Jiang, Z. Cao, What can help pedestrian detection? in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017.
[29] W. Nam, P. Dollár, J.H. Han, Local decorrelation for improved pedestrian detection, in: Advances in Neural Information Processing Systems, 2014, pp. 424–432.
[30] M. Oliveira, V. Santos, A.D. Sappa, Multimodal inverse perspective mapping, Inf. Fusion 24 (2015) 108–121.
[31] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, T. Poggio, Pedestrian detection using wavelet templates, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 1997, pp. 193–199.
[32] K. Park, S. Kim, K. Sohn, Unified multi-spectral pedestrian detection based on probabilistic fusion networks, Pattern Recognit. 80 (2018) 143–155.
[33] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: International Conference on Machine Learning, 2013, pp. 1310–1318.
[34] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.
[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[36] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, International Conference on Learning Representations, 2015.
[37] A. Torabi, G. Massé, G.-A. Bilodeau, An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications, Comput. Vision Image Understanding 116 (2) (2012) 210–221.
[38] J. Wagner, V. Fischer, M. Herman, S. Behnke, Multispectral pedestrian detection using deep fusion convolutional neural networks, in: European Symposium on Artificial Neural Networks, 2016, pp. 509–514.
[39] X. Wang, M. Wang, W. Li, Scene-specific pedestrian detection for static video surveillance, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 361–374.
[40] B. Wu, F. Iandola, P.H. Jin, K. Keutzer, SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving, IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 446–454.
[41] L. Zhang, L. Lin, X. Liang, K. He, Is faster r-cnn doing well for pedestrian detection? in: European Conference on Computer Vision, Springer, 2016, pp. 443–457.
[42] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, Towards reaching human performance in pedestrian detection, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2017) 1–11.
[43] S. Zhang, R. Benenson, B. Schiele, Filtered channel features for pedestrian detection, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015.
[44] S. Zhang, R. Benenson, B. Schiele, Citypersons: A diverse dataset for pedestrian detection, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017.
[45] M. Zinkevich, M. Weimer, L. Li, A.J. Smola, Parallelized stochastic gradient descent, in: Advances in neural information processing systems, 2010, pp. 2595–2603.