applied optics

Exploiting fusion architectures for multispectral pedestrian detection and segmentation

DAYAN GUAN,[1,2] YANPENG CAO,[1,2,*] JIANGXIN YANG,[1,2] YANLONG CAO,[1,2] AND CHRISTEL-LOIC TISSE[3]

[1]State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou, China
[2]Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, School of Mechanical Engineering, Zhejiang University, Hangzhou, China
[3]ULIS, ZI Les Iles Cordees-BP27, 38113 Veurey-Voroize, France
*Corresponding author: caoyp@zju.edu.cn

Received 1 February 2018; revised 21 March 2018; accepted 22 March 2018; posted 23 March 2018 (Doc. ID 321237); published 18 April 2018

Recent research has demonstrated that the fusion of complementary information captured by multi-modal sensors (visible and infrared cameras) enables robust pedestrian detection under various surveillance situations (e.g., daytime and nighttime). In this paper, we investigate a number of fusion architectures in an attempt to identify the optimal way of incorporating multispectral information for joint semantic segmentation and pedestrian detection. We made two important findings: (1) the sum fusion strategy, which computes the sum of two feature maps at the same spatial locations, delivers the best performance of multispectral detection, while the most commonly used concatenation fusion surprisingly performs the worst; and (2) two-stream semantic segmentation without multispectral fusion is the most effective scheme to infuse semantic information as supervision for learning human-related features. Based on these studies, we present a unified multispectral fusion framework for joint training of semantic segmentation and target detection that outperforms state-of-the-art multispectral pedestrian detectors by a large margin on the KAIST benchmark dataset.

OCIS codes: (150.4232) Multisensor methods; (330.1880) Detection; (100.4996) Pattern recognition, neural networks.

https://doi.org/10.1364/AO.57.00D108

1. INTRODUCTION

Pedestrian detection is an essential functionality in many human-centric applications such as urban monitoring and autonomous driving. Given sensing data captured in different surveillance situations, pedestrian detectors generate bounding boxes to accurately locate individual pedestrian instances. Many successful pedestrian detectors are trained using visible images only, and their performance is highly dependent on the illumination condition of a scene. In comparison, an infrared image captures radiation emitted by an object itself and remains robust against variations of illumination and shading [1]. As a result, thermal information is more suitable for target detection in low-light scenes (e.g., nighttime) [2,3].

Recently, a number of multispectral fusion solutions have been proposed [4–6]. Multispectral images provide complementary information about objects of interest, and human detection based on multi-cues is more robust under various surveillance situations (e.g., daytime and nighttime). Developing a fusion architecture that can adaptively incorporate human-related features extracted on visible and thermal channels is critical to achieving optimal detection performance, but it is not a trivial task.

In this paper, we first present a unified multispectral fusion framework for joint training of semantic segmentation and pedestrian detection. It consists of three major components: feature extraction/fusion, pedestrian detection, and semantic segmentation, as illustrated in Fig. 1. The feature extraction/fusion module extracts visible and thermal features individually and then combines them to generate the multispectral ones. Pedestrian detection takes the multispectral feature maps as input and generates predictions of pedestrian instances (confidence scores and bounding boxes). The semantic segmentation module outputs a number of segmentation masks that provide weak but helpful supervision to make two-stream features (visible and infrared) become more distinctive. Our proposed end-to-end method is trained using a multi-task loss function.

Based on this framework, we then experimentally explore the most effective configurations for multispectral detection and segmentation tasks. For multispectral pedestrian detection, we evaluate three different multispectral feature fusion strategies (concatenation, maximum, and sum). It is observed that the sum fusion strategy, which computes the sum of the two feature maps at the same spatial locations, delivers the best performance of multispectral detection, while the most commonly used concatenation fusion surprisingly performs the worst. Moreover, we explore two different multispectral semantic segmentation infusion architectures: fused semantic segmentation (FSS) and

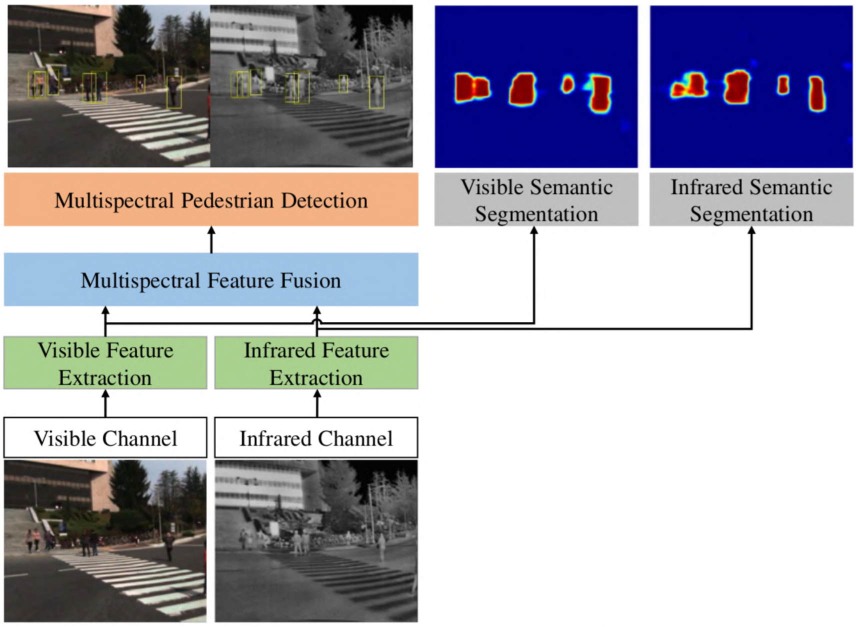

Multispectral Pedestrian Detection
Visible Semantic Segmentation
Infrared Semantic Segmentation
Multispectral Feature Fusion
Visible Feature Extraction
Infrared Feature Extraction
Visible Channel
Infrared Channel

Fig. 1. Architecture of our proposed multispectral fusion framework for joint training of semantic segmentation and pedestrian detection.

two-stream semantic segmentation (TSS). We found that the TSS scheme without multispectral fusion can better infuse semantic information to supervise the training of visible and thermal features. We demonstrate the effectiveness of this multispectral fusion framework on KAIST benchmark. Experimental results show that our proposed method outperforms state-of-the-art multispectral pedestrian detectors and achieves higher detection accuracy using less runtime. In summary, our contributions are as follows:

1. We evaluate a number of feature fusion strategies for multispectral pedestrian detection. Although concatenation is the most commonly used technique for feature fusion, the sum operation of two feature maps at the same spatial locations can combine two weak features to generate a stronger multispectral one and deliver better detection results.
2. We experimentally reveal that a simple TSS infusion architecture, which directly infuses visible/thermal semantic information into their corresponding feature maps without multispectral fusion, provides the most effective scheme to make use of semantics as supervision for visual feature learning.
3. Based on the above studies, we present a unified multispectral fusion framework for joint training of semantic segmentation and target detection that is trained end-to-end using a multi-task loss function. Our method achieves lower miss rate and faster runtime compared with state-of-the-art multispectral pedestrian detectors [7–9].

The remainder of our paper is organized as follows. We review some existing pedestrian detection solutions (both visible and multispectral) in Section 2. The details of our proposed multispectral fusion framework are presented in Section 3. An extensive evaluation of various fusion configurations and experimental comparison of multispectral pedestrian detectors are provided in Section 4, and Section 5 concludes this paper.

2. RELATED WORKS

Research most closely related to our work includes pedestrian detection solutions based on visible and multispectral images. A review of the recent research on these topics is presented below.

A. Visible Pedestrian Detection

In recent years, a large variety of methods have been proposed to improve visible pedestrian detection performance. Dalal and Triggs [10] designed the histograms of oriented gradient (HOG) descriptors with a cascaded linear support-vector network [11] for visible pedestrian detection. Dollár *et al.* [12] improved the HOG descriptors by building integrate channel features (ICF) with multi-channel feature pyramids followed by the AdaBoost classifier [13]. The multi-channel feature representations of ICF have been further improved in a number of studies, including ACF [14], LDCF [15], SCF [16], and Checkerboards [17]. Recently, deep-learning-based approaches for object detection [18–21] have been successfully adopted to boost the performance of visible pedestrian detection. Li *et al.* [22] presented a Scale-aware Fast R-CNN [19] model that adaptively combines the outputs from multiple built-in subnetworks to generate robust detection results. Cai *et al.* [23] developed a unified architecture of multi-scale deep neural networks, which provides a number of receptive fields to match pedestrians in different scales, to combine complementary scale-specific detectors together. Zhang *et al.* [24] presented a coarse-to-fine classification scheme for pedestrian detection by applying region proposal networks [20] followed by the AdaBoost classifier [13]. Mao *et al.* [25] developed a powerful deep neural network (DNN) architecture by using the information of aggregating extra features to boost detection performance without extra inputs in inference. Brazil *et al.* [26] proposed a novel multi-task learning scheme to improve pedestrian detection performance with joint supervision on semantic segmentation and pedestrian detection. It is proven that the simple box-based segmentation masks can provide helpful supervision information to achieve additional performance gains.

B. Multispectral Pedestrian Detection

It is experimentally demonstrated that pedestrian detectors trained using both visible and infrared images are able to generate more robust detection results than using visible images alone. Hwang *et al.* [7] built up a large-scale multispectral pedestrian benchmark dataset named KAIST by capturing well-aligned visible and thermal image pairs. The author further developed multispectral aggregated features (ACF+T+THOG) followed by the AdaBoost classifier [13] for classification. ACF+T+THOG concatenate the visible features (ACF) [14] and the infrared ones (T+THOG), which use the thermal image intensity (T) and the thermal HOG [10] features (THOG). Wagner *et al.* [27] considered the detections in Ref. [7] as proposals and classified them with a two-stream R-CNN [18], applying concatenation fusion in the late stage. The authors further compared the performance of architectures with different fusion stages and found that the late-stage fusion is the optimal architecture. Liu *et al.* [8] adopted the faster R-CNN [20] model for multispectral pedestrian detection tasks and developed four DNN-based fusion architectures in which two-branch networks are concatenated at different stages. Experimental results showed that the halfway fusion model that merges two-stream networks at a high-level convolutional stage performs best. König *et al.* [9] modified the visible pedestrian detector designed by Zhang *et al.* [24] to build Fusion RPN+BDT architecture for multispectral pedestrian detection. The Fusion RPN concatenated the two-branch region proposal network (RPN) on the high-level convolutional stage and achieved the current state-of-the-art performance on the KAIST multispectral dataset. Our method is distinctly different from the above approaches in two aspects. First, we investigate a number of fusion architectures in an attempt to identify the optimal way for incorporating complementary information from multispectral channels. Second, we develop a powerful multispectral pedestrian detector based on joint learning of pedestrian detection and semantic segmentation.

3. PROPOSED METHOD

A unified multispectral fusion framework is presented for joint training of semantic segmentation and pedestrian detection. It consists of three major components including feature extraction/fusion, pedestrian detection, and semantic segmentation. The details of each component are provided in the following subsections.

A. Multispectral Feature Extraction and Fusion

Figure 2 shows the architecture of DNNs for multispectral feature extraction and fusion. Given a pair of well-aligned visible and thermal images, we make use of the two-stream deep convolutional neural networks presented by König *et al.* [9] to extract semantic feature maps in individual channels. Note that each feature extraction stream consists of five convolutional layers and pooling ones (Conv1-V to Conv5-V in the visible stream and Conv1-T to Conv5-T in the thermal stream), which are designed based on the architectures of Conv1-5 from VGG-16 [28].

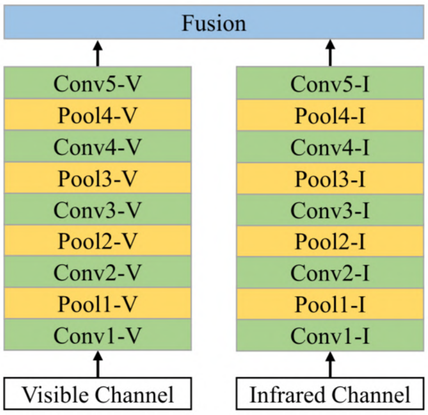


Fig. 2. Architecture of two-stream DNNs for multispectral feature extraction and fusion. Note that green boxes represent convolutional layers, yellow boxes represent pooling layers, and the blue box represents the fusion layer.

Previous studies revealed that a late-stage fusion scheme can generate semantic feature maps that are more suitable for high-level object detection tasks [8,9,18]. Based on this consideration, we deploy a fusion layer (Fusion) after the Conv5-V and Conv5-T layers to integrate the semantic feature maps from visible and infrared channels. The multispectral feature maps are obtained as

$$\mathbf{y} = f(\mathbf{p}, \mathbf{q}), \tag{1}$$

where f is the fusion function, $\mathbf{p} \in \mathbb{R}^{B_1 \times C_1 \times H_1 \times W_1}$ and $\mathbf{q} \in \mathbb{R}^{B_2 \times C_2 \times H_2 \times W_2}$ are the semantic feature maps extracted on the visible and thermal images, respectively, and $\mathbf{y} \in \mathbb{R}^{B \times C \times H \times W}$ is the fused multispectral feature maps. Note that B, C, H, and W are the batch size, number of channels, height, and weight of feature maps, respectively. Since two feature extraction streams of the visible and thermal channels use the same architecture, we have $B_1 = B_2$, $C_1 = C_2$, $W_1 = W_2$, $H_1 = H_2$. Here we consider three different fusion functions: f^{cat} (concatenation fusion), f^{max} (max fusion), and f^{sum} (sum fusion).

Concatenation fusion: $\mathbf{y}^{\text{cat}} = f^{\text{cat}}(\mathbf{p}, \mathbf{q})$ stacks the visible semantic feature maps $\mathbf{p}$ and the infrared ones $\mathbf{q}$ at the same spatial locations h, w and batch b, but across the feature channels as

$$\begin{cases} y^{\text{cat}}_{b,2c-1,h,w} = p_{b,c,h,w} \\ y^{\text{cat}}_{b,2c,h,w} = q_{b,c,h,w} \end{cases}, \tag{2}$$

where $b \in (1, 2, \ldots, B)$, $c \in (1, 2, \ldots, C)$, $h \in (1, 2, \ldots, H)$, $w \in (1, 2, \ldots, W)$, and $\mathbf{y} \in \mathbb{R}^{B \times 2C \times H \times W}$. Given stacked visible and thermal feature maps, their correlation is learned in the subsequent convolutional layer by minimizing the defined objective function. It is worth mentioning that the concatenation operation is the most commonly used technique to incorporate feature maps extracted in multispectral channels [7–9,18].

Max fusion: $\mathbf{y}^{\text{max}} = f^{\text{max}}(\mathbf{p}, \mathbf{q})$ outputs the maximum response of visible feature maps $\mathbf{p}$ and the infrared ones $\mathbf{q}$ at the same spatial locations h, w, channel number c, and batch b as

$$y^{\max}_{b,c,h,w} = \max\{p_{b,c,h,w}, q_{b,c,h,w}\}, \tag{3}$$

where the fused feature maps $\mathbf{y} \in \mathbb{R}^{B\times C\times H\times W}$ have the same size of visible $\mathbf{p}$ and thermal $\mathbf{q}$. Through max fusion, the more distinct features in the visible and infrared channels are selected to generate the multispectral representation.

Sum fusion: $\mathbf{y}^{\text{sum}} = f^{\text{sum}}(\mathbf{p}, \mathbf{q})$ calculates the sum of visible feature maps $\mathbf{p}$ and the infrared ones $\mathbf{q}$ at the same spatial locations h, w, channels c, and batch b as

$$y^{\text{sum}}_{b,c,h,w} = p_{b,c,h,w} + q_{b,c,h,w}, \tag{4}$$

and similarly, $\mathbf{y} \in \mathbb{R}^{B\times C\times H\times W}$ have the same size of $\mathbf{p}$ and $\mathbf{q}$. The sum fusion scheme combines the feature maps of visible and infrared channels using equal weights. Consequently, two weak features in visible and infrared channels might be added up to generate a strong one.

The abovementioned three fusion schemes (f^{cat}, $f^{\max}$, f^{sum}) have different working principles and will lead to different detection performance. Their comparative evaluation is provided in Section 4.C.

B. Multispectral Detection Networks

We combine the multispectral feature extraction and fusion DNN (Section 3.A) with the RPN model [24], which is a good performing pedestrian detector, to build fused region proposal network (FRPN) for multispectral pedestrian detection. The architecture of the FRPN is shown in Fig. 3.

Given the multispectral feature maps from the fusion layer (Fusion-Det), the FRPN generates classification scores (Cls in figures) and bounding boxes (Bbox in figures) as pedestrian detection results. Conv-Det is designed as a single 3 × 3 convolutional layer to generate human-related features from the multispectral semantic feature maps. Pedestrian detection results (classification scores and bounding boxes) are generated using two sibling 1 × 1 convolutional layers (Conv-C and Conv-B), which are attached after the Conv-Det. To train the FRPN, we define the detection loss term L_{det} as

$$L_{\text{det}} = \sum_{i\in S} L_{\text{cls}}(c_i, \hat{c}_i) + \lambda_r \sum_{i\in S} \hat{c}_i \cdot L_{\text{reg}}(b_i, \hat{b}_i), \tag{5}$$

where c_i is the classification score, b_i is the predicted bounding box, L_{cls} is the classification loss term, L_{reg} is the regression loss term, λ_r is the trade-off coefficient of loss term L_{reg}, and S is the set of training samples. A sample is labeled as a positive if the intersection-over-union (IoU) ratio between the sample's bounding box and any ground truth is greater than 0.5. Otherwise, the sample is labeled as a negative. We set training label $\hat{c}_i = 1$ for positive samples and $\hat{c}_i = 0$ for negative ones. In Eq. (5), the L_{cls} is the cross-entropy loss over the predicted score c and the training label $\hat{c}_i$, defined as

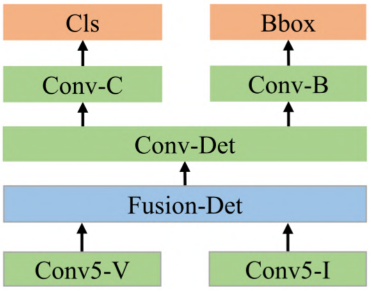


Fig. 3. Architecture of the FRPN for multispectral pedestrian detection. Note that green boxes represent convolutional layers, blue boxes represent fusion layers, and orange boxes represent detection results.

$$L_{\text{cls}}(c, \hat{c}) = -\hat{c} \cdot \log(c) - (1 - \hat{c}) \log(1 - c), \tag{6}$$

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

$$L_{\text{reg}}(b, \hat{b}) = \sum_{j\in\{x,y,w,h\}} \text{smooth}_{L_1}(b_j, \hat{b}_j), \tag{7}$$

where $b = (b_x, b_y, b_w, b_h)$ are the parameterized coordinates of the predicted bounding box, $\hat{b} = (\hat{b}_x, \hat{b}_y, \hat{b}_w, \hat{b}_h)$ are the coordinates of the ground-truth box, and smooth_{L_1} is the robust $L1$ loss function defined in Ref. [19].

C. Multispectral Segmentation Infusion Networks

Inspired by the success of multi-task (detection and segmentation) infusion framework for DNN-based object detection [21,26], we combine a multispectral semantic segmentation module with the FRPN for joint training of multispectral pedestrian detection and semantic segmentation. We adopt the single-channel box-based segmentation architecture presented by Brazil *et al.* [26] to facilitate multispectral semantic segmentation. The predicted segmentation masks provide weak but helpful supervision information for training more distinctive feature maps.

Given feature maps extracted on visible and infrared channels, learning multispectral semantic segmentation using the fused features or two individual ones would have different impacts on the infusion framework. To investigate the optimal semantic infusion architecture for multispectral pedestrian detection tasks, we combine the FRPN with two different multispectral semantic segmentation architectures, denoted FSS and TSS. As shown in Fig. 4, the FRPN+FSS generates multispectral segmentation masks using a single segmentation infusion layer (Conv-S) attached after the fusion layer (Fusion-Seg), while the FRPN+TSS applies two individual segmentation infusion layers (Conv-S-V and Convs-S-I) to generate different segmentation masks for visible and thermal channels. Note there are three options for the fusion layer, as discussed in Section 3.A. As suggested by Brazil *et al.* [26], a single 1 × 1 convolutional layer design is utilized in the segmentation infusion layers (Conv-S, Conv-S-V, and Conv-S-I), which directly infuses multispectral segmentation masks and produces the highest impact on the training of feature maps from Conv5-V and Conv5-I.

The segmentation loss term L_{seg} is defined as

$$L_{\text{seg}} = \sum_{i\in C} \sum_{j\in B} [-\hat{s}_j \cdot \log(s_{ij}) - (1 - \hat{s}_j) \cdot \log(1 - s_{ij})], \tag{8}$$

where s_{ij} is the predicted segmentation mask, C are the segmentation streams (note FSS contains only one segmentation stream while TSS contains two streams), and B are the box-based segmentation training samples in one mini-batch. If the sample is within a ground-truth bounding box, we set $\hat{s}_j = 1$, otherwise $\hat{s}_j = 0$. Comparative results of different multispectral semantic segmentation architectures are provided in Section 4.D.

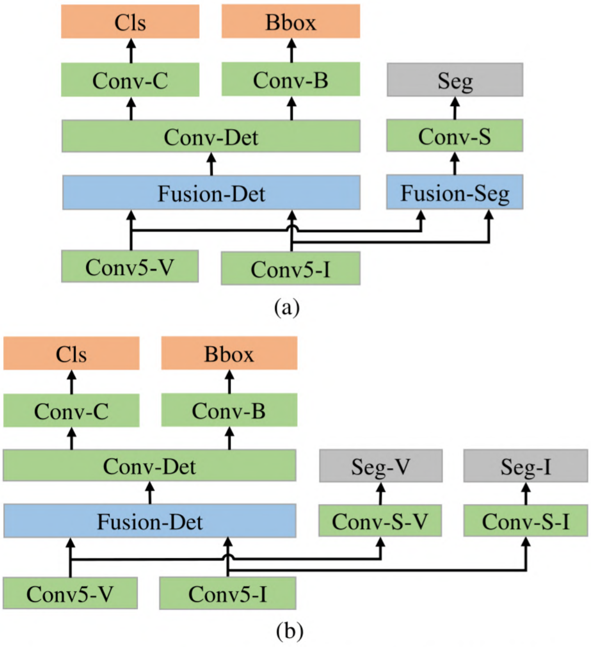


Fig. 4. Architectures of (a) FRPN+FSS and (b) FRPN+TSS for joint learning of multispectral pedestrian detection and semantic segmentation. Note that green boxes represent convolutional layers, blue boxes represent fusion layers, gray boxes represent segmentation masks, and orange boxes represent detection results.

To perform multi-task learning of multispectral pedestrian detection and semantic segmentation, we combine the detection loss term L_{det} defined in Eq. (5) with the segmentation loss term L_{seg} defined in Eq. (8) to build our final multi-task loss function as follows:

$$L = L_{\mathrm{det}} + \lambda_s L_{\mathrm{seg}}, \tag{9}$$

where λ_s is the trade-off coefficient of loss term L_{seg}. This final loss function is used for joint training of multispectral pedestrian detection and semantic segmentation.

4. EXPERIMENTAL RESULTS AND DISCUSSION

A. Dataset and Evaluation Metric

We evaluate our method on the public KAIST multispectral pedestrian benchmark [7]. The training dataset of KAIST contains 50,172 well-aligned visible–infrared image pairs captured under different lighting conditions with 13,853 pedestrian annotations. We sample the training images every two frames following [8,9]. The testing dataset of KAIST contains 2,252 image pairs with 1,356 pedestrian annotations. The ground-truth label whose bounding box is larger 55 pixels and not/partially occluded is considered as a pedestrian label. All the other labels are considered as *ignore* ones.

The log-average miss rate (MR) [14] is utilized to evaluate the quantitative performance of multispectral pedestrian detectors. A detection is counted as a true positive if the IoU ratio between the bounding box of the detection and any pedestrian label is greater than 50% [7–9]. Unmatched detections and unmatched pedestrian labels are counted as false positives and false negatives, respectively. Specifically, unmatched *ignore* labels are not counted as false negatives according to the evaluation protocol presented by Dollár *et al.* [14]. The MR is computed by averaging miss rate (false negative rate) at nine false positives per image (FPPI) rates that are evenly spaced in log-space from the range 10^{-2} to 10^{0} [7–9].

B. Implementation Details

We apply the image-centric training scheme to generate mini-batches and set the batch size $B = 1$, according to the method presented by Zhang *et al.* [24]. The feature extraction layers in each stream of FRPN (Conv1-V to Conv5-V in the visible stream and Conv1-I to Conv5-I in the infrared stream) are initialized using the parameters of the feature extraction layers in VGG-16 [28] (from Conv1 to Conv5) pre-trained on the ImageNet dataset [29] in parallel. All the other convolutional layers are initialized with standard Gaussian distribution following the method presented by Ren *et al.* [20]. We set $\lambda_r = 5$ in Eq. (5) according to the pedestrian detection method presented by Zhang *et al.* [24] and $\lambda_s = 1$ in Eq. (8) according to the single channel segmentation approach presented by Brazil *et al.* [26]. The Caffe package [30] framework is utilized to train our multispectral pedestrian detectors. All the detectors are trained with stochastic gradient descent (SGD) [31] for two epochs with learning rate of 0.001 and one more epoch with learning rate of 0.0001. To avoid exploding gradient problems [32], we clip gradients when the L2 norm of the gradients exceeds 10.

C. Evaluation of Fused Region Proposal Networks

We first compare the FRPN with different fusion layer (Fusion-Det) architectures in an attempt to identify the optimal feature fusion strategy for multispectral pedestrian detection. As described in Section 3.A, three different feature fusion strategies (concatenation, maximum, and sum) are used in the FRPN architectures, denoted FRPN-Cat, FRPN-Max, and FRPN-Sum, respectively. The detection loss term L_{det} defined in Eq. (5) is used to train the FRPN models. The performance (log-average MR [14]) of FRPN-Cat, FRPN-Max, and FRPN-Sum are quantitatively compared in Table 1. In addition, we conduct the qualitative comparison by displaying some sample detection examples of different architectures in Fig. 5.

It is observed that the selection of feature fusion strategy affects the performance of multispectral pedestrian detection. Our experimental results demonstrate that the FRPN-Sum performs better than the FRPN-Cat and FRPN-Max by achieving lower MR and more accurate detection results. Surprisingly, the concatenation fusion architecture, which is widely used in the current state-of-the-art deep-learning-based multispectral pedestrian detectors [8,9,27], performs the worst. The concatenation fusion directly stacks the two feature maps at the same spatial locations across feature channels. The correlation between visible and infrared feature maps is further learned in

Table 1. Quantitative Comparison (MR [14]) of FRPN Employing Different Feature Fusion Architectures

Model	All-Day (%)	Daytime (%)	Nighttime (%)
FRPN-Cat	32.60	33.80	30.53
FRPN-Max	31.54	32.66	29.43
FRPN-Sum	**30.49**	**31.27**	**28.29**

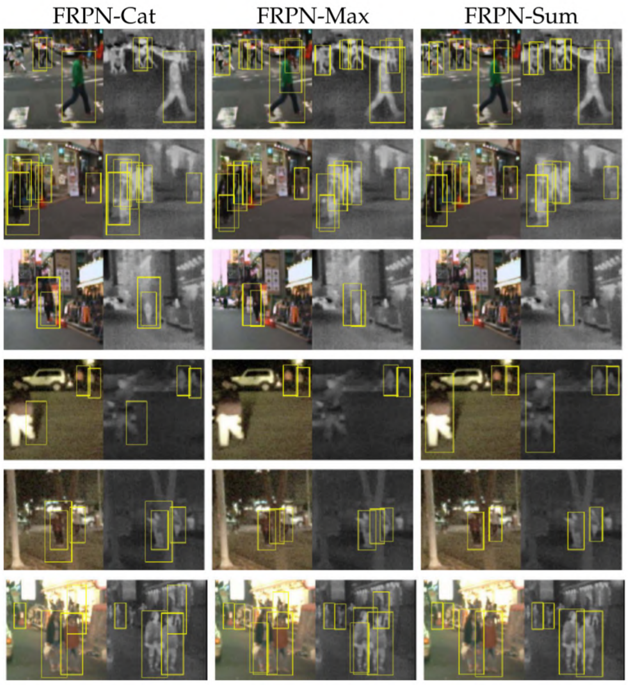


Fig. 5. Qualitative comparison of multispectral pedestrian detection results of FRPN-Cat, FRPN-Max, and FRPN-Sum. Note that yellow bounded boxes show pedestrian detections.

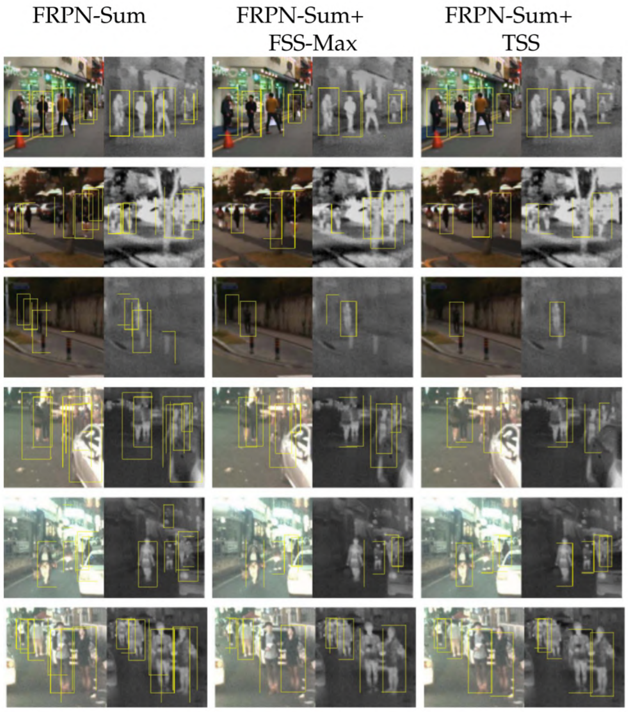


Fig. 6. Qualitative comparison of multispectral pedestrian detection results of FRPN-Sum, FRPN-Sum+FSS-Max, and FRPN-Sum+TSS. Note that yellow bounded boxes show pedestrian detections.

the subsequent layers (e.g., Conv-Det). Intuitively, such correlation is scene dependent. For instance, this correlation should be different for daytime and nighttime scenes. However, the defined loss function contains only detection and segmentation loss terms and does not provide effective information to supervise the learning of this scene-dependent correlation. Another reasonable explanation for this phenomenon is that it is difficult to make use of a single Conv-Det layer to learn both human-related features and correlation between visible and infrared feature maps. In comparison, better detection results are achieved by FRPN-Max and FRPN-Sum architectures in which the correlation between visible and infrared feature maps is defined through either a max or sum function. We also noticed that FRPN-Sum can better recognize instances whose characteristic features are weak in both visible and infrared channels. Different from the max fusion scheme, which keeps only the more distinct features in either the visible or infrared channel, the sum fusion function can combine two weak but complementary features to generate a stronger cue for classification.

D. Evaluation of Multispectral Segmentation Infusion Networks

We further evaluate the performance gain by incorporating the multispectral semantic segmentation information into the FRPN model. For this purpose, we design two different multispectral pedestrian detectors, FRPN-Sum+FSS and FRPN-Sum+TSS, by combining the best performing FRPN-Sum with FSS and TSS, respectively. Furthermore, we apply the different feature fusion strategies (concatenation, maximum, and sum) described in Section 3.A in the fusion layer to design three FSS architectures, denoted FSS-Cat, FSS-Max, and FSS-Sum, respectively. The quantitative performance (log-average MR [14]) of FRPN-Sum, FRPN-Sum+FSS-Cat, FRPN-Sum+FSS-Max, FRPN-Sum+FSS-Sum, and FRPN-Sum+TSS are compared in Table 2.

We observe that better detection performance can generally be achieved through the joint training of pedestrian detection and multispectral semantic segmentation. The underlying reason is that the semantic segmentation masks are able to provide additional supervision to facilitate the training of more distinct human-related features for pedestrian detection [26]. Another important observation is that TSS without multispectral fusion provides the most effective way to infuse semantic information as supervision for learning human-related features, performing better than FSS-Cat, FSS-Max, and FSS-Sum architectures. A logical explanation is that the TSS generates the most effective supervision information for feature training by directly infusing

Table 2. Quantitative Comparison (MR [14]) of FRPN-Sum with Different Multispectral Semantic Segmentation Architectures[a]

Model	All-Day (%)	Daytime (%)	Nighttime (%)
FRPN-Sum	30.49	31.27	28.29
FRPN-Sum+FSS-Cat	28.32	28.29	27.44
FRPN-Sum+FSS-Sum	28.12	28.80	25.85
FRPN-Sum+FSS-Max	27.65	28.47	25.53
FRPN-Sum+TSS	**26.67**	**26.75**	**25.24**

[a] Note that FSS-Cat, FSS-Max, and FSS-Sum apply concatenation, maximum, and sum fusion, respectively, while TSS represents the two-stream semantic segmentation.

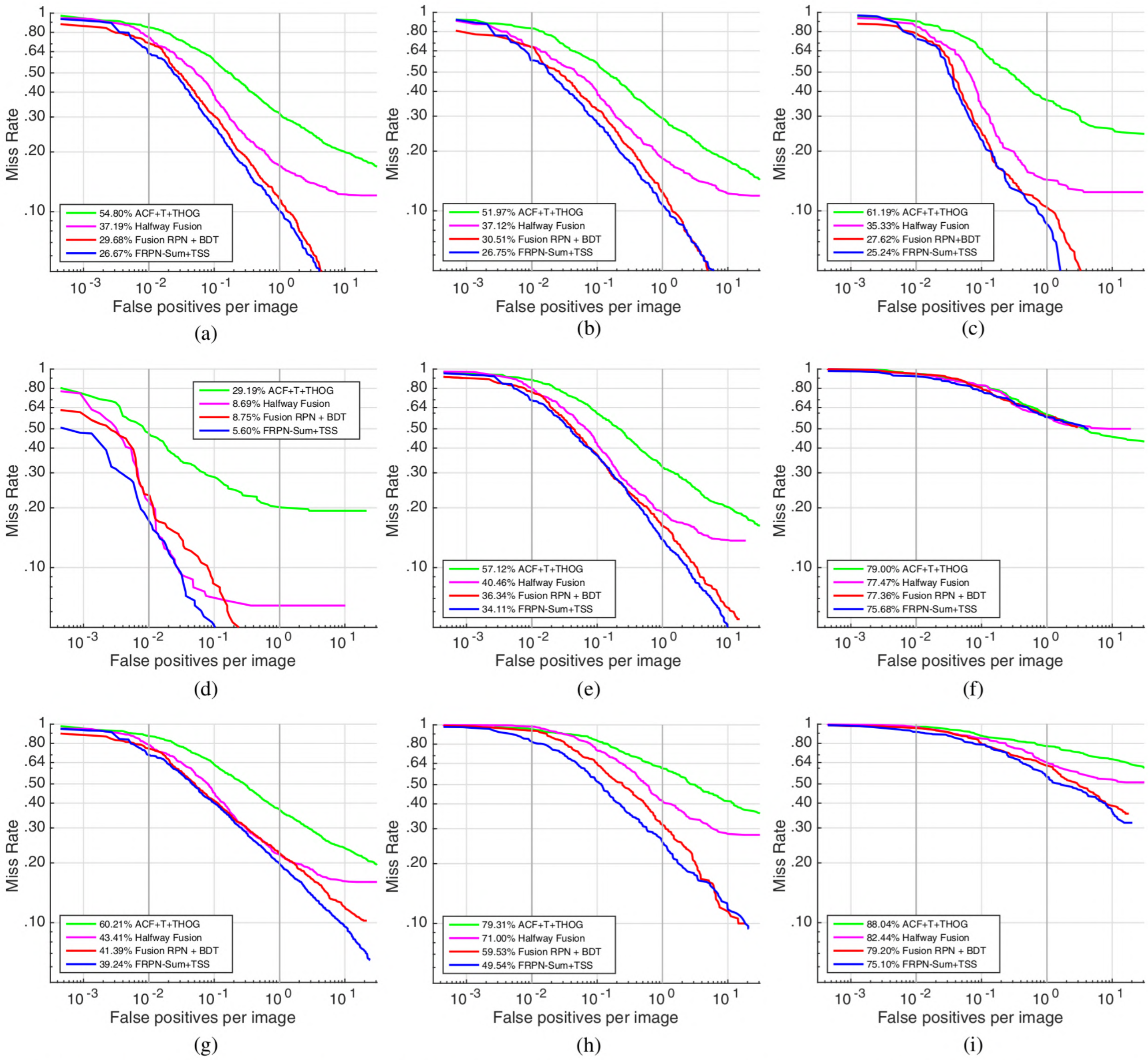
visible/infrared semantic information into their corresponding feature maps.

In Fig. 6, we show some sample detection results of FRPN-Sum, FRPN-Sum+FSS-Max, and FRPN-Sum+TSS. It is noted that FRPN-Sum+TSS significantly reduces false detections under cluttered background by incorporating substantial two-stream multispectral semantic segmentation information into FRPN-Sum.

E. Comparison with the State-of-the-Art

We compare the proposed FRPN-Sum+TSS with three other state-of-the-art multispectral pedestrian detection methods: ACF+T+THOG [7], Halfway Fusion [8], and Fusion RPN +BDT [9]. The ACF+T+THOG [7] and Fusion RPN +BDT [9] models are re-implemented and trained according to the original papers, and the detection results of the Halfway Fusion method [8] are kindly provided by the authors.

To quantitatively evaluate different multispectral pedestrian detectors, we consider the reasonable, scale, and occlusion subsets of the KAIST test dataset [7] to plot MR against FPPI [14] by varying the threshold on detection scores in Fig. 7. Our proposed FRPN-Sum+TSS achieves an impressive 26.67% MR on the reasonable subset (all-day scenes). The quantitative performance gain is a relative improvement rate of 10% compared with the current state-of-the-art multispectral pedestrian detection method Fusion RPN+BDT (29.68%). Our proposed detector surpasses the state-of-the-art method in both reasonable daytime (28.75% versus 30.51%) and nighttime (25.24% versus 27.62%) scenes. Moreover, our proposed FRPN-Sum+TSS



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 7. Comparing the quantitative performance of FRPN-Sum+TSS with the current state-of-the-art methods by plotting miss rate against FPPI (using log-log plots) [14] on the various subsets of KAIST test dataset [7] (legends indicate MR). (a) Reasonable all, (b) reasonable day, (c) reasonable night, (d) near scale, (e) medium scale, (f) far scale, (g) no occlusion, (h) partial occlusion, and (i) heavy occlusion.

outperforms Fusion RPN+BDT [9] on different scale subsets (*near*, 5.60% versus 8.75%; *medium*, 34.11% versus 36.34%; *far*, 75.68% versus 77.36%) and on three occlusion ones (*no*, 39.24% versus 41.39%; *partial*, 49.54% versus 59.53%; *heavy*, 75.10% versus 79.20%). These comparative results indicate that our proposed multispectral pedestrian detector achieves more robust performance under various surveillance situations.

We also compare the computing efficiency of FRPN-Sum+TSS with state-of-the-art methods in Table 3. The efficiency of FRPN-Sum+TSS surpasses the current state-of-the-art deep-learning approaches for multispectral pedestrian detection by a large margin, with 0.23 s/image versus 0.40 s/image on average.

Table 3. Comparing the Quantitative Performance (MR [14] in All-Day) and Runtime Performance of FRPN-Sum+TSS with State-of-the-Art Methods[a]

Model	MR (%)	Runtime (s)	Method
Halfway Fusion [8]	37.19	0.40	DL
Fusion RPN+BDT [9]	29.68	0.80	DL+AB
FRPN-Sum+TSS	**26.67**	**0.23**	**DL**

[a] A single Titan X GPU is utilized to evaluate the computation efficiency. Note that DL denotes deep learning and AB denotes AdaBoost [13].

We qualitatively evaluate the multispectral pedestrian detectors by visualizing some sample detection results in Fig. 8. Comparing with the current state-of-the-art multispectral pedestrian detection methods, our proposed FRPN-Sum+TSS is able to generate more accurate pedestrian detections (predicting more true positives and fewer false positives). It is worth mentioning that FRPN-Sum+TSS can successfully detect pedestrian instances that are falsely missed in the KAIST testing dataset, as illustrated in Fig. 8. As a result, some of our correct detection results will be considered as false positives. In the future, we plan to improve annotations of the KAIST testing dataset to facilitate better evaluation of multispectral pedestrian detectors.

5. CONCLUSION

In this paper, we present a powerful multispectral pedestrian detector based on multi-task learning of pedestrian detection and semantic segmentation. For the task of multispectral pedestrian detection, different strategies (concatenation, maximum,

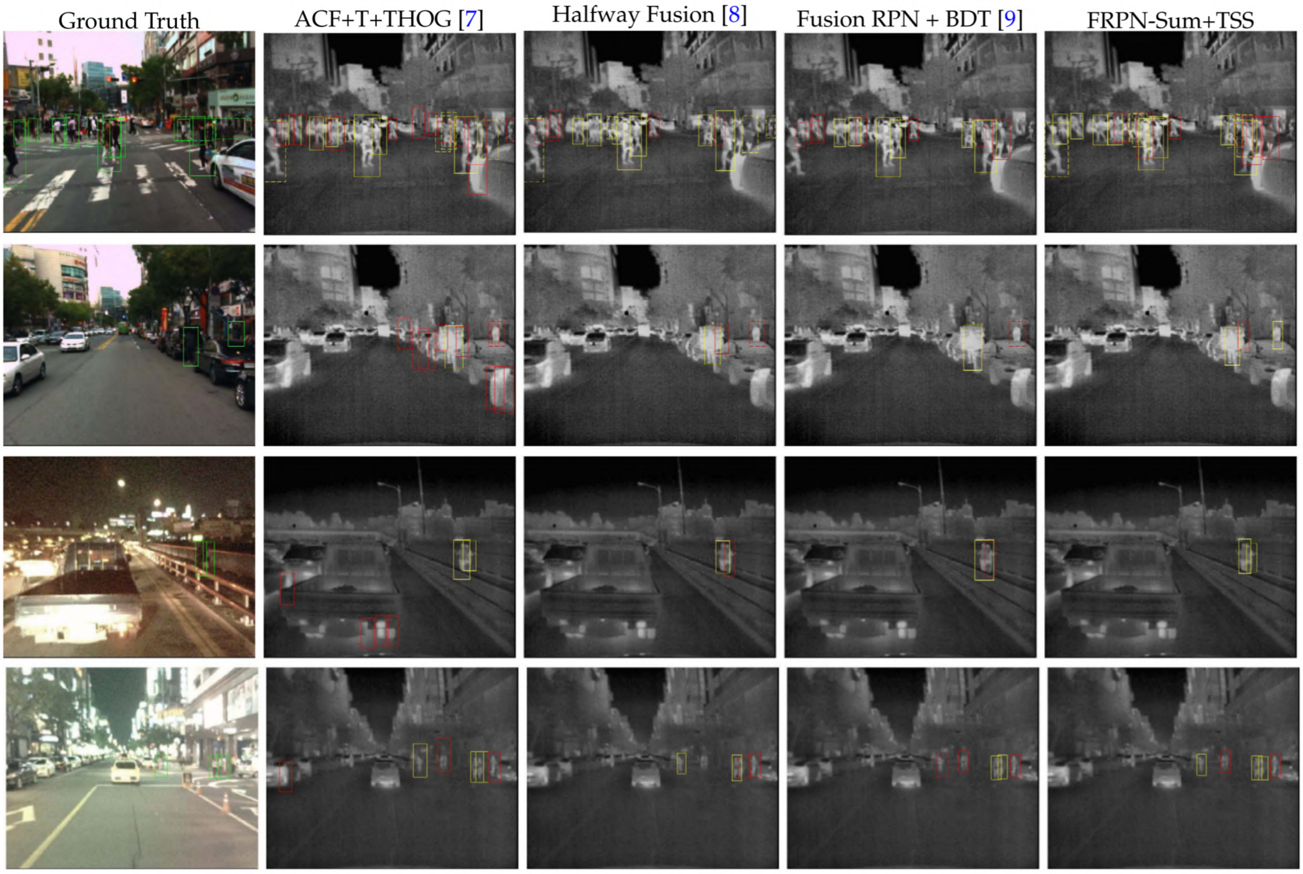


Fig. 8. Qualitative comparison of multispectral pedestrian detection results in the KAIST testing dataset with other state-of-the-art approaches. The first column shows the ground truth (displayed using the visible channel), and the others show detection results of ACF+T+THOG [7], Halfway Fusion [8], Fusion RPN+BDT [9], and our proposed FRPN-Sum+TSS (displayed using the thermal channel). Note that green bounding boxes (BBs) in solid line show positive labels, green BBs in dashed line show *ignore* ones, yellow BBs in solid line show true positives, yellow BBs in dashed line show *ignore* detections, red BBs in solid line show false positives, and red BBs in dashed line show false negatives. It is observed that our approach is able to generate more accurate detection results compared with current state-of-the-art multispectral pedestrian detectors [7–9]. Some detected persons are not even annotated by human observers.

and sum) are systematically investigated to explore the optimal multispectral feature fusion architecture. We experimentally demonstrate that the sum operation performs the best and the widely used concatenation fusion scheme surprisingly performs the worst. We extend an existing single-channel semantic segmentation architecture to handle multispectral data. More accurate detection results can be obtained by infusing the multispectral semantic segmentation masks as supervision for learning human-related features. Moreover, we explore four different architectures for multispectral semantic segmentation and reveal that TSS is the most effective infusion scheme for multispectral pedestrian detection. Experimental results on the KAIST benchmark show that both the quantitative and qualitative performance of our proposed FRPN-Sum+TSS is better than ones of the current state-of-the-art multispectral pedestrian detectors.

Funding. National Natural Science Foundation of China (NSFC) (51575486, 51605428, U1664264); Fundamental Research Funds for the Central Universities.

REFERENCES

1. A. Nabatchian, E. Abdel-Raheem, and M. Ahmadi, "Illumination invariant feature extraction and mutual-information-based local matching for face recognition under illumination variation and occlusion," Pattern Recogn. **44**, 2576–2587 (2011).
2. J. Ge, Y. Luo, and G. Tei, "Real-time pedestrian detection and tracking at nighttime for driver-assistance systems," IEEE Trans. Intell. Transp. Syst. **10**, 283–298 (2009).
3. Y. Li, P. Li, and Q. Shen, "Real-time infrared target tracking based on l1 minimization and compressive features," Appl. Opt. **53**, 6518–6526 (2014).
4. X. Yan, H. Qin, J. Li, H. Zhou, J.-G. Zong, and Q. Zeng, "Infrared and visible image fusion using multiscale directional nonlocal means filter," Appl. Opt. **54**, 4299–4308 (2015).
5. Z. Zhou, M. Dong, X. Xie, and Z. Gao, "Fusion of infrared and visible images for night-vision context enhancement," Appl. Opt. **55**, 6480–6490 (2016).
6. Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: recent advances and future prospects," Inf. Fusion **42**, 158–173 (2018).
7. S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: benchmark dataset and baseline," in *IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1037–1045.
8. J. Liu, S. Zhang, S. Wang, and D. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *British Machine Vision Conference* (2016), pp. 73.1–73.13.
9. D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2017), pp. 243–250.
10. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition* (2005), pp. 886–893.
11. C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn. **20**, 273–297 (1995).
12. P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *British Machine Vision Conference* (2009).
13. Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory* (1995), pp. 23–37.
14. 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).
15. W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural Information Processing Systems* (2014), pp. 424–432.
16. R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *European Conference on Computer Vision* (2014), pp. 613–627.
17. S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition* (2015).
18. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 580–587.
19. R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision* (2015), pp. 1440–1448.
20. 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), pp. 91–99.
21. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision* (2017).
22. 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 (2018).
23. Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision* (2016), pp. 354–370.
24. L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *European Conference on Computer Vision* (2016), pp. 443–457.
25. J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *IEEE Conference on Computer Vision and Pattern Recognition* (2017).
26. G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *IEEE International Conference on Computer Vision* (2017).
27. J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *European Symposium on Artificial Neural Networks* (2016), pp. 509–514.
28. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations* (2015).
29. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," Int. J. Comput. Vis. **115**, 211–252 (2015).
30. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: convolutional architecture for fast feature embedding," in *International Conference on Multimedia* (2014), pp. 675–678.
31. M. Zinkevich, M. Weimer, L. Li, and A. J. Smola, "Parallelized stochastic gradient descent," in *Advances in Neural Information Processing Systems* (2010), pp. 2595–2603.
32. R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning* (2013), pp. 1310–1318.