CHAPTER 5

Multimodal Fusion Architectures for Pedestrian Detection

Dayan Guan[*,†], **Jiangxin Yang**[*,†], **Yanlong Cao**[*,†], **Michael Ying Yang**[‡], **Yanpeng Cao**[*,†]

[*]State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou, China [†]Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, School of Mechanical Engineering, Zhejiang University, Hangzhou, China [‡]Scene Understanding Group, University of Twente, Enschede, The Netherlands

Contents

5.1 Introduction 101
5.2 Related Work 105
5.2.1 Visible Pedestrian Detection 105
5.2.2 Infrared Pedestrian Detection 107
5.2.3 Multimodal Pedestrian Detection 108
5.3 Proposed Method 110
5.3.1 Multimodal Feature Learning/Fusion 110
5.3.2 Multimodal Pedestrian Detection 112
5.3.2.1 *Baseline DNN model 112*
5.3.2.2 *Scene-aware DNN model 113*
5.3.3 Multimodal Segmentation Supervision 116
5.4 Experimental Results and Discussion 118
5.4.1 Dataset and Evaluation Metric 118
5.4.2 Implementation Details 118
5.4.3 Evaluation of Multimodal Feature Fusion 119
5.4.4 Evaluation of Multimodal Pedestrian Detection Networks 121
5.4.5 Evaluation of Multimodal Segmentation Supervision Networks 124
5.4.6 Comparison with State-of-the-Art Multimodal Pedestrian Detection Methods 125
5.5 Conclusion 130
Acknowledgment 130
References 130

5.1 Introduction

In recent years, pedestrian detection has received wide attention in the computer vision community [42,9,13,16,15,6]. Given images captured in various real-world environment, pedes-

Multimodal Scene Understanding
https://doi.org/10.1016/B978-0-12-817358-9.00011-1

Figure 5.1 : Detections generated by a well-trained visible pedestrian detector [60]. (A) Detections generated using an image from the public Caltech testing dataset [12]; (B) detections generated using a visible image from the public KAIST testing dataset [26] captured during daytime; (C) detections generated using a visible image from the public KAIST testing dataset [26] captured during nighttime in good illumination condition; (D) detections generated using a visible image from the public KAIST testing dataset [26] captured during nighttime in bad illumination condition. Please note that green bounding boxes (BBs) represent true positives, red BBs represent false positives, and red BBs in dashed line represent false negatives.

trian detection solution is needed to generate bounding boxes to identify individual pedestrian instances accurately. It supplies a vital functionality to assist a number of human-centric applications, such as video surveillance [54,2,36], urban scene analysis [17,7,63] and autonomous driving [56,35,61].

Although some improvements have been achieved in the past years, it remains a challenging task to develop a robust pedestrian detector for real applications. Most of the detectors existed are trained using visible images only, thus their performances are unstable in various illumination environments as illustrated in Fig. 5.1. In order to overcome the limitation,

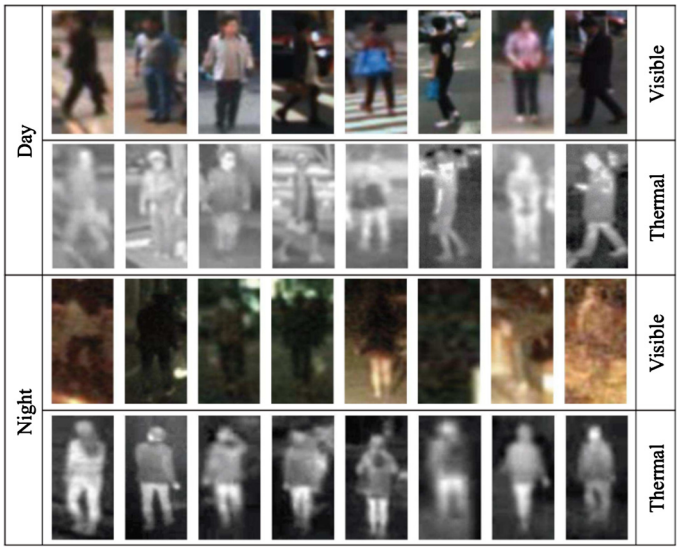


Figure 5.2 : Examples of pedestrian samples in the multimodal (visible and thermal) images captured in daytime and nighttime scenes [26]. It should be noted that multimodal data can supply complementary information about the target which could be effectively integrated to obtain more robust detection results.

multimodal pedestrian detection solutions are studied by many researchers to facilitate robust pedestrian detection for around-the-clock application [33,31,51,41,26,20]. The underlying reason is that multimodal images can supply complementary information about the target as shown in Fig. 5.2, therefore more robust detection results can be generated by the fusion of multimodal data. Designing an effect fusion architecture which can adaptively integrate multimodal features is critical to improving the detection results.

In Fig. 5.3, we illustrate the workflow of our proposed multimodal fusion framework for joint training of segmentation supervision and pedestrian detection. It contains three modules including feature learning/fusion, pedestrian detection, and segmentation supervision. The feature learning/fusion extracts features in individual channels (visible and thermal) and then integrate them to obtain multimodal feature maps. The pedestrian detection module generates predictions of the targets (confidence scores and bounding boxes) utilizing the generated multimodal feature maps. The segmentation supervision module improves the distinctiveness of features in individual channels (visible and thermal) through the supervision learning of segmentation masks. Pedestrian detection and segmentation supervision are trained end-to-end using a multitask loss function.

Based on this baseline framework, we organize experiments to explore a number of multimodal fusion models to identify the most effective scheme for the joint learning task of

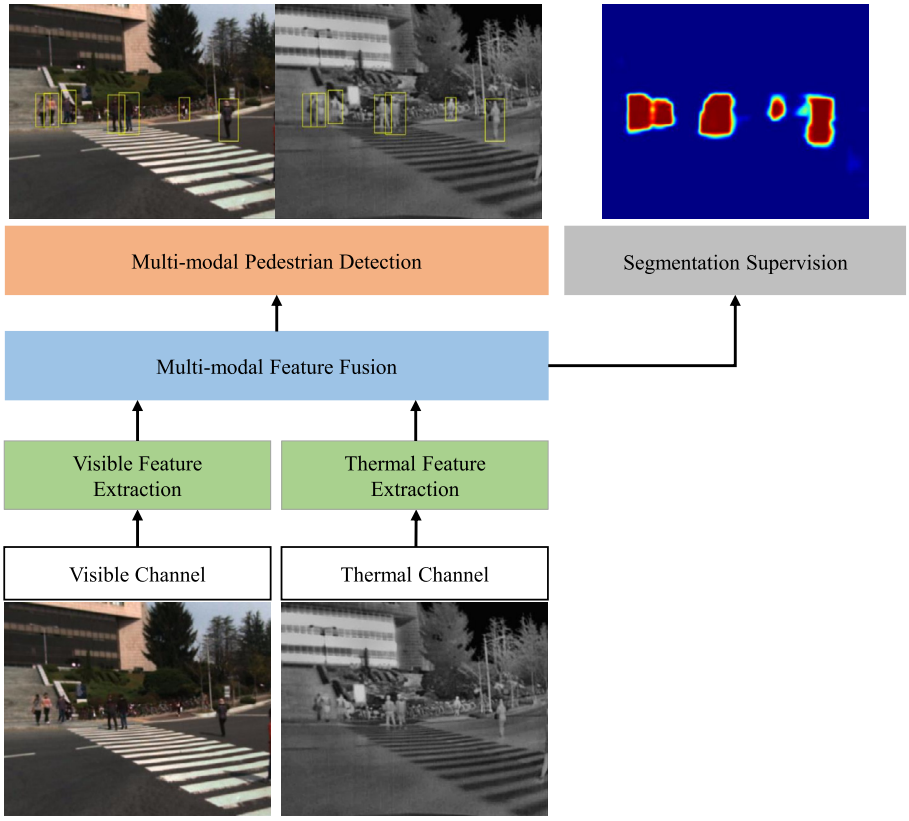


Figure 5.3 : The architecture of our proposed multimodal fusion framework for joint training of pedestrian detection and segmentation supervision.

pedestrian detection and segmentation supervision. For multimodal pedestrian detection, we evaluate three different multimodal feature fusion schemes (concatenation, maximization, and summation) and two deep neural network (DNN) models with and without a scene-aware mechanism. It can be observed that integrating the concatenation fusion scheme with a global scene-aware mechanism leads to better learning of both human-related features and correlation between visible and thermal feature maps. Moreover, we explore four different multimodal segmentation supervision infusion architectures (fused segmentation supervision with/without a scene-aware mechanism and two-stream semantic segmentation with/without a scene-aware mechanism). We found that the two-stream segmentation supervision with a scene-aware mechanism can better infuse semantic information to supervise the training of visible and thermal features. Our proposed method outperforms state-of-the-art multimodal pedestrian detectors and achieves higher detection accuracy using less runtime on public KAIST benchmark. In summary, our **contributions** are as follows:

1. We evaluate a number of feature fusion strategies for multimodal pedestrian detection. Compared with other basic fusion algorithms, integrating the concatenation with a global scene-aware mechanism leads to better learning of both human-related features and correlation between visible and thermal feature maps.
2. We experimentally reveal that the two-stream segmentation supervision infusion architecture, which individually infuses visible/thermal semantic information into their corresponding feature maps without multimodal fusion, provides the most effective scheme to make use of semantic information as supervision for visual feature learning.
3. We present a unified framework for joint training of multimodal segmentation supervision and target detection using a multitask loss function. Our method achieves the most accurate results with the least runtime compared with the current state-of-the-art multimodal pedestrian detectors [26,28,30].

This chapter further extends our previous work [22,23] in two aspects. First, we systematically evaluate three basic multimodal feature fusion schemes (concatenation, maximization, and summation) and two fusion models with and without a scene-aware mechanism. Our experimental results reveal that the basic concatenation fusion scheme combined with a global scene-aware mechanism performs better than other alternatives. Second, we utilize the original KAIST testing dataset and the improved one provided by Liu et al. [37] for quantitative evaluation. In both datasets, our proposed method achieves the most accurate results with the least runtime compared with the current state-of-the-art multimodal pedestrian detectors [26, 28,30].

The remainder of our chapter is organized as follows. Some existing research work as regards pedestrian detection using visible, thermal and multimodal images is summarized in Sect. 5.2. We present our proposed multimodal fusion architectures in detail in Sect. 5.3. The evaluation of a number of multimodal fusion architectures and the experimental comparison of multimodal pedestrian detectors are presented in Sect. 5.4. This chapter is concluded in Sect. 5.5.

5.2 Related Work

Pedestrian detection methods using visible, thermal and multimodal images are closely relevant to our work. We present a review of the recent researches on these topics below.

5.2.1 Visible Pedestrian Detection

A large variety of methods have been proposed to perform pedestrian detection using visible information in the last 20 years. Papageorgiou et al. [43] firstly abandoned motion cues

to train a visible pedestrian detection system, utilizing Haar wavelets transform and support-vector machines (SVMs) [8]. The authors also collected a new visible pedestrian database (MIT) to evaluate their method. Viola and Jones [52] designed the integral images with a cascade structure for fast feature computation and efficient visible pedestrian detection, applying the AdaBoost classifier for automatic and accurate feature selection. Wu et al. [55] firstly proposed sliding window detectors for pedestrian detection, applying multiscale Haar wavelets and support-vector machines (SVMs) [8]. Dalal et al. [9] proposed the histograms of oriented gradient (HOG) descriptors along with a cascaded linear support-vector network for more accurate pedestrian detection. Dollar et al. [11] improved the HOG descriptors by designing the integrate channel features (ICFs) descriptors with an architecture of multichannel feature pyramids followed by the AdaBoost classifier. The feature representations of ICF have been further improved through multiple techniques including ACF [12], LDCF [39], SCF [1], and Checkerboards [62].

In recent years, DNN based approaches for object detection [19,18,45,25] have been widely adopted to improve the performance of visible pedestrian detection. Sermanet et al. [47] applied convolutional sparse coding to pretrain convolutional neural networks (CNNs) for feature extraction and fully-connected (FC) layers for classification. Tian et al. [50] developed extensive part detectors using weakly annotated humans to handle occlusion problems occurred in pedestrian detection. Li et al. [34] presented a scale-aware DNN which adaptively combines the outputs from a large-size sub-network and a small-size one to generate robust detection results at different scales. Cai et al. [5] designed the unified multiscale DNN to combine complementary scale-specific ones, applying various receptive fields to identify pedestrians in different scales. Zhang et al. [60] proposed a coarse-to-fine classification scheme for visible pedestrian detection by applying region proposal networks [45] to generate high-resolution convolutional feature maps, which are followed by the AdaBoost classifier for final classification. Mao et al. [38] designed a powerful deep convolutional neural networks architecture by utilizing the information of aggregating extra features to improve pedestrian detection performance without additional inputs in inference. Brazil et al. [4] developed a novel multitask learning scheme to improve visible pedestrian detection performance with the joint supervision on weakly box-based semantic segmentation and pedestrian detection, indicating that the box-based segmentation masks can provide sufficient supervision information to achieve additional performance gains.

We summarize the mentioned visible pedestrian detection methods in Table 5.1. However, pedestrian detectors trained on visible images are sensitive to changes of illumination, weather, and occlusions. These detectors are very likely to be stuck with images captured during nighttime.

Table 5.1: The summarization of visible pedestrian detection methods.

	Feature Extractor	Feature Classifier	Highlight	Year
Papageorgiou et al. [43]	Haar wavelets	SVM	No motion cues	1999
Viola and Jones [52]	Haar wavelets	AdaBoost	Cascade structure	2004
Wu et al. [55]	Haar wavelets	SVM	Sliding window	2005
Dalal et al. [9]	HOG	SVM	HOG descriptors	2005
Dollar et al. [11]	ICF	AdaBoost	ICF descriptors	2009
Dollar et al. [12]	ACF	AdaBoost	ACF descriptors	2012
Sermanet et al. [47]	CNN	FC	Convolutional sparse coding	2013
Nam et al. [39]	LDCF	AdaBoost	LDCF descriptors	2014
Benenson et al. [1]	SCF	AdaBoost	SCF descriptors	2014
Zhang et al. [62]	Checkerboards	AdaBoost	Checkerboards descriptors	2015
Tian et al. [50]	CNN	FC	Extensive part detectors	2015
Li et al. [34]	CNN	FC	Scale-aware mechanism	2015
Cai et al. [5]	CNN	FC	Multiscale CNN	2016
Mao et al. [38]	CNN	FC	Aggregating extra features	2017
Brazil et al. [4]	CNN	FC	Segmentation supervision	2017

5.2.2 Infrared Pedestrian Detection

Infrared imaging sensors, which provide excellent visible cues during nighttime, have found their importance for around-the-clock robotic applications, such as autonomous vehicle and surveillance system.

Nanda et al. [40] presented a real-time pedestrian detection system that works on low quality thermal videos. Probabilistic templates are utilized to capture the variations in human targets, working well especially for the case when object contrast is low and body parts are missing. Davis et al. [10] utilized a generalized template and an AdaBoosted ensemble classifier to detect people in widely varying thermal imagery. The authors also collected a challenging thermal pedestrian dataset (OSU-T) to test their method. Suard et al. [49] developed image descriptors, based on histograms of oriented gradients (HOG) with a support-vector machine (SVM) classifier, for pedestrian detection applied to stereo thermal images. This approach achieved good results for window classification in a video sequence. Zhang et al. [59] investigated the approaches derived from visible spectrum analysis for the task of thermal pedestrian detection. The author extended two feature classes (edgelets and HOG features) and two classification models (AdaBoost and SVM cascade) to the thermal images. Lee et al. [32] presented a nighttime part-based pedestrian detection method which divides a pedestrian into parts for a moving vehicle with one camera and one near-infrared lighting projector. The confidence of detected parts can be enhanced by analyzing the spatial relationship between every pair of parts, and the overall pedestrian detection result is refined by a block-based segmentation method. Zhao et al. [64] proposed a robust approach utilizing the shape distribution histogram (SDH) feature and the modified sparse representation classification (MSRC) for

Table 5.2: The summarization of infrared pedestrian detection methods.

	Feature Extractor	Feature Classifier	Highlight	Year
Nanda et al. [40]	Templates	Bayesian	Probabilistic templates	2003
Papageorgiou et al. [10]	Sobel	AdaBoost	Template-based method	2005
Suard et al. [49]	HOG	SVM	Stereo infrared application	2006
Zhang et al. [59]	Edgelets/HOG	AdaBoost/SVM	Experimental analysis	2007
Lee et al. [64]	HOG	SVM	Near-infrared application	2015
Zhao et al. [64]	SDH	MSRC	MSRC classifier	2015
Biswas et al. [3]	LSK	SVM	LSK descriptors	2017
Kim et al. [29]	TIR-ACF	SVM	TIR-ACF descriptors	2018

thermal pedestrian detection. Biswas et al. [3] proposed the multidimensional templates based on local steering kernel (LSK) descriptors to improve the pedestrian detection accuracies in low resolution and noisy thermal images. Kim et al. [29] presented a new thermal infrared radiometry aggregated channel feature (TIR-ACF) to detect pedestrians in the far thermal images at night.

We summarize the above-mentioned infrared pedestrian detection methods in Table 5.2. However, false detections are frequently caused due to strong solar radiation and background clutters in the daytime thermal images. With the development of multimodal sensing technology, it is possible to generate more stable detection results by simultaneously capturing multimodal information (e.g., visible, thermal and depth) of the same scene, which provide complementary information about objects of interest.

5.2.3 Multimodal Pedestrian Detection

It is experimentally demonstrated that multimodal images provide complementary information about objects of interest. As a result, pedestrian detectors trained using multimodal images can generate more robust detection results than using the visible or thermal images alone.

Grassi et al. [21] proposed a novel information fusion approach to detecting pedestrians, by determining the regions of interest in the video data through a lidar sensor and a thermal camera. Yuan et al. [58] proposed a multispectral based pedestrian detection approach which employs latent variable support-vector machines (L-SVMs) to train the multispectral (visible and near-thermal) pedestrian detection model. A large-size multispectral pedestrian dataset (KAIST) is presented by Hwang et al. [26]. The KAIST dataset contains well-aligned visible and thermal image pairs with dense pedestrian annotations. The authors further developed a new multimodal aggregated feature (ACF+T+THOG) followed by the AdaBoost classifier for target classification. The ACF+T+THOG concatenate the visible features generated by

Table 5.3: The summarization of multimodal pedestrian detection methods.

	Feature Extractor	Feature Classifier	Highlight	Year
Grassi et al. [21]	Invariant vectors	SVM	Invariant feature extraction	2011
Yuan et al. [58]	HOG	L-SVM	NIR-RGB application	2015
Hwang et al. [26]	ACF+T+THOG	AdaBoost	ACF+T+THOG descriptors	2015
Wagner et al. [53]	CNN	FC	Two-stream R-CNN	2016
Liu et al. [28]	CNN	FC	Two-stream Faster R-CNN	2016
Xu et al. [57]	CNN	FC	Cross-modal representations	2017
König et al. [30]	CNN	FC+BDT	Two-stream RPN	2017

ACF [12] and the thermal one generated by T+THOG, which contains the thermal image intensity (T) and the thermal HOG [9] features (THOG).

Recently, DNN based approaches for visible pedestrian detection have been widely adopted to design the multimodal pedestrian detectors. Wagner et al. [53] proposed the first application of DNN for multimodal pedestrian detection. The detections in [26] are considered as proposals, which are classified with a two-stream R-CNN [19] applying concatenation fusion in the late stage. The authors further evaluated the performance of architectures with different fusion stages, and the optimal architecture is in the late fusion stage. Liu et al. [28] investigated how to adopt the Faster R-CNN [45] model for the task of multimodal pedestrian detection and designed four different fusion architectures in which two-stream networks are integrated at different stages. Experimental results showed that the optimal architecture is the Halfway Fusion model which merges two-stream networks at a high-level convolutional stage. Xu et al. [57] designed a method to learn and transfer cross-modal deep representations in a supervised manner for robust pedestrian detection against bad illumination conditions. However, this method is based on information of visible channel only (during the testing stage) and its performance is not comparable with ones based multispectral data (e.g., Halfway Fusion model [28]) König et al. [30] modified the visible pedestrian detector RPN+BDT [60] to build Fusion RPN+BDT architecture for multimodal pedestrian detection. The Fusion RPN concatenates the two-stream RPN on the high-level convolutional stage and achieves the state-of-the-art performance on KAIST multimodal dataset.

We summarize the mentioned multimodal pedestrian detection methods in Table 5.3. It is worth it to mention in this chapter that our approach is definitely different from the above methods in two aspects. Firstly, a number of feature fusion strategies for multimodal pedestrian detection is firstly evaluated comparing with other basic fusion schemes. Secondly, we make use of semantic segmentation information as supervision for multimodal feature learning.

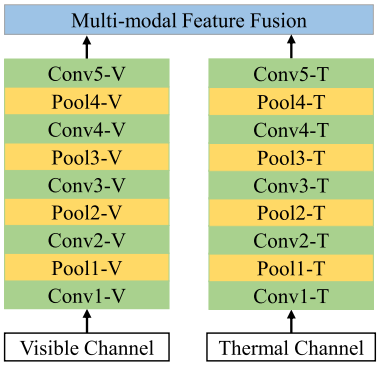


Figure 5.4 : The architecture of two-stream deep convolutional neural networks (TDCNNs) for multimodal feature learning and fusion. Please note that convolutional layers are represented by green boxes, pooling layers are represented by yellow boxes, and the feature fusion layer is represented by blue box. This figure is best seen in color.

5.3 Proposed Method

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

5.3.1 Multimodal Feature Learning/Fusion

The architecture of two-stream deep convolutional neural networks (TDCNNs) for multimodal feature learning and fusion is illustrated in Fig. 5.4. TDCNN learn semantic feature descriptions in the visible and thermal channels individually. The visible stream of TDCNN contains five convolutional layers (from Conv1-V to Conv5-V) and pooling ones (from Pool1-V to Pool4-V), which is the same as the thermal stream of TDCNN (from Conv1-T to Conv5-T and from Pool1-V to Pool4-V). Each stream of TDCNN takes the first five convolutional layers (from Conv1 to Conv5) and pooling ones (from Pool1 to Pool4) in VGG-16 as the backbone.

Previous researches revealed that feature fusion at a late stage can generate semantic feature maps which are appropriate for complex high-level detection tasks [53,28,30]. On the basis of this conclusion, the multimodal feature fusion layer is deployed after the Conv5-V and Conv5-T layers in TDCNN to combine the feature maps generated by the visible and thermal streams. The multimodal feature maps are generated as

$$\mathbf{m} = \mathcal{F}(\mathbf{v}, \mathbf{t}), \tag{5.1}$$

where $\mathcal{F}$ is the feature fusion function, $\mathbf{v} \in \mathbb{R}^{B_v \times C_v \times H_v \times W_v}$ and $\mathbf{t} \in \mathbb{R}^{B_t \times C_t \times H_t \times W_t}$ are the feature maps generated by the visible and thermal streams, respectively, and $\mathbf{m} \in \mathbb{R}^{B_m \times C_m \times H_m \times W_m}$ is the multimodal feature maps. It should be noted that B, C, H and W denote the number of batch size, channels, heights and widths of the respective feature maps. Considering that the feature maps generated by the visible and thermal streams are the same size, we set $B_t = B_c$, $C_v = C_t$, $W_v = W_t$, $H_v = H_t$. Three different feature fusion functions are considered: concatenation ($\mathcal{F}^{\text{concat}}$), maximization ($\mathcal{F}^{\max}$) and summation ($\mathcal{F}^{\text{sum}}$).

Concatenation fusion. Concatenation function $\mathcal{F}^{\text{concat}}$ is the most widely used operation to integrate feature maps generated in visible and thermal channels [53,28,30]. $\mathbf{m}^{\text{concat}} = \mathcal{F}^{\text{concat}}(\mathbf{v}, \mathbf{t})$ stacks $\mathbf{v}$ and $\mathbf{t}$ across the individual channels c as

$$\begin{cases} \mathbf{m}^{\text{concat}}_{B,2c-1,H,W} = \mathbf{v}_{B,c,H,W}, \\ \mathbf{m}^{\text{concat}}_{B,2c,H,W} = \mathbf{t}_{B,c,H,W}, \end{cases} \tag{5.2}$$

where $c \in (1, 2, ..., C)$ and $\mathbf{m} \in \mathbb{R}^{B \times 2C \times H \times W}$. Considering that the feature maps generated in visible and thermal streams are directly stacked across feature channels using concatenation function, the correlation of features generated in visible and thermal streams will be further learned in subsequent convolutional layers.

Maximization fusion. $\mathbf{m}^{\max} = \mathcal{F}^{\max}(\mathbf{v}, \mathbf{t})$ takes the maximum of $\mathbf{v}$ and $\mathbf{t}$ at the same spatial locations h, w as

$$\mathbf{m}^{\max}_{B,C,h,w} = \max\{\mathbf{v}_{B,C,h,w}, \mathbf{t}_{B,C,h,w}\}, \tag{5.3}$$

where $h \in (1, 2, ..., H)$, $w \in (1, 2, ..., W)$ and $\mathbf{m} \in \mathbb{R}^{B \times C \times H \times W}$. The maximization fusion function is utilized to select the features which are most distinct in either the visible or thermal streams.

Summation fusion. $\mathbf{m}^{\text{sum}} = \mathcal{F}^{\text{sum}}(\mathbf{p}, \mathbf{q})$ computes the summation of $\mathbf{v}$ and $\mathbf{t}$ at the same spatial locations h, w as

$$\mathbf{m}^{\max}_{B,C,h,w} = \mathbf{v}_{B,C,h,w} + \mathbf{t}_{B,C,h,w}, \tag{5.4}$$

where $h \in (1, 2, ..., H)$, $w \in (1, 2, ..., W)$ and $\mathbf{m} \in \mathbb{R}^{B \times C \times H \times W}$. The summation fusion function is utilized to integrate the feature maps generated in visible and thermal streams using equal weighting scheme. Thus, multimodal feature maps will be stronger by combining the weak features generated in the visible and thermal streams. The performances of multimodal pedestrian detectors utilizing these three fusion functions ($\mathcal{F}^{\text{concat}}$, $\mathcal{F}^{\max}$, $\mathcal{F}^{\text{sum}}$) are comparatively evaluated in Sect. 5.4.3.

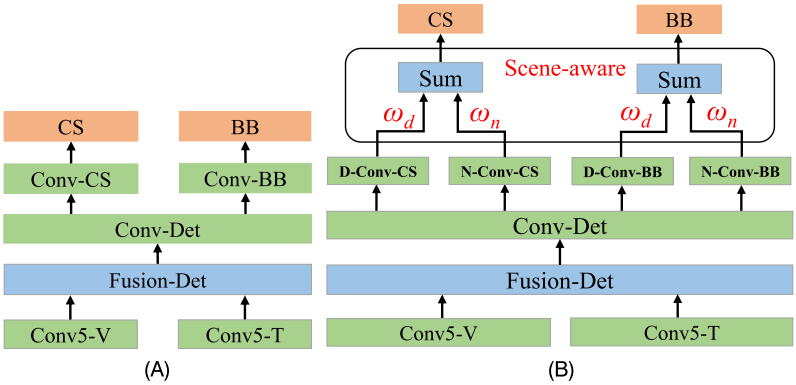


Figure 5.5 : Comparison of the baseline (A) and scene-aware (B) architectures. Please note that ω_d and ω_n are the computed scene-aware weighting parameters, convolutional layers are represented by green boxes, fusion layers are represented by blue boxes, and output layers are represented by orange boxes. This figure is best seen in color.

5.3.2 Multimodal Pedestrian Detection

In this subsection, we design two different DNN based models with and without a scene-aware mechanism for the task of multimodal pedestrian detection. Their overall architectures are shown in Fig. 5.5.

5.3.2.1 Baseline DNN model

The baseline DNN model is designed based on the TDCNN. The region proposal network (RPN) model [60] is adopted as two-stream region proposal network (TRPN) to generate multimodal pedestrian detection results due to its superior performance for large-scale target detection. Given a pair of well-aligned multimodal images, TRPN generate numbers of bounding boxes along with predicted scores following target classification and box regression. The architecture of TRPN is shown in Fig. 5.5A.

Given the multimodal feature maps generated utilizing the fusion layer (Fusion-Det), TRPN outputs classification scores (CSs) and bounding boxes (BBs) as multimodal pedestrian detections. A 3×3 convolutional layer (Conv-Det) is designed to encode pedestrian related features from the multimodal feature maps. Attached after the Conv-Det, two sibling 1×1 convolutional layers (Conv-CS and Conv-BB) are designed to generate multimodal pedestrian detections (CS and BB). In order to train the baseline DNN model, we utilize the loss term for detection $\mathcal{L}_{Det}$ as

$$\mathcal{L}_{Det} = \sum_{i \in S} \mathcal{L}_{Cls}(c_i, \hat{c}_i) + \lambda_r \sum_{i \in S} \hat{c}_i \cdot \mathcal{L}_{Reg}(b_i, \hat{b}_i), \tag{5.5}$$

where S denotes the set of training samples, c_i denotes the computed CS, b_i denotes the predicted BB, L_{cls} denotes the loss term for classification, L_{reg} denotes the loss term for box regression, and λ_r denotes the trade-off coefficient. The training label $\hat{c}_i$ is set to 1 for a positive sample. Otherwise, we set $\hat{c}_i = 0$ for a negative sample. A training sample is considered as a positive one in the case that the maximum intersection-over-union (IoU) ratio between every ground-truth label and the sample is larger than a set threshold. Otherwise, the training sample is considered a negative one. In Eq. (5.5), the loss term for classification L_{cls} is defined as

$$\mathcal{L}_{Cls}(c, \hat{c}) = -\hat{c} \cdot \log(c) - (1 - \hat{c})\log(1 - c), \tag{5.6}$$

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

$$\mathcal{L}_{Reg}(b, \hat{b}) = \sum_{j \in \{x,y,w,h\}} smooth_{L_1}(b_j, \hat{b}_j), \tag{5.7}$$

where $b = (b_x, b_y, b_w, b_h)$ represents the parameterized coordinates of the generated bounding box, $\hat{b} = (\hat{b}_x, \hat{b}_y, \hat{b}_w, \hat{b}_h)$ represents the coordinates of the bounding box label, and $smooth_{L_1}$ represents the robust $L1$ loss function defined in [18].

5.3.2.2 Scene-aware DNN model

Pedestrian samples have significantly different multimodal characteristics in daytime and nighttime scenes as shown in Fig. 5.6. Therefore, it is reasonable to deploy multiple scene-based sub-networks to handle intra-class object variances. Based on this observation, we further present a scene-aware DNN model to improve detection performance in daytime and nighttime scenes by considering the scene information encoded in multimodal images.

For this purpose, we firstly design a simple scene prediction model to estimate the probability of being daytime scene or nighttime one. As shown in Fig. 5.7, the scene prediction network (SPN) contains one pooling layer (SP-Pool), three continuous fully-connected layers (SP-FC1, SP-FC2, SP-FC3), and the classification layer (Soft-max). Each pair of multimodal images are entered into the first five convolutional layers and four pooling ones of TRPN to extract feature maps in individual visible and thermal channels. The two-stream feature maps are integrated utilizing the concatenate fusion layer (Concat). Inspired by the spatial pyramid pooling layer which can resize the feature maps to the same spatial resolution [24], the pooling layer SP-Pool resizes the multimodal feature maps to a fixed spatial size of 7×7 using symmetric bilinear interpolation from the nearest neighbors. Attached after the SP-FC3 is Soft-max, which is the classification layer of the scene prediction model. The outputs of Soft-max are ω_d and $\omega_n = (1 - \omega_d)$, which compute the probability of being day scene or night one. We define the error term of scene prediction $\mathcal{L}_{SP}$ as

$$\mathcal{L}_{SP} = -\hat{\omega}_d \cdot \log(\omega_d) - \hat{\omega}_n \cdot \log(\omega_n), \tag{5.8}$$

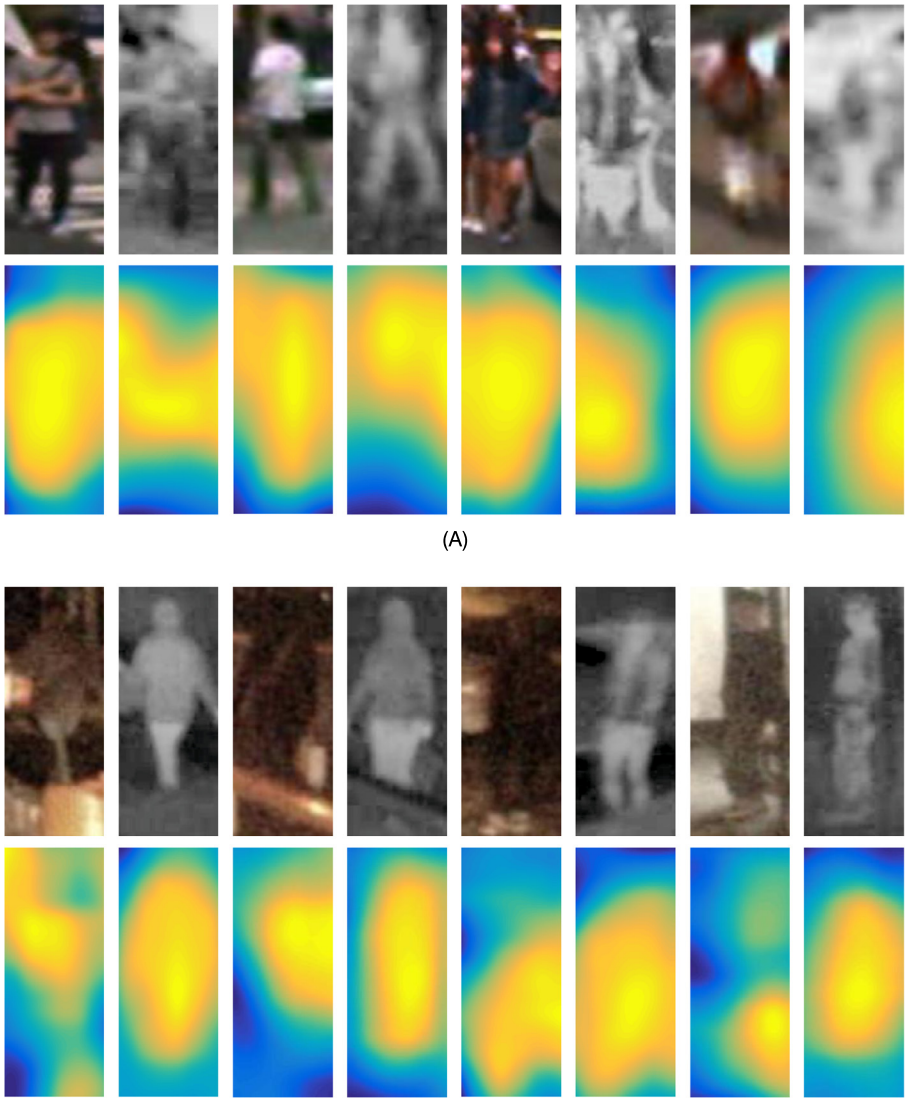


Figure 5.6 : Characteristics of multimodal pedestrian samples captured in (A) daytime and (B) nighttime scenes. The first rows display the multimodal images of pedestrian samples. The second rows illustrate the visualization of feature maps in the corresponding multimodal images. The feature maps are generated utilizing the RPN [60] well trained in their corresponding channels. It is observed that multimodal pedestrian samples have significantly different human-related characteristics under daytime and nighttime scenes.

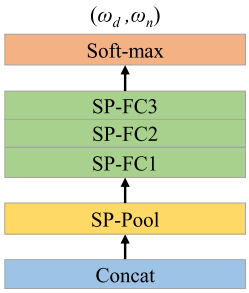


Figure 5.7 : Architecture of the scene prediction network (SPN). Please note that convolutional and fully-connected layers are represented by green boxes, pooling layers are represented by yellow boxes, fusion layers are represented by blue boxes, classification layer is represented by an orange box. This figure is best seen in color.

where ω_d and $\omega_n = (1 - \omega_d)$ are the predicted scene weights for daytime and nighttime, $\hat{\omega}_d$ and $\hat{\omega}_n = (1 - \hat{\omega}_d)$ are the scene labels. The $\hat{\omega}_d$ is set to 1 when the scene labels are annotated as daytime. Otherwise, we set $\hat{\omega}_d = 0$.

We incorporate scene information into the baseline DNN model in order to generate more accurate and robust results for various scene conditions. Specifically, the scene-aware two-stream region proposal networks (STRPNs) consist of four sub-networks (D-Conv-CS, N-Conv-CS, D-Conv-BB, and N-Conv-BB) to generate scene-aware detection results (CS and BB) as shown in Fig. 5.5B. D-Conv-CS and N-Conv-CS generate feature maps for classification under daytime and nighttime scenes, respectively. The outputs of D-Conv-CS and N-Conv-CS are combined utilizing the weights computed in the SPN model to produce the scene-ware classification scores (CSs). D-Conv-BB and N-Conv-BB generate feature maps for box regression in day and night scene conditions, respectively. The outputs of D-Conv-BB and N-Conv-BB are integrated using the weights computed in the SPN model to calculate the scene-ware bounding boxes (BBs). The loss term for detection L_{det} is also defined based on Eq. (5.5), while c_i^s is computed as the weighted sum of classification score in daytime scene c_i^d and one in nighttime scene c_i^n as

$$c_i^s = \omega_d \cdot c_i^d + \omega_n \cdot c_i^n, \tag{5.9}$$

and b_i^s is the scene-aware weighted combination of b_i^d and b_i^n, which are calculated by D-Conv-BB and N-Conv-BB sub-networks, respectively, as

$$b_i^s = \omega_d \cdot b_i^d + \omega_n \cdot b_i^n. \tag{5.10}$$

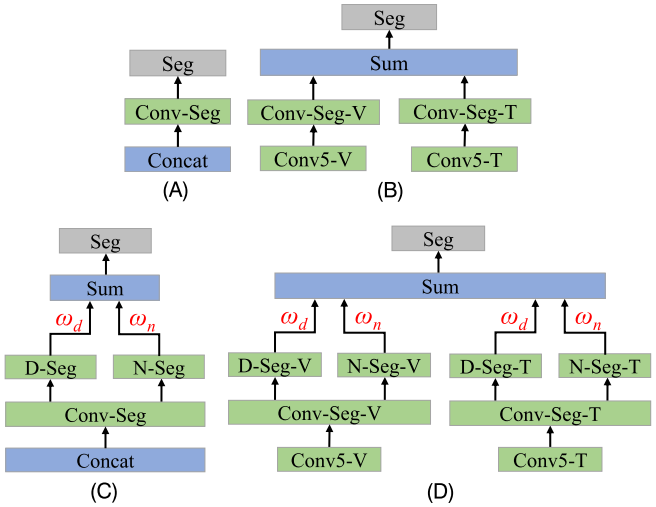


Figure 5.8 : The comparison of MS-F (A), MS (B), SMS-F (C) and SMS (D) architectures. Please note that ω_d and ω_n are the computed scene-aware weighting parameters, convolutional layers are represented by green boxes, fusion layers are represented by blue boxes, and output layers are represented by gray boxes. This figure is best seen in color.

5.3.3 Multimodal Segmentation Supervision

According to some recent research work [25,4], the performance of anchor-based object detectors can be improved using the information of semantic segmentation as a strong cue. The underlying reason is that box-based segmentation masks are able to supply additional effective supervision to make the feature maps in the shared layers more distinctive for the downstream object detectors. We integrate the segmentation supervision scheme with TRPN and STRPN to obtain more accurate multimodal pedestrian detection results.

Given feature maps generated from two-stream visible and thermal channels, different segmentation supervision results will be achieved by integrating the feature maps in different levels (feature-level and decision-level). In order to explore the optimal fusion architecture for the multimodal segmentation supervision task, we develop two different multimodal segmentation supervision architectures, denoted as feature-level multimodal segmentation supervision (MS-F) and decision-level multimodal segmentation supervision (MS). As illustrated in Fig. 5.8A, the MS-F contains a concatenate fusion layer (Concat), a 3×3 convolutional layer (Conv-Seg) and the output layer (Seg). Each pair of multimodal images are entered into the first five convolutional layers and four pooling ones of TRPN to extract feature maps in individual visible and thermal channels. The two-stream feature maps are integrated utilizing the Concat layer. The MS-F generate segmentation prediction (Seg) as a supervision

to make the two-stream feature maps more distinctive for multimodal pedestrian detector. In comparison, as illustrated in Fig. 5.8B, the MS utilizes two 3×3 convolutional layers (Conv-Seg-V and Conv-Seg-T) to generate visible and thermal segmentation prediction as the supervision to make the visible and thermal feature maps more distinctive for multimodal pedestrian detector. The segmentation outputs of the MS is a combination of Conv-Seg-V and Conv-Seg-T.

It is reasonable to explore whether the performance of segmentation supervision able to be improved by integrating the scene information. Two different scene-aware multimodal segmentation supervision architectures (SMS-F and SMS) are developed based on the architectures of multimodal segmentation supervision models (MS-F and MS). As shown in Fig. 5.8C–D, multimodal segmentation outputs are generated using daytime and nighttime segmentation sub-networks (D-Seg and N-Seg) applying the scene-aware weighting mechanism. It should be noted that SMS-F consists of two sub-networks, while SMS consists of four sub-networks. The performance of these four different multimodal segmentation supervision architectures (SM-F, SM, SMS-F, and SMS) are evaluated in Sect. 5.4.

The loss term of segmentation supervision is defined as

$$L_{Seg} = \sum_{i \in C} \sum_{j \in B} [-\hat{s}_j \cdot \log(s_{ij}^s) - (1 - \hat{s}_j) \cdot \log(1 - s_{ij}^s)], \tag{5.11}$$

where s_{ij}^s is the segmentation prediction, C is the segmentation streams (MS-F and SMS-F consist of one multimodal stream while MS and SMS consist of one visible stream and one thermal stream), B are training samples for box-based segmentation supervision. We set $\hat{s}_j = 1$ if the sample is located in the region within any bounding box of ground truth. Otherwise, we set $\hat{s}_j = 0$. As for the scene-aware multimodal segmentation supervision architectures SMS-F and SMS, the scene-aware segmentation prediction s_{ij}^s is computed by the scene-aware weighted combination of daytime and nighttime segmentation prediction s_{ij}^d and s_{ij}^n, respectively, as

$$s_{ij}^s = \omega_d \cdot s_{ij}^d + \omega_n \cdot s_{ij}^n. \tag{5.12}$$

The loss terms defined in Eqs. (5.7), (5.2), and (5.11) are combined to conduct multitask learning of scene-aware pedestrian detection and segmentation supervision. The final multitask loss function is defined as

$$\mathcal{L} = \mathcal{L}_{Det} + \lambda_{sp} \cdot \mathcal{L}_{SP} + \lambda_{seg} \cdot \mathcal{L}_{Seg} \tag{5.13}$$

where λ_{sp} and λ_{seg} are the trade-off coefficient of loss term $\mathcal{L}_{SP}$ and $\mathcal{L}_{Seg}$, respectively.

5.4 Experimental Results and Discussion

5.4.1 Dataset and Evaluation Metric

The public KAIST multimodal pedestrian benchmark dataset [26] are utilized to evaluate our proposed methods. The KAIST dataset consists of 50,172 well-aligned visible-thermal image pairs with 13,853 pedestrian annotations in training set and 2252 image pairs with 1356 pedestrian annotations in the testing set. All the images in KAIST dataset are captured under daytime and nighttime lighting conditions. According to the related research work [28,30], we sample the training images every two frames. The original annotations under the "reasonable" evaluation setting (pedestrian instances are larger than 55 pixels height and over 50% visibility) are used for quantitative evaluation. Considering that many problematic annotations (e.g., missed pedestrian targets and inaccurate bounding boxes) are existed in the original KAIST testing dataset, we also utilize the improved annotations manually labeled by Liu et al. [37] for quantitative evaluation.

The log-average miss rate (MR) proposed by Dollar et al. [12] is used as the standard evaluation metric to evaluate the quantitative performance of multimodal pedestrian detectors. According to the related research work [26,28,30], a detection is considered as a true positive if the IoU ratio between the bounding boxes of the detection and any ground-truth label is greater than 50% [26,28,30]). Unmatched detections and unmatched ground-truth labels are considered as false positives and false negatives, respectively. The MR is computed by averaging miss rate at nine false positives per image (FPPI) rates which are evenly spaced in log-space from the range 10^{-2} to 10^{0} [26,28,30].

5.4.2 Implementation Details

According to the related research work [60,28,30], the image-centric training framework is applied to generate mini-batches and each mini-batch contains one pair of multimodal images and 120 randomly selected anchor boxes. In order to make the right balance between foreground and background training samples, we set the ratio of positive and negative anchor boxes to 1:5 in each mini-batch. The first five convolutional layers in each stream of TDCNN (from Conv1-V to Conv5-V in the visible stream and from Conv1-T to Conv5-T in the thermal one) are initialized using the weights and biases of VGG-16 [48] DNN pretrained on the ImageNet dataset [46] in parallel. Following the RPN designed by Zhang et al. [60], other convolutional layers are initialized with a standard Gaussian distribution. The number of channels in SP-FC1, SP-FC2, SP-FC3 are empirically set to 512, 64, 2, respectively. We set $\lambda_r = 5$ in Eq. (5.5) following [60] and $\lambda_s = 1$ in Eq. (5.13) according to the visible segmentation supervision method proposed by Brazil et al. [4]. All of multimodal pedestrian detectors are trained using the Caffe deep learning framework [27] with stochastic gradient descent [65]

Table 5.4: The quantitative comparison (MR [12]) of TRPN using different feature fusion functions on the original KAIST testing dataset [26].

Model	All-day (%)	Daytime (%)	Nighttime (%)
TRPN-Concat	32.60	33.80	30.53
TRPN-Max	31.54	32.66	29.43
TRPN-Sum	**30.49**	**31.27**	**28.29**

Table 5.5: The quantitative comparison (MR [12]) of TRPN using different feature fusion functions on the improved KAIST testing dataset [37].

Model	All-day (%)	Daytime (%)	Nighttime (%)
TRPN-Concat	21.12	20.66	22.81
TRPN-Max	19.90	18.45	23.29
TRPN-Sum	**19.45**	**17.94**	**22.37**

for the first two epochs in the learning rate (LR) of 0.001 and one more epoch in the LR of 0.0001. We clip gradients when the L2 norm of the back-propagation gradient is larger than 10 to avoid exploding gradient problems [44].

5.4.3 Evaluation of Multimodal Feature Fusion

In order to explore the optimal feature fusion scheme for multimodal pedestrian detection, we compare the TRPN with different feature fusion layers. As exposed in Sect. 5.3.1, three different feature fusion functions (concatenation, maximization, and summation) are utilized to build the three different TRPN models (TRPN-Concat, TRPN-Max, and TRPN-Sum). The TRPN models are trained using the loss term of detection $\mathcal{L}_{Det}$. The detection performances of TRPN-Concat, TRPN-Max, and TRPN-Sum are quantitatively compared in Table 5.4 and Table 5.5 using the log-average miss rate (MR) proposed by Dollar et al. [12]. In addition, qualitative comparisons of the detection performances of the three different TRPN models are conducted by showing some detection results in Fig. 5.9.

We can observe that the multimodal pedestrian detection performance is affected by the functions of feature fusion. Our experimental comparisons show that the performance of TRPN-Sum is better than the TRPN-Concat and TRPN-Max, resulting in lower MR on both original and improved KAIST testing datasets. Surprisingly, the widely used fusion function (concatenation) to integrate feature maps generated in visible and thermal channels [53,28,30] performs worst among the three different feature fusion functions. As described in Sect. 5.3.1, the feature maps generated in visible and thermal streams are directly stacked across feature

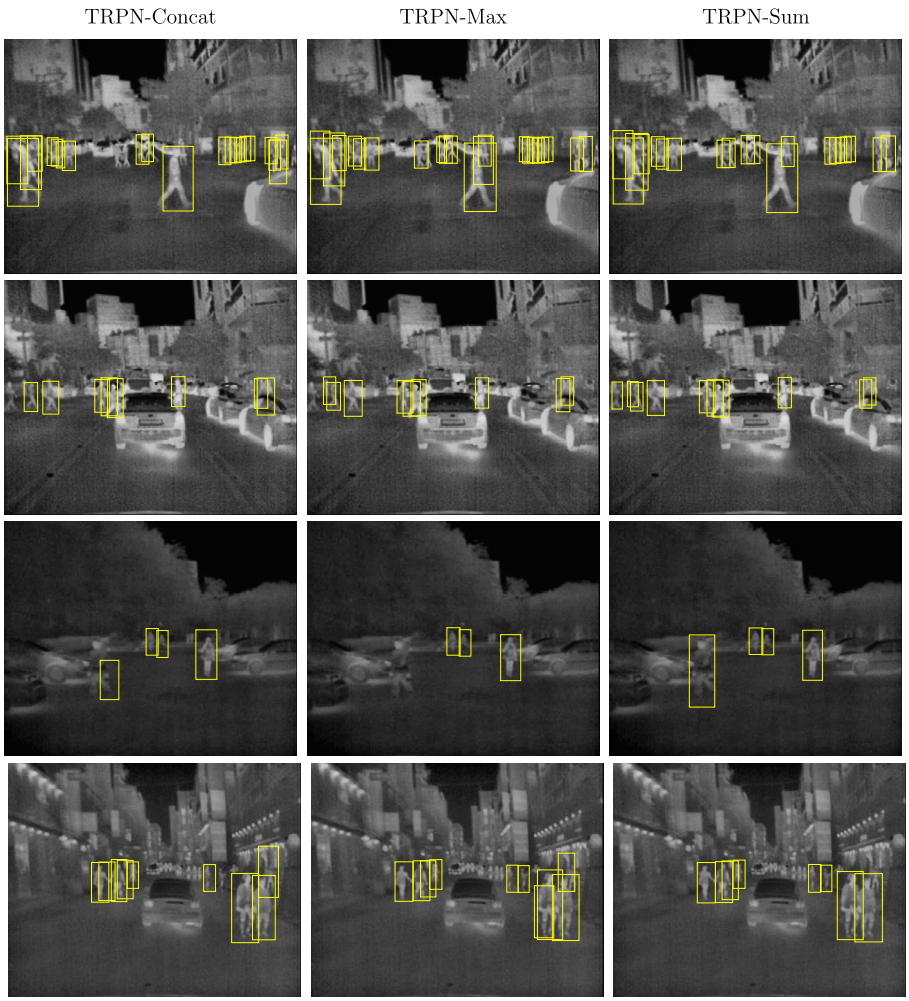


Figure 5.9 : Qualitative comparison of multimodal pedestrian detections of TRPN-Concat, TRPN-Max, and TRPN-Sum (only displayed in the thermal images). Note that yellow BB in solid line represents the pedestrian detections. This figure is best seen in color.

channels using concatenation function. Thus, the correlation of features generated in visible and thermal streams will be further learned using the Conv-Det layer in the TRPN-Concat model. As shown in Fig. 5.6, such correlation is different for daytime and nighttime multimodal pedestrian characteristics. It is difficult to build up the correlation between visible and thermal feature maps using a simple convolutional encoder (Conv-Det). On comparison, TRPN-Max and TRPN-Sum models achieve better detection results by using either maximum or summation function to define the correlation between visible and thermal feature maps. It

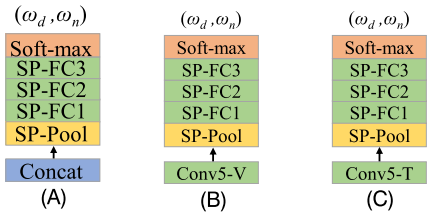


Figure 5.10 : The architecture of SPN (A), SPN-V (B) and SPN-T (C). Please note that convolutional and fully-connected layers are represented by green boxes, pooling layers are represented by yellow boxes, fusion layers are represented by blue boxes, classification layer is represented by an orange box. This figure is best seen in color.

should be noticed that the TRPN-Sum can successfully detect the pedestrian targets whose human-related characteristics are not distinct in both visible and thermal images. Different from the maximum function that selects the most distinct features in the visible and thermal streams, the summation function is able to integrate the weak features generated in the visible and thermal streams to generate a stronger multimodal one to facilitate more accurate pedestrian detection results.

5.4.4 Evaluation of Multimodal Pedestrian Detection Networks

The scene prediction network (SPN) is essential and fundamental in our proposed scene-aware multimodal pedestrian detection networks. We first evaluate whether the scene prediction networks (SPNs) can accurately compute the scene weights ω_d and $\omega_n = (1 - \omega_d)$, which supply vital information to integrate the scene-aware sub-networks. The KAIST testing dataset is utilized to evaluate the performance of SPN. Please note that the KAIST testing dataset consists of 1455 image pairs captured in daytime scene conditions and 797 in nighttime. Given a pair of multimodal images, the SPN computes a daytime scene weight ω_d. The scene condition will be predicted correctly if $\omega_d > 0.5$ during daytime or $\omega_d < 0.5$ during nighttime. In order to investigate whether visible images or thermal ones can supply the most reliable information to predict the scene conditions, the performances of scene prediction utilizing the feature maps generated only in the visible stream (SPN-V) or thermal one (SPN-T) are evaluated individually. We illustrate the architectures of SPN-V, SPN-T, and SPN in Fig. 5.10. The prediction results of these three scene prediction networks are compared in Table 5.6.

We can observe that the SPN-V are able to generate accurate prediction of scene conditions for daytime scenes (97.94%) and nighttime ones (97.11%). The underlying reason is that the visible images display different brightness in daytime and nighttime scene conditions. The scene prediction performance using SPN-T are not comparable with SPN-T (daytime

Table 5.6: Accuracy of scene prediction utilizing SPN-V, SPN-T, and SPN.

	Daytime (%)	Nighttime (%)
SPN-V	97.94	97.11
SPN-T	93.13	94.48
SPN	**98.35**	**99.75**

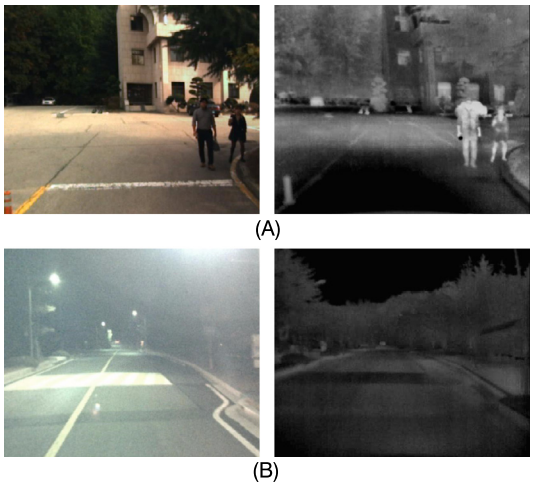


Figure 5.11 : Samples of false scene prediction results during (A) daytime and (B) nighttime. When the illumination condition is not good during daytime or street lights provide good illumination during nighttime, the SPN model will generate false prediction results.

93.13% vs. 97.94% and nighttime 94.48% vs. 97.11%). This is a reasonable result as the relative temperature between the pedestrian, and surrounding environment is not very different in daytime and nighttime scenes. Although thermal images cannot be utilized for scene prediction individually, it supplies supplementary information for the visible images to boost the performance of scene prediction. By integrating the complementary information of visible and thermal images, the SPN can generate more accurate prediction of scene conditions compared with SPN-V (daytime 98.35% vs. 97.94% and nighttime 99.75% vs. 97.11%). In addition, we show some fail cases using SPN in Fig. 5.11. The SPN may generate false scene predictions when the brightness is very low in daytime or the street lights provide high illumination in nighttime. Overall, the scene conditions can be accurately predicted utilizing the SPN by integrating visible and thermal features.

We further evaluate whether the performance of multimodal pedestrian detector can be boosted by applying our proposed scene-aware weighting mechanism by comparing the performance of the baseline TRPN and scene-aware TRPN (STRPN) models. The loss term for

Table 5.7: The calculated MRs of different TRPN and STRPN models on the original KAIST testing dataset [26].

	All-day (%)	Daytime (%)	Nighttime (%)
TRPN-Sum	30.49	31.27	28.29
STRPN-Sum	30.50	30.89	28.91
TRPN-Max	31.54	32.66	29.43
STRPN-Max	30.70	31.40	28.32
TRPN-Concat	32.60	33.80	30.53
STRPN-Concat	**29.62**	**30.30**	**26.88**

Table 5.8: The calculated MRs of Tdifferent TRPN and STRPN models on the improved KAIST testing dataset [37].

	All-day (%)	Daytime (%)	Nighttime (%)
TRPN-Sum	19.45	17.94	22.37
STRPN-Sum	17.67	17.76	18.16
TRPN-Maxn	19.90	18.45	23.29
STRPN-Max	18.40	18.15	18.69
TRPN-Concat	21.12	20.66	22.81
STRPN-Concat	**17.58**	**17.36**	**17.79**

scene prediction defined in Eq. (5.8) and loss term for detection defined in Eq. (5.2) are combined to jointly train SPN and STRPN. Although the TRPN model is an effective framework to integrate visible and thermal streams for robust pedestrian detection, it cannot differentiate human-related features under daytime and nighttime scenes when generating detection results. With comparison, STRPN utilize the scene-aware weighting mechanism to adaptively integrate multiple scene-aware sub-networks (D-Conv-CS, N-Conv-CS, and D-Conv-BB, N-Conv-BB) to generate detection results (CS and BB).

As shown in Tables 5.7 and 5.8, the quantitative comparative results of TRPN and STRPN are conducted using the log-average miss rate (MR) as the evaluation protocols. Applying the scene-aware weighting mechanism, the detection performances of the STRPN are significantly improved comparing with TRPN for all scenes on both original and improved KAIST testing datasets. We can observe that integrating the Concat fusion scheme with a global scene-aware mechanism, instead of a single Conv-Det layer, facilitates better learning of both human-related features and correlation between visible and thermal feature maps. It also worth mentioning that such performance gain (TRPN-Concat 32.60% MR v.s. STRPN-Concat 29.62% MR on the original KAIST testing dataset [26] and TRPN-Concat 21.12% MR v.s. STRPN-Concat 17.58% MR on the improved KAIST testing dataset [37]) is achieved at the cost of small computational overhead. Measured on a single Titan X GPU, the STRPN

Table 5.9: Detection results (MR) of STRPN, STRPN+MS-F, STRPN+MS, STRPN+SMS-F, and STRPN+SMS on the original KAIST testing dataset [26].

	All-day (%)	Daytime (%)	Nighttime (%)
STRPN	29.62	30.30	26.88
STRPN+MS-F	29.17	29.92	26.96
STRPN+MS	27.21	27.56	25.57
STRPN+SMS-F	28.51	28.98	27.52
STRPN+SMS	**26.37**	**27.29**	**24.41**

Table 5.10: Detection results (MR) of STRPN, STRPN+MS-F, STRPN+MS, STRPN+SMS-F, and STRPN+SMS on the improved KAIST testing dataset [26].

	All-day (%)	Daytime (%)	Nighttime (%)
STRPN	17.58	17.36	17.79
STRPN+MS-F	16.54	15.83	17.28
STRPN+MS	15.88	15.01	16.45
STRPN+SMS-F	16.41	15.17	16.91
STRPN+SMS	**14.95**	**14.67**	**15.72**

model takes 0.24 s to process a pair of multimodal images in KAIST dataset while TRPN needs 0.22 s. These experimental results show that the global scene information can be infused into scene-aware sub-networks to boost the performance of the multimodal pedestrian detector.

5.4.5 Evaluation of Multimodal Segmentation Supervision Networks

In this section, we investigate whether the performance of multimodal pedestrian detection can be improved by incorporating the segmentation supervision scheme with STRPN. As described in Sect. 5.3.3, four different multimodal segmentation supervision (MS) models including MS-F (feature-level MS), MS (decision-level MS), SMS-F (scene-aware feature-level MS) and SMS (scene-aware decision-level MS) are combined with STRPN to build STRPN, STRPN+MS-F, STRPN+MS and STRPN+SMS-F and STRPN+SMS respectively. Multimodal segmentation supervision models generate the box-based segmentation prediction and supply the supervision to make the multimodal feature maps more distinctive. The detection results (MR) of the STRPN, STRPN+MS-F, STRPN+MS and STRPN+SMS-F and STRPN+SMS are compared in Table 5.9 and Table 5.10.

Through the joint training of segmentation supervision and pedestrian detection, all the multimodal segmentation supervision models, except for STRPN+MS-F in nighttime scene,

achieve performance gains. The reason is that semantic segmentation masks provide additional supervision to facilitate the training of more sophisticated features to enable more robust detection [4]. Meanwhile, we observe that the choice of fusion architectures (feature-level or decision-level) can significantly affect the performance of multimodal pedestrian detection results. As illustrated in Tables 5.9 and 5.10, the detection results of decision-level multimodal segmentation supervision models (MS and SMS) are much better than the feature-level models (MS-F and SMS-F). The reason is that decision-level models can generate more effective supervision information by infusing visible and thermal segmentation information directly for learning human-related features for multimodal pedestrian detection. It will be our future research work to explore the optimal segmentation fusion scheme for the effective supervision of multimodal pedestrian detection. More importantly, we can observe that the performance of segmentation supervision can be significantly improved by applying the scene-aware weighting mechanism. More accurate segmentation prediction can be generated by adaptively integrating the scene-aware segmentation sub-networks. Some comparative segmentation predictions utilizing MS-F, MS, SMS-F, and SMS are shown in Fig. 5.12. The STRPN+SMS can generate more accurate segmentation predictions which supply better supervision information to facilitate the training of human-related features for multimodal pedestrian detection task.

In order to show the improvements gains achieved by different scene-aware modules, we visualize the feature maps of TRPN, STRPN, and STRPN+SMS in Fig. 5.13. It can be observed that STRPN can generate more distinctive human-related features by incorporating scene-aware weighting mechanism into TRPN for better learning of multimodal human-related features. More importantly, further improvements can be achieved by STRPN+SMS through the scene-aware segmentation modules to supervise the learning of multimodal human-related features.

5.4.6 Comparison with State-of-the-Art Multimodal Pedestrian Detection Methods

We compare the designed STRPN and STRPN+SMS models with the current state-of-the-art multimodal pedestrian detection algorithms: ACF+T+THOG [26], Halfway Fusion [28] and Fusion RPN+BDT [30]. The log–log figure of MR vs. FPPI is plotted for performance comparison in Fig. 5.14. It can be observed that our proposed STRPN+SMS achieves the best detection accuracy (26.37% MR on the original KAIST testing dataset [26] and 14.95% MR on the improved KAIST testing dataset [37]) in all-day scenes (see Fig. 5.15). The performance gain on the improved KAIST testing dataset achieves a relative improvement rate of 18% compared with the best of current state-of-the-art algorithm Fusion RPN+BDT (18.23% MR on the improved KAIST testing dataset [37]). In addition, the detection performance of our proposed STRPN is comparable with the current state-of-art models. Furthermore, some

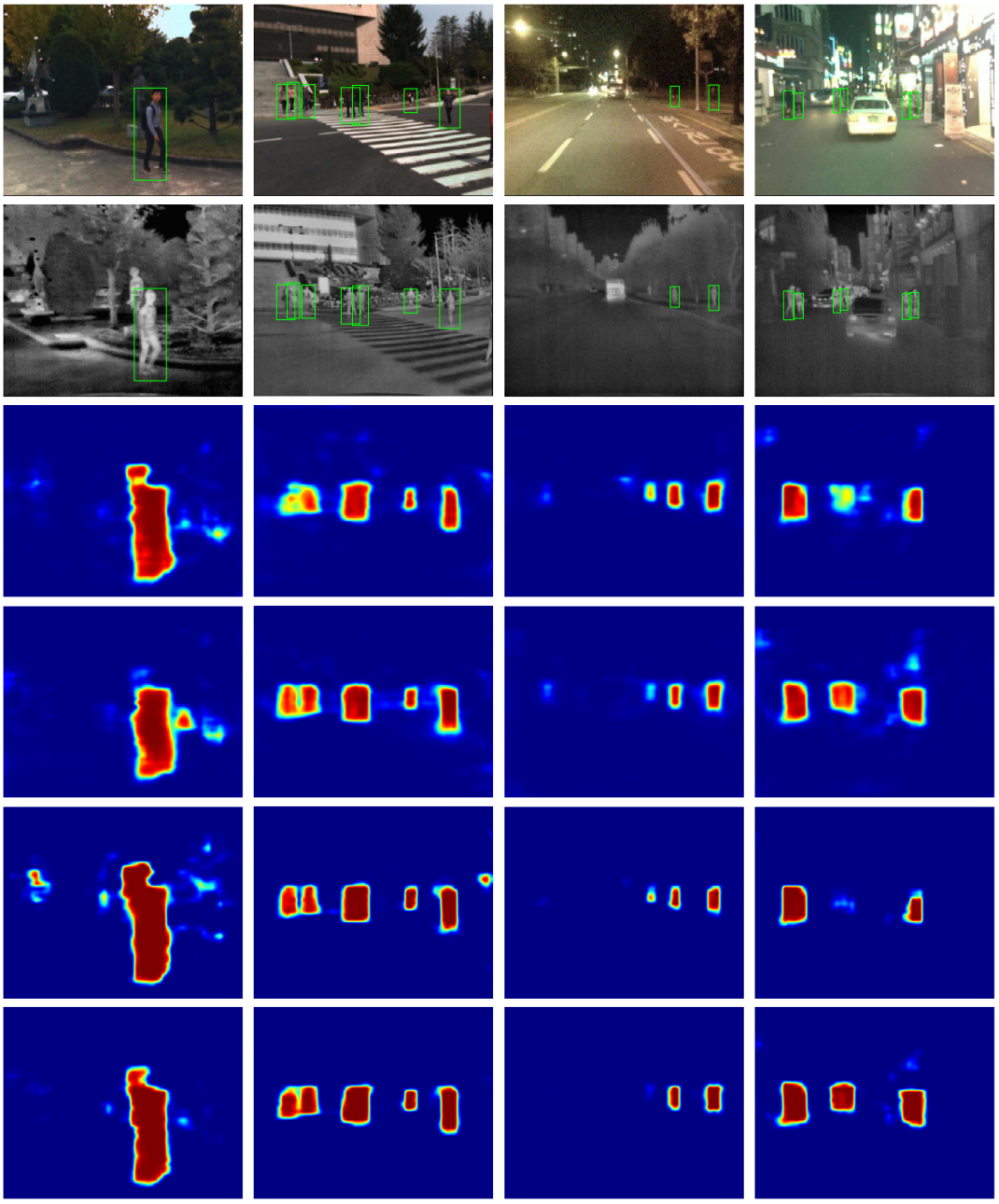
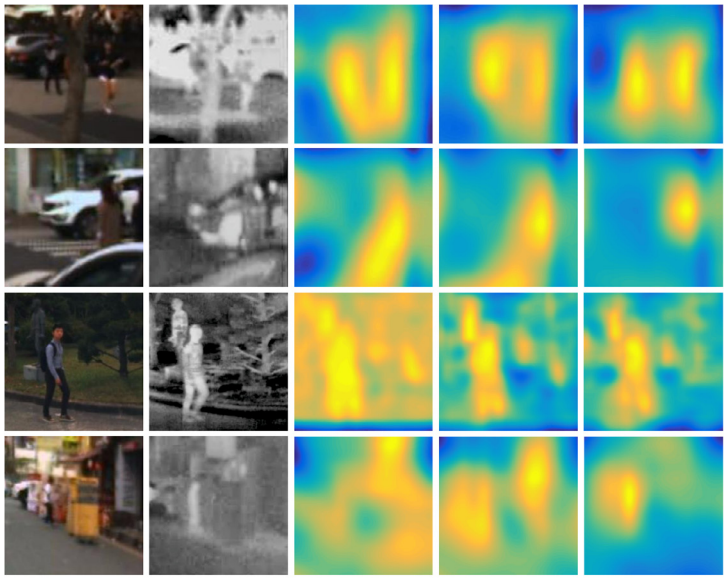
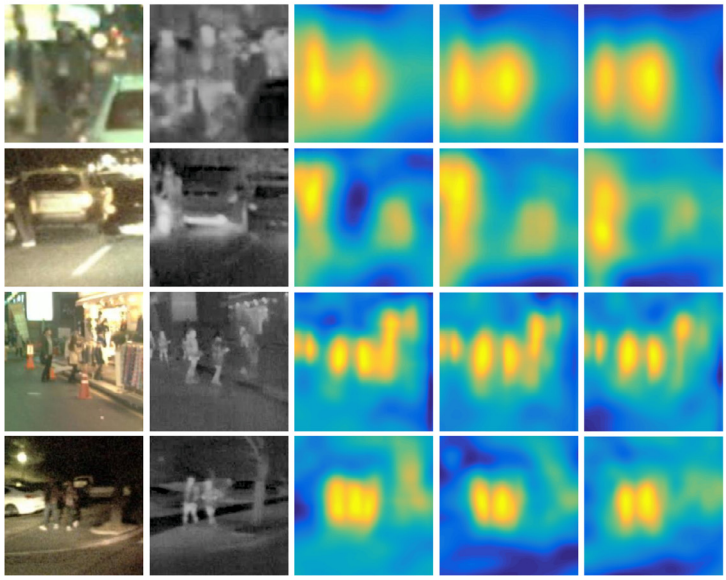


Figure 5.12 : Examples of pedestrian segmentation prediction utilizing four different multimodal segmentation supervision models. (A) Daytime. (B) Nighttime. The first two rows in (A) and (B) illustrate the visible and thermal images, respectively. Other rows in (A) and (B) illustrate the pedestrian segmentation prediction results of MS-F, MS, SMS-F, and SMS respectively. Please note that green BB presents pedestrian labels. This figure is best seen in color.



(A)



(B)

Figure 5.13 : Examples of multimodal pedestrian feature maps which are promoted by scene-aware mechanism captured in (A) daytime and (B) nighttime scenes. The first two columns in (A) and (B) show the pictures of visible and thermal pedestrian instances, respectively. The third to the fifth columns in (A) and (B) show the feature map visualizations generated from TRPN, STRPN, and STRPN+SMS respectively. It is noticed that the feature maps of a multimodal pedestrian become more distinct by using our proposed scene-aware modules (STRPN and SMS).

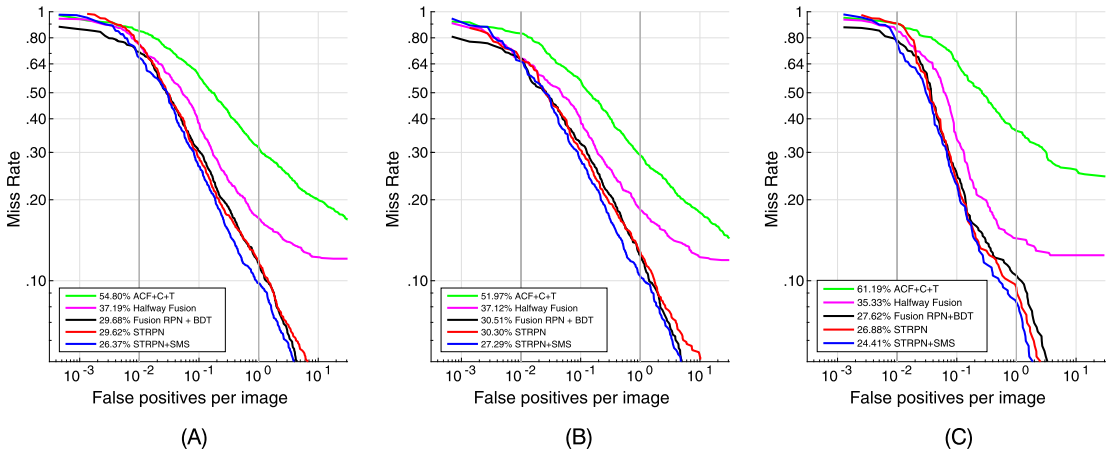


Figure 5.14 : Comparisons of the original KAIST test dataset under the reasonable evaluation setting during all-day (A), daytime (B), and nighttime (C) (legends indicate MR).

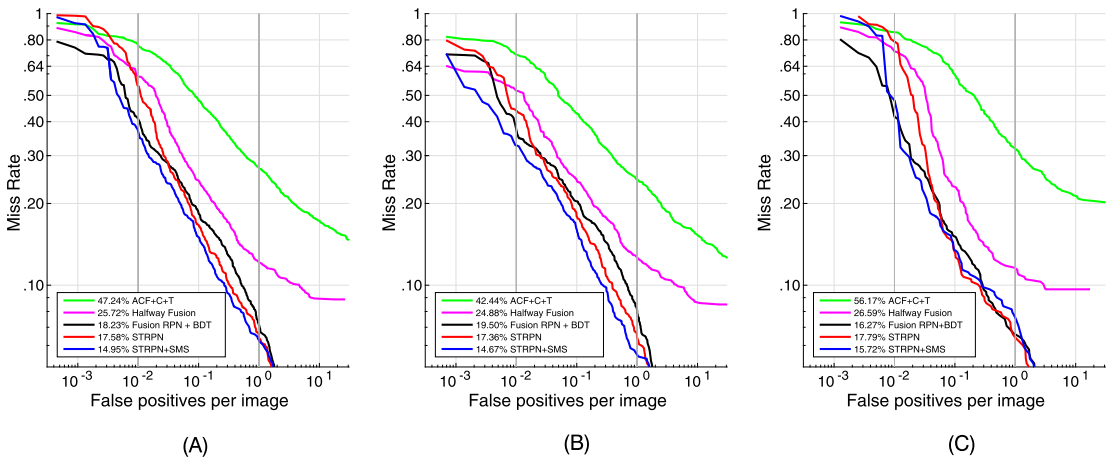


Figure 5.15 : Comparisons of the improved KAIST test dataset under the reasonable evaluation setting during all-day (A), daytime (B), and nighttime (C) (legends indicate MR).

detection results of the Fusion RPN+BDT and STRPN+SMS are visualized for qualitative comparison as shown in Fig. 5.16. We can observe that our proposed STRPN+SMS model can generate more accurate detection results in both daytime and nighttime scene conditions.

The computation efficiency of STRPN, STRPN+SMS and the current state-of-the-art methods [28,30] are illustrated in Table 5.11. Every method is executed 1000 times to compute the averaged runtime. We can observe that our proposed STRPN+SMS results in least runtime comparing with the current state-of-the-art methods [28,30] (STRPN+SMS 0.25 s vs.

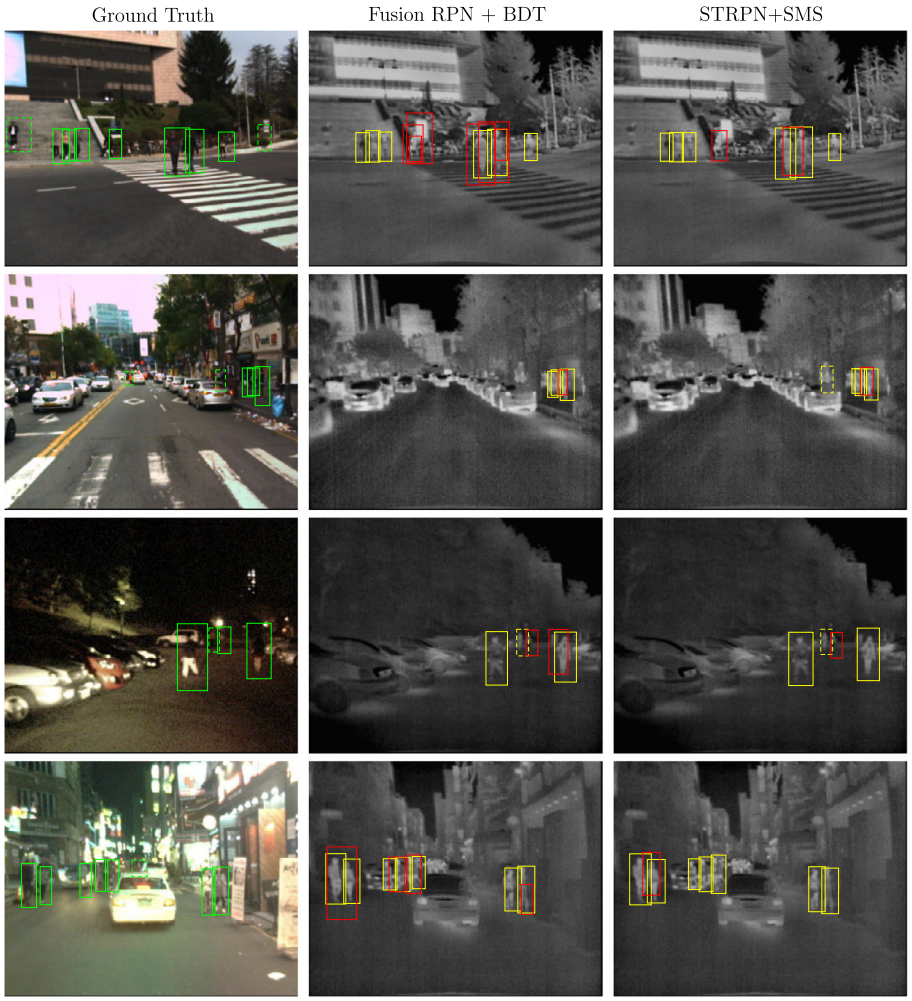


Figure 5.16 : Comparison with the current state-of-the-art multimodal pedestrian detector (Fusion RPN+BDT). The first column shows the multimodal images along with the ground truth (displayed in the visible image) and the other columns show the detection results of Fusion RPN+BDT, STRPN, and STRPN+SMS (displayed in the thermal image). Please note that green BBs in solid line show positive labels, green BBs in dashed line show ignore ones, yellow BB in solid line show true positives, and red BB show false positives. This figure is best seen in color.

Halfway Fusion 0.40 s vs. Fusion RPN+BDT 0.80 s). The reason is that both the extra Fast R-CNN module [18] utilized in the Halfway Fusion model and the boosted decision trees [14] algorithm utilized in Fusion RPN+BDT model significantly decrease the computation efficiency. It should be noticed that our proposed scene-aware architectures can significantly

Table 5.11: Comprehensive comparison of STRPN and STRPN+SMS with state-of-the-art multimodal pedestrian detectors [28,30] on the improved KAIST testing dataset. A single Titan X GPU is used to evaluate the computation efficiency. Note that DL represents deep learning and BDT represents boosted decision trees [14].

	MR (%)	Runtime (s)	Method
Halfway Fusion [28]	25.72	0.40	DL
Fusion RPN+BDT [30]	18.23	0.80	DL+BDT
TRPN	21.12	0.22	DL
STRPN	17.58	0.24	DL
STRPN+SMS	14.95	0.25	DL

boost the multimodal pedestrian detection results while only cause a small computational overhead (TRPN 0.24 s vs. STRPN 0.24 s vs. STRPN+SMS 0.25 s).

5.5 Conclusion

In this chapter, we present a new multimodal pedestrian detection method based on multi-task learning of scene-aware target detection and segmentation supervision. To achieve better learning of both human-related features and correlation between visible and thermal feature maps, we propose a feature fusion scheme utilizing the concatenation with a global weighting scene-aware mechanism. Experimental results illustrate that multimodal pedestrian detections can be improved by applying our proposed scene-aware weighting mechanism. Moreover, we design four different multimodal segmentation supervision architectures and conclude that scene-aware decision-level multimodal segmentation (SMS) module can generate the most accurate prediction as supervision for visual feature learning. Experimental evaluation on public KAIST benchmark shows that our method achieves the most accurate results with least runtime compared with the current state-of-the-art multimodal pedestrian detectors.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (No. 51605428, No. 51575486, U1664264) and DFG (German Research Foundation) YA 351/2-1.

References

[1] Rodrigo Benenson, Mohamed Omran, Jan Hosang, Bernt Schiele, Ten years of pedestrian detection, what have we learned?, in: European Conference on Computer Vision, 2014, pp. 613–627.

[2] M. Bilal, A. Khan, M.U. Karim Khan, C.M. Kyung, A low-complexity pedestrian detection framework for smart video surveillance systems, IEEE Transactions on Circuits and Systems for Video Technology 27 (10) (2017) 2260–2273.
[3] Sujoy Kumar Biswas, Peyman Milanfar, Linear support tensor machine with lsk channels: pedestrian detection in thermal infrared images, IEEE Transactions on Image Processing 26 (9) (2017) 4229–4242.
[4] Garrick Brazil, Xi Yin, Xiaoming Liu, Illuminating pedestrians via simultaneous detection & segmentation, in: IEEE International Conference on Computer Vision, IEEE, 2017.
[5] Zhaowei Cai, Quanfu Fan, Rogerio S. Feris, Nuno Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: European Conference on Computer Vision, 2016, pp. 354–370.
[6] Yanpeng Cao, Dayan Guan, Weilin Huang, Jiangxin Yang, Yanlong Cao, Yu Qiao, Pedestrian detection with unsupervised multispectral feature learning using deep neural networks, Information Fusion 46 (2019) 206–217.
[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, Bernt Schiele, The cityscapes dataset for semantic urban scene understanding, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 3213–3223.
[8] Corinna Cortes, Vladimir Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.
[9] Navneet Dalal, Bill Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 886–893.
[10] James W. Davis, Mark A. Keck, A two-stage template approach to person detection in thermal imagery, in: Application of Computer Vision, 2005, vol. 1, Seventh IEEE Workshops on, WACV/MOTIONS'05, IEEE, 2005, pp. 364–369.
[11] Piotr Dollár, Zhuowen Tu, Pietro Perona, Serge Belongie, Integral channel features, in: British Machine Vision Conference, 2009.
[12] Piotr Dollar, Christian Wojek, Bernt Schiele, Pietro Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (4) (2012) 743–761.
[13] Andreas Ess, Bastian Leibe, Luc Van Gool, Depth and appearance for mobile scene analysis, in: IEEE International Conference on Computer Vision, IEEE, 2007, pp. 1–8.
[14] Yoav Freund, Robert 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.
[15] Andreas Geiger, Philip Lenz, Raquel 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.
[16] David Geronimo, Antonio M. Lopez, Angel D. Sappa, Thorsten Graf, Survey of pedestrian detection for advanced driver assistance systems, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (7) (2010) 1239–1258.
[17] Samuel Gidel, Paul Checchin, Christophe Blanc, Thierry Chateau, Laurent Trassoudaine, Pedestrian detection and tracking in an urban environment using a multilayer laser scanner, IEEE Transactions on Intelligent Transportation Systems 11 (3) (2010) 579–588.
[18] Ross Girshick, Fast r-cnn, in: IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[19] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[20] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, Antonio M. López, Pedestrian detection at day/night time with visible and fir cameras: a comparison, Sensors 16 (6) (2016) 820.
[21] F.P. León, A.P. Grassi, V. Frolov, Information fusion to detect and classify pedestrians using invariant features, Information Fusion 12 (2011) 284–292.
[22] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, Christel-Loic Tisse, Exploiting fusion architectures for multispectral pedestrian detection and segmentation, Applied Optics 57 (2018) 108–116.

[23] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, Michael Ying Yang, Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection, Information Fusion 50 (2019) 148–157.
[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, in: European Conference on Computer Vision, Springer, 2014, pp. 346–361.
[25] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick, Mask r-cnn, in: IEEE International Conference on Computer Vision, 2017.
[26] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, In So Kweon, Multispectral pedestrian detection: benchmark dataset and baseline, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1037–1045.
[27] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, Trevor Darrell, Caffe: convolutional architecture for fast feature embedding, in: International Conference on Multimedia, 2014, pp. 675–678.
[28] Liu Jingjing, Zhang Shaoting, Wang Shu, Metaxas Dimitris, Multispectral deep neural networks for pedestrian detection, in: British Machine Vision Conference, 2016, pp. 73.1–73.13.
[29] Taehwan Kim, Sungho Kim, Pedestrian detection at night time in fir domain: comprehensive study about temperature and brightness and new benchmark, Pattern Recognition 79 (2018) 44–54.
[30] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, Michael Teutsch, Fully-convolutional region proposal networks for multispectral person detection, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 243–250.
[31] Stephen J. Krotosky, Mohan Manubhai Trivedi, Person surveillance using visual and infrared imagery, IEEE Transactions on Circuits and Systems for Video Technology 18 (8) (2008) 1096–1105.
[32] Yi-Shu Lee, Yi-Ming Chan, Li-Chen Fu, Pei-Yung Hsiao, Near-infrared-based nighttime pedestrian detection using grouped part models, IEEE Transactions on Intelligent Transportation Systems 16 (4) (2015) 1929–1940.
[33] Alex Leykin, Yang Ran, Riad 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.
[34] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast r-cnn for pedestrian detection, IEEE Transactions on Multimedia 20 (4) (2018) 985–996.
[35] Xiaofei Li, Lingxi Li, Fabian Flohr, Jianqiang Wang, Hui Xiong, Morys Bernhard, Shuyue Pan, Dariu M. Gavrila, Keqiang Li, A unified framework for concurrent pedestrian and cyclist detection, IEEE Transactions on Intelligent Transportation Systems 18 (2) (2017) 269–281.
[36] Xudong Li, Mao Ye, Yiguang Liu, Feng Zhang, Dan Liu, Song Tang, Accurate object detection using memory-based models in surveillance scenes, Pattern Recognition 67 (2017) 73–84.
[37] Jingjing Liu, Shaoting Zhang, Shu Wang, Dimitris Metaxas, Improved annotations of test set of kaist, http://paul.rutgers.edu/~jl1322/multispectral.htm/, 2018.
[38] Jiayuan Mao, Tete Xiao, Yuning Jiang, Zhimin Cao, What can help pedestrian detection?, in: IEEE Conference on Computer Vision and Pattern Recognition, July 2017.
[39] Woonhyun Nam, Piotr Dollár, Joon Hee Han, Local decorrelation for improved pedestrian detection, in: Advances in Neural Information Processing Systems, 2014, pp. 424–432.
[40] H. Nanda, L. Davis, Probabilistic template based pedestrian detection in infrared videos, in: Intelligent Vehicle Symposium, vol. 1, 2003, pp. 15–20.
[41] Miguel Oliveira, Vitor Santos, Angel D. Sappa, Multimodal inverse perspective mapping, Information Fusion 24 (2015) 108–121.
[42] Michael Oren, Constantine Papageorgiou, Pawan Sinha, Edgar Osuna, Tomaso Poggio, Pedestrian detection using wavelet templates, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 1997, pp. 193–199.
[43] Constantine Papageorgiou, Tomaso Poggio, Trainable pedestrian detection, in: Image Processing, 1999, Proceedings. 1999 International Conference on, vol. 4, ICIP 99, IEEE, 1999, pp. 35–39.

[44] Razvan Pascanu, Tomas Mikolov, Yoshua Bengio, On the difficulty of training recurrent neural networks, in: International Conference on Machine Learning, 2013, pp. 1310–1318.
[45] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster r-cnn: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.
[47] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, Yann Lecun, Pedestrian detection with unsupervised multi-stage feature learning, in: Computer Vision and Pattern Recognition, 2013, pp. 3626–3633.
[48] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.
[49] Frédéric Suard, Alain Rakotomamonjy, Abdelaziz Bensrhair, Alberto Broggi, Pedestrian detection using infrared images and histograms of oriented gradients, in: Intelligent Vehicles Symposium, 2006 IEEE, IEEE, 2006, pp. 206–212.
[50] Yonglong Tian, Ping Luo, Xiaogang Wang, Xiaoou Tang, Deep learning strong parts for pedestrian detection, in: IEEE International Conference on Computer Vision, 2015, pp. 1904–1912.
[51] Atousa Torabi, Guillaume Massé, Guillaume-Alexandre Bilodeau, An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications, Computer Vision and Image Understanding 116 (2) (2012) 210–221.
[52] P. Viola, M. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.
[53] Jörg Wagner, Volker Fischer, Michael Herman, Sven Behnke, Multispectral pedestrian detection using deep fusion convolutional neural networks, in: European Symposium on Artificial Neural Networks, 2016, pp. 509–514.
[54] Xiaogang Wang, Meng Wang, Wei Li, Scene-specific pedestrian detection for static video surveillance, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2) (2014) 361–374.
[55] B. Wu, R. Nevatia, Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part det, in: IEEE Intl. Conf. Computer Vision, 2005.
[56] Bichen Wu, Forrest Iandola, Peter H. Jin, Kurt Keutzer, Squeezedet: unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving, arXiv preprint, arXiv:1612.01051, 2016.
[57] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, Nicu Sebe, Learning cross-modal deep representations for robust pedestrian detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5363–5371.
[58] Yuan Yuan, Xiaoqiang Lu, Xiao Chen, Multi-spectral pedestrian detection, Signal Processing 110 (2015) 94–100.
[59] Li Zhang, Bo Wu, R. Nevatia, Pedestrian detection in infrared images based on local shape features, in: Computer Vision and Pattern Recognition, 2007, IEEE Conference on, CVPR '07, 2007, pp. 1–8.
[60] Liliang Zhang, Liang Lin, Xiaodan Liang, Kaiming He, Is faster r-cnn doing well for pedestrian detection?, in: European Conference on Computer Vision, 2016, pp. 443–457.
[61] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, Towards reaching human performance in pedestrian detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (4) (2018) 973–986.
[62] Shanshan Zhang, Rodrigo Benenson, Bernt Schiele, Filtered channel features for pedestrian detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[63] Shanshan Zhang, Rodrigo Benenson, Bernt Schiele, Citypersons: a diverse dataset for pedestrian detection, in: The IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, CVPR, 2017, p. 3.
[64] Xinyue Zhao, Zaixing He, Shuyou Zhang, Dong Liang, Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification, Pattern Recognition 48 (6) (2015) 1947–1960.
[65] Martin Zinkevich, Markus Weimer, Lihong Li, Alex J. Smola, Parallelized stochastic gradient descent, in: Advances in Neural Information Processing Systems, 2010, pp. 2595–2603.