Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection

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

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

ARTICLE INFO

ABSTRACT

Effective 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 present a novel box-level segmentation supervised learning framework for accurate and real-time multispectral pedestrian detection by incorporating features extracted in visible and infrared channels. Specifically, our method takes pairs of aligned visible and infrared images with easily obtained bounding box annotations as input and estimates accurate prediction maps to highlight the existence of pedestrians. It offers two major advantages over the existing anchor box based multispectral detection methods. Firstly, it overcomes the hyperparameter setting problem occurred during the training phase of anchor box based detectors and can obtain more accurate detection results, especially for small and occluded pedestrian instances. Secondly, it is capable of generating accurate detection results using small-size input images, leading to improvement of computational efficiency for real-time autonomous driving applications. Experimental results on KAIST multispectral dataset show that our proposed method outperforms state-of-the-art approaches in terms of both accuracy and speed.

1. Introduction

Pedestrian detection has received much attention within the field of computer vision and robotics in recent years (Oren et al., 1997; Dalal and Triggs, 2005; Dollár et al., 2012; Angelova et al., 2015; Geiger et al., 2012; Jafari and Yang, 2016; Cordts et al., 2016; Zhang et al., 2017b). Given images captured in various real-world surveillance situations, pedestrian detectors are required to accurately locate human regions. It provides an important functionality to facilitate human-centric applications such as autonomous driving, video surveillance, and urban monitoring (Wu et al., 2016; Li et al., 2017a,b; Zhang et al., 2017a; Wang et al., 2014; Bu and Chan, 2005; Shirazi and Morris, 2017).

Although significant improvements have been accomplished during recent years, it still remains a challenging task to develop a robust pedestrian detector ready for practical applications. It can be noticed that most existing pedestrian detection methods are based on visible

information alone. Their performances are sensitive to changes of the environmental brightness (daytime or nighttime). To overcome the aforementioned limitations, multispectral information (*e.g.,* visible and infrared), which can supply complementary information about the targets of interest, are considering to build more robust pedestrian detectors under various illumination conditions. In the past few years, multispectral pedestrian detection solutions are developed by many research works to achieve more accurate and stable pedestrian detection results for around-the-clock application (Leykin et al., 2007; Krotosky and Trivedi, 2008; Torabi et al., 2012; Oliveira et al., 2015; Hwang et al., 2015; González et al., 2016).

It is noted that most existing multispectral pedestrian detection approaches are built upon anchor box based detectors such as region proposal networks (RPN) (Zhang et al., 2016) or Faster R-CNN (Ren et al., 2017), localizing each human target using a bounding box. During the training phase, a large number of anchor boxes are needed to ensure sufficient overlap with most ground truth boxes and will
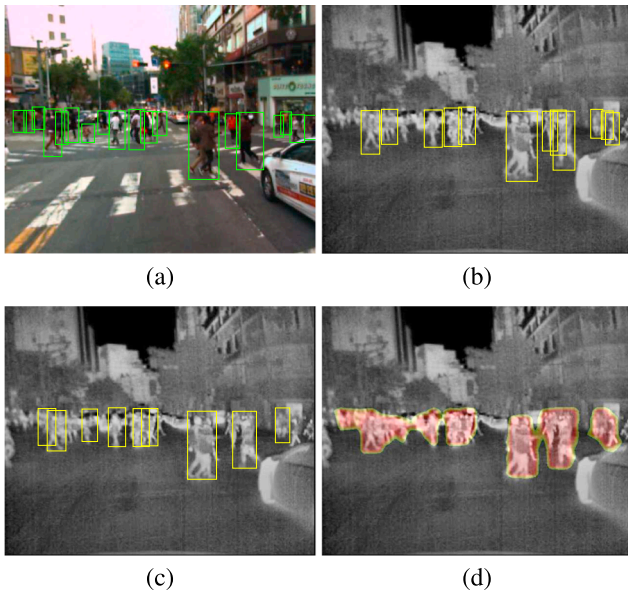
Fig. 1. (a) Ground truth detection results (displayed using the visible channel); (b) Bounding box detection results using 640 × 512 images (displayed using the thermal channel); (c) Bounding box detection results using 320 × 256 images; (d) Detection results of our proposed method using 320 × 256 images. Note that green bounding boxes show ground truth boxes, yellow bounding boxes show bounding box detections. A score threshold of 0.5 is used to display the detections. It is observed that the proposed box-level segmentation supervised learning framework produces more accurate detection results and successfully localizes far-scale human targets even when the input is small-size images. All images are resized to the same resolution for visualization.

cause severe imbalance between positive and negative anchor boxes and slow down the training process (Lin et al., 2018). Moreover, the state-of-the-art pedestrian detection techniques only perform well when the input is large-size images. Their performances will drop significantly when applied to small-size images since it is difficult to make use of anchor boxes to generate positive samples for small-size targets. A simple solution is to increase the size of input images and human targets through image up-scaling, however such practice will adversely decrease the computational efficiency which is critical for real-time autonomous driving applications.

To overcome the problems mentioned above, we present a novel box-level segmentation supervised learning framework for accurate and real-time multispectral pedestrian detection. Our approach takes pairs of aligned visible and infrared images with easily obtained bounding box annotations as input and computes heat maps to predict the existence of human targets. In Fig. 1, we show some comparative detection results of our method with the state-of-the-art anchor box based detector. It is noticed that the proposed box-level segmentation supervised learning framework produces more accurate detection results, successfully locating far-scale human targets even when the input is small-size images. It is also worth mentioning that our proposed method can process more than 30 images per second on a single NVIDIA Geforce Titan X GPU which is sufficient for real-time applications in autonomous vehicles. The contributions of this work are as follows.

Overall, the **contributions** of this paper are summarized as follows:

1 Our box-level segmentation supervised framework completely eliminates the complex hyperparameter settings of anchor boxes (e.g., box size, aspect ratio, stride, and intersection-over-union threshold) required in existing anchor box based detectors. To the best of our knowledge, this is the first attempt to train deep learning based multispectral pedestrian detectors without using anchor boxes.

2 We demonstrate that box-level approximate segmentation masks provide better supervision information than anchored boxes to train two-stream deep neural networks for distinguishing pedestrians from the background, particularly for small human targets. As the result, our method is capable of generating accurate detection results even using small-size input images.

3 Our method achieves significantly higher detection accuracy compared with the state-of-the-art multispectral pedestrian detectors (König et al., 2017; Jingjing et al., 2016a; Guan et al., 2018b,a; Li et al., 2018a). Moreover, this efficient framework can process more than 30 images per second on a single NVIDIA Geforce Titan X GPU to facilitate real-time applications in autonomous vehicles.

The remainder of our paper is structured as follows. Section 2 reviews the existing research work on multispectral pedestrian detection. The details of our proposed box-level segmentation supervised deep neural networks are presented in Section 3. An extensive evaluation of our method and experimental comparison of methods for multispectral pedestrian detection are provided in Section 4. We conclude our paper in Section 5.

2. Related works

Pedestrian detection facilitates various applications in robotics, automotive safety, surveillance, and autonomous vehicles. A large variety of visible-channel pedestrian detectors have been proposed. Schindler et al. (2010) developed a visual stereo system, which consists of various probabilistic models to fuse evidence from 3D points and 2D images, for accurate detection and tracking of pedestrians in urban traffic scenes. Dollár et al. (2009) developed the Integrate Channel Features (ICF) detector using feature pyramids and boosted classifiers for visible images. The feature representations of ICF have been further improved through various techniques, including aggregated channel features (ACF) (Dollár et al., 2014), locally decorrelated channel features (LDCF) (Nam et al., 2014), Checkerboards (Zhang et al., 2015), etc. Klinger et al. (2017) addressed the problems of target occlusion and imprecise visual observation by building up a new predictive model on the basis of Gaussian process regression, and by combining generic object detection with instance-specific classification for refined localization. Object detection based on deep neural networks (Girshick, 2015; Ren et al., 2017; He et al., 2017) have achieved state-of-the-art results on various challenging benchmarks, thus they have been adopted for the task of human-target detection. Li et al. (2018b) developed a scale-aware fast region-based convolutional neural networks (SAF R-CNN) method which combines a large-size sub-network and a small-size one into a unified architecture using a scale-aware weighting mechanism to capture unique pedestrian features at different scales. Zhang et al. (2016) proposed an effective baseline for pedestrian detection using region proposal networks (RPN) followed by boosted classifiers, which utilizes high-resolution convolutional feature maps generated by RPN for classification. Mao et al. (2017) proposed a powerful deep neural networks framework by implementing representations of channel features to boost pedestrian detection accuracy without extra inputs in inference. Brazil et al. (2017) developed an effective segmentation infusion network to improve pedestrian detection performance through the joint training of target detection and semantic segmentation.

Recently, multispectral pedestrian detection becomes a promising solution to narrow the gap between automatic pedestrian detectors and human observers. Multi-modal sensors (visible and infrared) supply complementary information about the targets of interest thus lead to more robust and accurate detection results. Hwang et al. (2015) published the first large-scale multispectral pedestrian dataset (KAIST) which contains well-aligned visible and infrared image pairs with dense pedestrian annotations. Wagner et al. (2016) presented the first application of deep neural networks for multispectral pedestrian detection.

Two decision networks, one for early-fusion and the other for late-fusion, were proposed to classify the proposals generated by ACF + T + THOG (Hwang et al., 2015) and achieved more accurate detections. Jingjing et al. (2016a) systematically evaluated the performance of four ConvNet fusion architectures which integrate two-branch ConvNets on different DNNs stages and found the optimal architecture is the Halfway Fusion model that merges two-branch ConvNets on the middle-level convolutional features. König et al. (2017) adopt the architecture of RPN + BDT (Zhang et al., 2016) to build Fusion RPN + BDT, which merges the two-branch RPN on the middle-level convolutional features, for multispectral pedestrian detection. Recently, researchers explore illumination information of a scene and proposed illumination-aware weighting mechanism to boost multispectral pedestrian detection performances (Guan et al., 2018b; Li et al., 2019). Guan et al. (2018a) presented a unified multispectral fusion framework for joint training of semantic segmentation and target detection. More accurate detection results were obtained by infusing the multispectral semantic segmentation masks as supervision for learning human-related features. Li et al. (2018a) further deployed subsequent multispectral classification network to distinguish pedestrian instances from hard negatives.

It is noted that most existing multispectral pedestrian detection approaches are built upon anchor box based detectors such as region proposal networks (RPN) (Zhang et al., 2016) or Faster R-CNN (Ren et al., 2017), using a number of bounding boxes to localize human pedestrians. However, the use of anchor boxes will cause severe imbalance between positive and negative training samples (Lin et al., 2018) and involve complex hyperparameter settings (e.g., box size, aspect ratio, stride, and intersection-over-union threshold) (Law and Deng, 2018). Our method is very different from the existing anchor box based multispectral pedestrian detectors (König et al., 2017; Jingjing et al., 2016a; Li et al., 2019; Guan et al., 2018b,a; Li et al., 2018a) in two major aspects. Firstly, we make use of the ground truth bounding boxes (manually annotated) to generate coarse box-level segmentation masks, which are utilized to replace the anchor bounding boxes for the training of two-stream deep neural networks to learn human-relative characteristic features. Secondly, our method estimates a prediction heat map instead of a number of bounding boxes to localize pedestrians in the surrounding space, which can be easily used to support perceptive autonomous driving applications such as path planning or collision avoidance. It is worth mentioning that a large number of semantic segmentation techniques have been proposed to generate accurate boundary between foreground objects and background regions without using anchor boxes (Ha et al., 2017; Balloch et al., 2018; Jégou et al., 2017). However, these methods typically require the supervision of pixel-level accuracy mask annotations which are very time-consuming to obtain. Many researchers attempted to achieve competitive semantic segmentation accuracy by only using the easily obtained bounding box annotations (Dai et al., 2015; Rajchl et al., 2017). These methods involve iterative updates to gradually improve the accuracy of segmentation masks, which are slow and not suitable for real-time autonomous driving applications.

3. Our approach

We propose a novel box-level segmentation supervised framework for multispectral pedestrian detection. Given pairs of well-aligned visible and infrared images, we make use of two-stream deep neural networks to extract semantic features in individual channels. Visible and infrared feature maps are combined through the concatenation operation and then utilized to estimate heat maps to predict the existence of pedestrians as illustrated in Fig. 2. Note that image regions corresponding to human targets produce high confident scores (larger than 0.5).
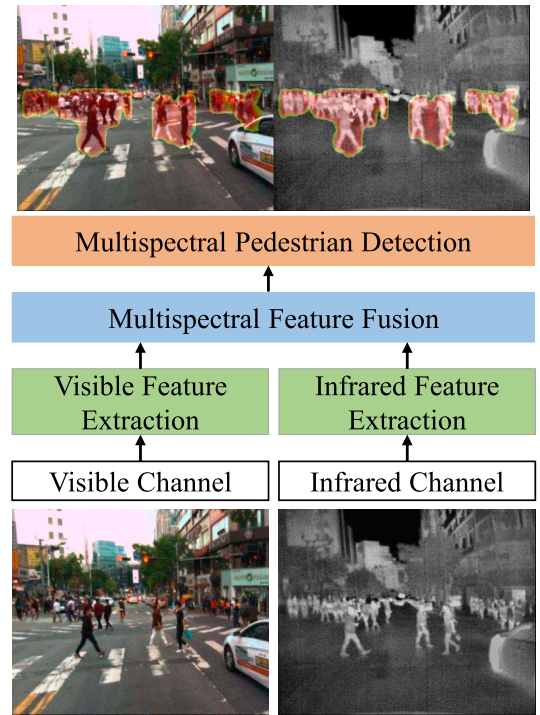

Fig. 2. The workflow of our proposed box-level segmentation supervised deep neural networks for multispectral pedestrian detection. Please note our method generates a prediction heat map (a score threshold of 0.5 is used to display the detected pedestrian regions) instead of a number of bounding boxes to localize pedestrians in the scene. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
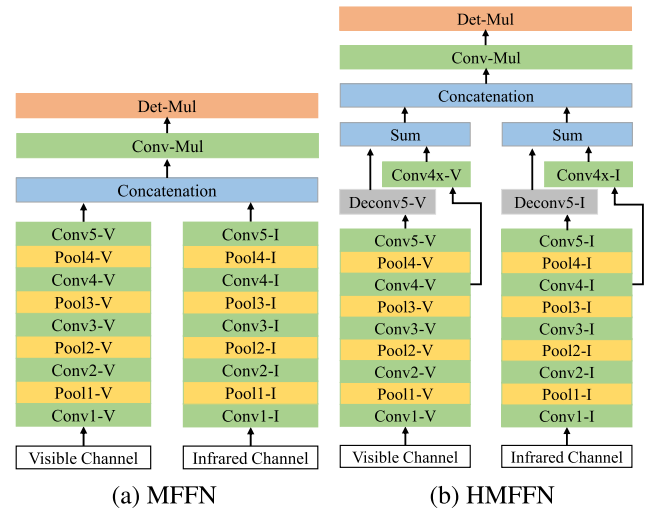
Fig. 3. Illustration of (a) MFFN and (b) HMFFN architectures. Note that green boxes represent convolutional layers, yellow boxes represent pooling layers, blue boxes represent fusion layers, gray boxes represent deconvolutional layers, and orange boxes represent soft-max layers. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Network architecture

Fig. 3(a) shows a baseline architecture of our proposed multispectral feature fusion network (MFFN) for pedestrian detection. Given a pair of well-aligned visible and infrared images, we make use of the two-stream deep convolutional neural networks presented by Jingjing

et al. (2016b) 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-I to Conv5-I in the infrared stream) which adopts the architectures of Conv1-5 from VGG-16 (Simonyan and Zisserman, 2014). The two single-channel feature maps are then fused using the concatenation layer followed by a 1×1 convolutional layer (Conv-Mul) to learn two-channel multispectral semantic features. We use a softmax layer (Det-Mul) to estimate the heat map to predict the location of pedestrians.

Inspired by the recent success of top-down architecture with lateral connections for object detection and segmentation (Pinheiro et al., 2016; Lin et al., 2017), we design another hierarchical multispectral feature fusion network (HMFFN) and its architecture is shown in Fig. 3(b). It is noted that the HMFFN architecture makes use of skip connections to associate the middle-level feature maps (output of Conv4-V/I layers) with the high-level ones (output of Conv5-V/I layers). The deconvolutional layers (Deconv5-V/I) are deployed to increase the spatial resolution of high-level feature maps by a factor of 2. Then, the upsampled high-level feature maps are merged with the corresponding middle-level ones (which undergoes 1×1 convolutional layers Conv4x-V/I to reduce channel dimensions) by element-wise addition. In deep convolutional neural networks, outputs of deeper layers encode high-level semantic information while shallower layers outputs capture rich low-level spatial patterns (Lin et al., 2017; Hou et al., 2017). Therefore, the proposed HMFFN architecture, combining feature maps from different levels, is capable of extracting informative multi-scale feature maps to achieve more accurate detection results. The comparative evaluation of MFFN and HMFFN architectures are provided in Section 4.3.

3.2. Box-level segmentation for supervised training

A common step of state-of-the-art anchor box based detectors is generating a large number of anchor boxes with various sizes and aspect ratios as potential detection candidates as illustrated in Fig. 4(a). However, the use of anchor boxes involves complex hyperparameter settings (e.g., box size, aspect ratio, stride, and intersection-over-union threshold) (Law and Deng, 2018) and causes severe imbalance between positive and negative training samples (Lin et al., 2018). Moreover, it is difficult to make use of discretely distributed anchor boxes (using a large stride) to generate positive samples for small-size targets. In comparison, our proposed method takes the easily obtained bounding box annotation as input and generates an unambiguous box-level segmentation mask for the training of two-stream deep neural networks to

(a)   (b)
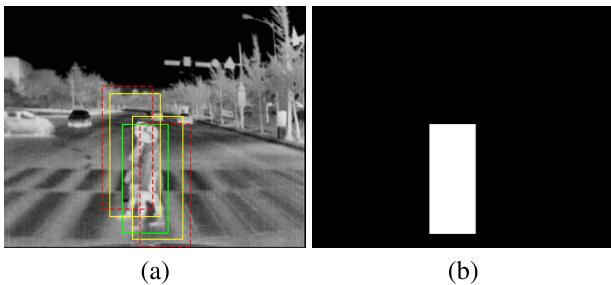
Fig. 4. Illustration of generating training labels using (a) Anchor boxes; (b) Box-level segmentation masks. The use of anchor boxes involves complex hyperparameter settings (e.g., box size, aspect ratio, stride, and intersection-over-union threshold). In comparison, our proposed method generates an unambiguous box-level segmentation mask for learning human-relative features. Note that green bounding boxes (BBs) represent BB ground truth, yellow BBs represent positive training samples, and red BBs in dashed line represent negative training samples. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

learn human-relative characteristic features as illustrated in Fig. 4(b). In our implementation, the obtained box-level segmentation masks are down-scaled to match with the size of final multispectral feature maps (outputs of the concatenation layer) through bilinear interpolation. It is worth mentioning that it is a challenging task to obtain pixel-level accurate annotations for visible and infrared image pairs since it is difficult to obtain perfectly aligned and synchronized multispectral data (Hwang et al., 2015). Therefore, we attempt to explore the easily obtained bounding box annotations as an alternative of supervision to train deep convolutional neural networks for multispectral target detection.

Let $\{(X, Y)\}$ denote the training images $X = \{x_i, i = 1, ..., M\}$ (M pixels) with box-level approximate segmentation masks $Y = \{y_i, i = 1, ..., M\}$, where $y_i = 1$ denotes the foreground pixel and $y_i = 0$ is the background pixel. The parameters θ of multispectral pedestrian detector are updated by minimizing the cross-entropy loss which is defined as

$$\mathcal{L}(\theta) = - \sum_{i \in Y_+} \log \Pr(y_i = 1 | X; \theta)$$
$$- \sum_{i \in Y_-} \log \Pr(y_i = 0 | X; \theta), \tag{1}$$

where Y_+ and Y_- represent the foreground and background pixels respectively, and $\Pr(y_i | X; \theta) \in [0, 1]$ is the confidence score of the prediction that measures probability of the pixel belong to pedestrian regions. The confidence score is calculated utilizing the softmax function as

$$\Pr(y_i = 1 | X; \theta) = \frac{e^{s_1}}{e^{s_0} + e^{s_1}}, \tag{2}$$

$$\Pr(y_i = 0 | X; \theta) = \frac{e^{s_0}}{e^{s_0} + e^{s_1}}, \tag{3}$$

where s_0 and s_1 are the computed values in our two-channel feature maps. The optimal parameters θ^* are obtained by minimizing the loss function $\mathcal{L}(\theta)$ through the gradient descent optimization algorithm as

$$\theta^* = \arg \min_\theta \mathcal{L}(\theta). \tag{4}$$

The output of our method is a full-size prediction heat map in which human target regions yields high confident scores (larger than 0.5) while the background regions produce low ones. Such perceptive information is useful for many autonomous driving applications such as path planning or collision avoidance. In comparison, it is difficult/impractical to use a number of bounding boxes to identify individual pedestrians in crowded urban scenes. Visual comparisons are provided in Fig. 1.

4. Experiments

4.1. Dataset and evaluation metric

All the detectors are evaluated using the public KAIST multispectral pedestrian benchmark (Hwang et al., 2015). We notice that CVC-14 (González et al., 2016) is another newly published multispectral pedestrian benchmark consisting of infrared and visible gray image pairs. However, the multispectral image pairs were not properly aligned thus the pedestrian annotations are individually labeled in infrared and visible images. It should be noted that some annotations are only generated in the infrared/visible image on the CVC-14 dataset. To the best of our knowledge, KAIST multispectral pedestrian benchmark is the only available pedestrian dataset which contains large-scale and well-aligned visible-infrared image pairs with accurate manual annotations.

Totally, KAIST training dataset consists of 50172 well aligned visible-infrared image pairs (640×512 resolution) captured in all-day traffic scenes with 13853 pedestrian annotations. The training images are sampled in every 2 frames following the other multispectral

pedestrian detection methods (Jingjing et al., 2016b; König et al., 2017; Guan et al., 2018b,a; Li et al., 2018a). The KAIST testing dataset contains 2252 image pairs with 1356 pedestrian annotations. Since the original KAIST testing dataset contains many problematic annotations (e.g., inaccurate bounding boxes and missed human targets), we make use of the improved annotations provided by Liu et al. (2018) for quantitative and qualitative evaluation. Specifically, we consider all reasonable, scale, and occlusion subset of the KAIST testing dataset (Hwang et al., 2015).

The output of our method is a full-size prediction heat map in which human target regions yields high confident scores while the background regions produce low ones. For a fair comparison, we transform the bounding box detection results with different prediction scores to the heat map representation, and the pixel-level average precision (AP) (Salton and McGill, 1986; Cordts et al., 2016) is utilized to evaluate the quantitative performance of multispectral pedestrian detectors in the pixel-level. The computed detection results are compared with the ground-truth annotation masks which are generated based on manually labeled bounding boxes. Pixels located in the ground-truth bounding boxes are defined as foreground ones, while other pixels are defined as background ones. Given the heat map predictions, true positive (TP) is the number of correctly predicted foreground pixels, false positive (FP) is the number of incorrectly predicted background pixels, and false negative (FN) is the number of incorrectly foreground background pixels. Precision is calculated as $TP/(TP + FP)$ and recall is computed as $TP/(TP + FN)$. The AP depicts the shape of the precision/recall curve, and is defined as the mean precision at a number of equally spaced recall levels by varying the threshold on detection scores. In our implementation, we average the precision values at 100 recall levels equally spaced between 0 and 1.

4.2. Implementation details

The image-centric training and testing strategy are applied to generate mini-batches without using image pyramids. The batch size is set to 1 according to the method presented by Guan et al. (2018a). Each stream of the feature extraction layers in MFFN and HMFFN are initialized using the weights and bias of VGG-16 net (Simonyan and Zisserman, 2014) pre-trained on the ImageNet dataset (Russakovsky et al., 2015). All the other convolutional layers use normalized initialization following the method presented by Glorot and Bengio (2010). We utilize the Jia et al. (2014) deep learning framework to train and test our proposed multispectral pedestrian detectors. All the models are fine-tuned using stochastic gradient descent (SGD) (Zinkevich et al., 2010) for the first two epochs with the learning rate of 0.001 and one more epoch with the learning rate of 0.0001. Adjustable gradient clipping technique is used in training to suppress exploding gradients (Pascanu et al., 2013).

4.3. Evaluation of multispectral feature fusion schemes

In this paper, we design two multispectral feature fusion schemes (MFFN and HMFFN). The HMFFN model makes use of skip connections to associate the middle-level feature maps (output of Conv4-V/I layers) with the high-level ones (output of Conv5-V/I layers). We experimentally evaluate the performance gain by incorporating middle-level feature maps into the baseline MFFN model. The quantitative performance (pixel-level AP (Salton and McGill, 1986)) of MFFN and HMFFN for different sizes of input images (640×512, 480×384, and 320×256) are compared in Table 1.

We observe that better detection performance is achieved through the hierarchical multispectral feature fusion. Moreover, the performance gain is more obvious when handling small-size input images. By incorporating the middle-level feature maps, AP index significantly increases from 0.748 (MFFN-320) to 0.817 (HMFFN-320) for 320×256 resolution input images in the Reasonable-all subset, while the

improvement is not obvious for 640×512 resolution input images (increasing from 0.844 to 0.854). The underlying reason is that the middle-level features from shallower layers (Conv4-V/I) encode rich small-scale image characteristics which are essential for accurate detection of small-size targets. Using a smaller size input image will significantly improve the computational efficiency for real-time autonomous driving applications.

Furthermore, we conduct the qualitative comparison of two multispectral feature fusion networks (MFFN-320 and HMFFN-320) by displaying detection results in various scenes in Fig. 5. It is observed that performance gains can generally be achieved (in both daytime and nighttime scenes and on different scale and occlusion subsets) by integrating middle-level feature maps with high-level ones. We evaluate MFFN-320 and HMFFN-320 models on testing subsets of different scales. Although both MFFN-320 and HMFFN-320 work well on the near scale subset, HMFFN-320 can better identify pedestrian targets in the medium and far scale subsets through incorporating image details extracted in middle-level layers (Conv4-V/I). Moreover, we test MFFN-320 and HMFFN-320 models on different occlusion subsets and observe that HMFFN-320 generates more accurate detection results when target objects are partially or heavily occluded. A reasonable explanation of this improvement is that low-level features extracted in shallower layers (Conv4-V/I) provide useful information of human parts and their relationships to handle the challenging target occlusion problem (Shu et al., 2012). The experimental results verify the effectiveness of the proposed HMFFN architecture, capable of extracting informative multiscale feature maps to achieve more precise object detection and remain more robust against scene variations.

4.4. Evaluation of box-level segmentation supervised framework

In this subsection, we evaluate the performance gain of using box-level segmentation masks instead of anchor boxes to train deep convolutional neural networks for multispectral target detection. For a fair comparison, we make use of the same architecture in HMFFN for multispectral feature extraction/fusion as shown in Fig. 3(b). Given the multispectral semantic features from Conv-Mul layer, the anchor box based detector RPN (Zhang et al., 2016) is utilized to generate confident scores and bounding boxes as detection results. In comparison, our proposed segmentation mask supervised method computes a prediction heat map to highlight the existence of human targets in a scene. The performances (pixel-level AP (Salton and McGill, 1986)) of our proposed box-level segmentation supervised method (HMFFN) and the one based on anchor boxes (RPN-HMFFN) on different sizes of input images (640×512, 480×384, and 320×256) are quantitatively compared in Table 2.

It is observed that HMFFN based on box-level segmentation masks performs better than RPN-HMFFN based on anchor boxes, achieving significantly higher AP indexes on various testing subsets and on images of different sizes (HMFFN-640 0.854 AP vs. RPN-HMFFN-640 0.756 AP on the reasonable all subset). Such improvements are particularly evident on some challenging detection tasks (HMFFN-640 0.166 AP vs. RPN-HMFFN-640 0.065 AP for far scale human target detection). Another advantage of our proposed HMFFN is that it directly computes a prediction heat map instead of confident scores and coordinates of bounding boxes, achieving faster inference speed (HMFFN-320 38.3 fps vs. RPN-HMFFN-320 32.0 fps).

Furthermore, we qualitatively show some sample detection results of HMFFN-640 and RPN-HMFFN-640 in Fig. 6. The output of our method is a full-size prediction heat map in which human target regions yields high confident scores. For a fair comparison, we also transform the bounding box detection results with different prediction scores to the heat map representation, utilizing different colors to show prediction scores of bounding boxes. Note we only show regions with confident scores larger than 0.5. It is noted that HMFFN-640 generate more precise detection results and fewer false positives compared with RPN-

Table 1
Quantitative performance (pixel-level AP (Salton and McGill, 1986)) of MFFN and HMFFN for different sizes of input images (640 × 512, 480 × 384, and 320 × 256).

| Model	Reasonable all	Reasonable day	Reasonable night	Near scale	Medium scale	Far scale	No occlusion	Partial occlusion	Heavy occlusion	Inference speed (fps)
MFFN-640	0.844	0.849	0.836	0.812	0.736	0.163	0.816	0.373	0.169	12.4
HMFFN-640	0.854	0.865	0.836	0.797	0.785	0.166	0.832	0.391	0.171	10.8
MFFN-480	0.825	0.837	0.812	0.799	0.705	0.100	0.790	0.328	0.152	20.3
HMFFN-480	0.843	0.866	0.805	0.796	0.764	0.148	0.818	0.373	0.152	18.5
MFFN-320	0.748	0.757	0.740	0.756	0.546	0.043	0.697	0.243	0.110	40.0
HMFFN-320	0.817	0.825	0.808	0.779	0.696	0.111	0.779	0.345	0.140	38.3

(a) Daytime
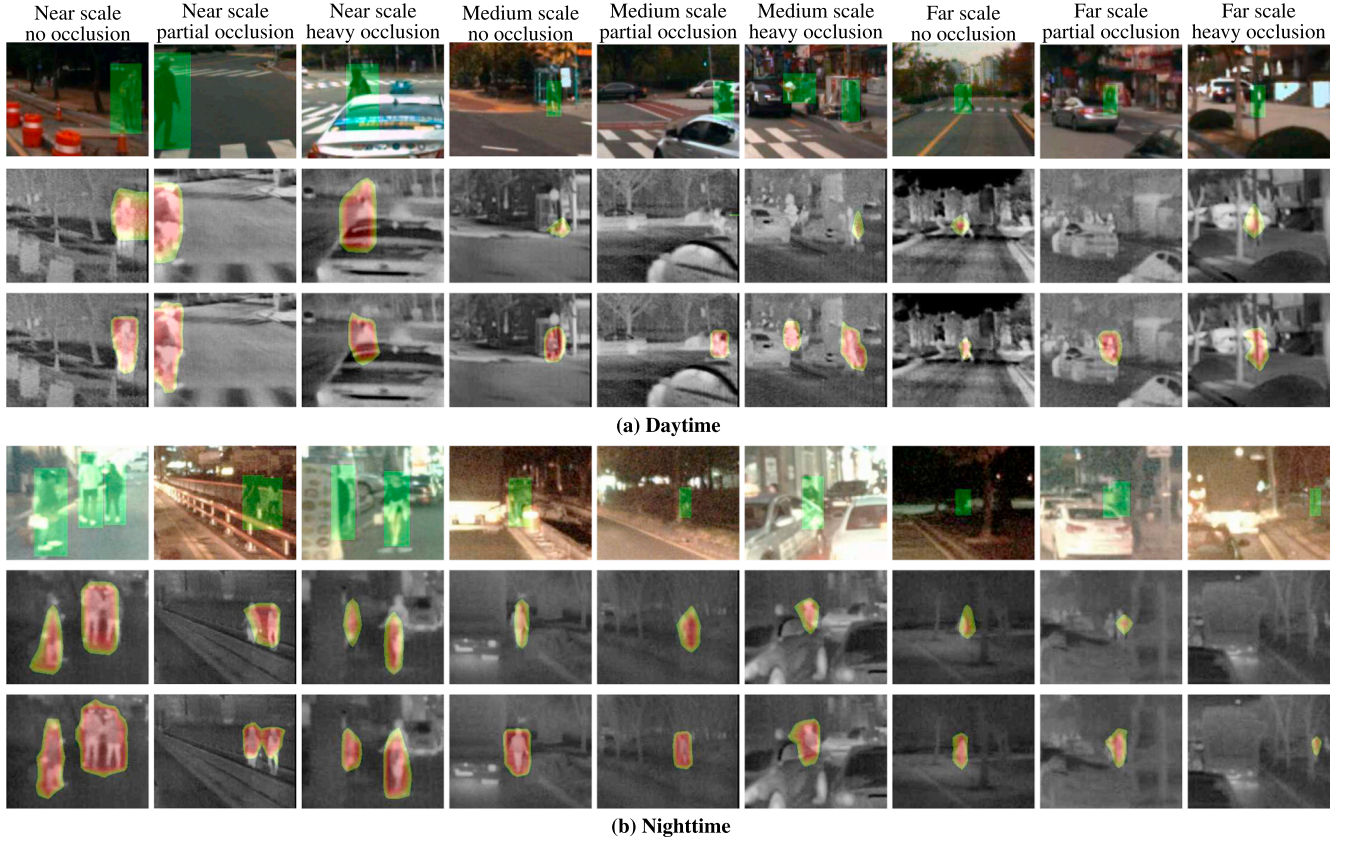
(b) Nighttime

Fig. 5. Qualitative comparison of multispectral pedestrian detection results of MFFN-320 and HMFFN-320 in the KAIST testing images captured in (a) daytime and (b) nighttime scenes. The first row shows the ground truth (displaying using the visible channel) and the others show detection results of MFFN-320 and HMFFN-320 respectively (displaying using the infrared channel). Note that the green regions represent ground-truth annotation masks which are generated based on manually labeled bounding boxes, and the detected pedestrian targets are visualized using the heat map representation with a 0.5 threshold. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

HMFFN-640. The use of anchor boxes involves complex hyperparameter settings (e.g., box size, aspect ratio, stride, and intersection-over-union threshold) will cause severe imbalance between positive and negative training samples and damage the learning of human-related features (Law and Deng, 2018). Moreover, we observe that HMFFN-640 can successfully identify some pedestrian instances on the far scale and heavy occlusion subsets, which are difficult to detect using the anchor box based RPN-HMFFN-640 or even based on visual observation. For small/occluded targets, it is difficult to generate enough positive samples using discretely distributed anchor boxes. In comparison, our

Table 2
Quantitative performance (pixel-level AP (Salton and McGill, 1986)) of our proposed box-based segmentation supervised detectors (HMFFN) with the anchor box based detectors (RPN-HMFFN) for different sizes of input images (640 × 512, 480 × 384, and 320 × 256).

| Model	Reasonable all	Reasonable day	Reasonable night	Near scale	Medium scale	Far scale	No occlusion	Partial occlusion	Heavy occlusion	Inference speed (fps)
RPN-HMFFN-640	0.756	0.761	0.741	0.607	0.662	0.065	0.705	0.263	0.149	9.4
HMFFN-640	0.854	0.865	0.836	0.797	0.785	0.166	0.832	0.391	0.171	10.8
RPN-HMFFN-480	0.75	0.755	0.743	0.591	0.64	0.046	0.7	0.282	0.142	16.5
HMFFN-480	0.843	0.866	0.805	0.796	0.764	0.148	0.818	0.373	0.152	18.5
RPN-HMFFN-320	0.718	0.717	0.713	0.638	0.571	0.057	0.672	0.225	0.124	32.0
HMFFN-320	0.817	0.825	0.808	0.779	0.696	0.111	0.779	0.345	0.140	38.3

(a) Daytime

(b) Nighttime
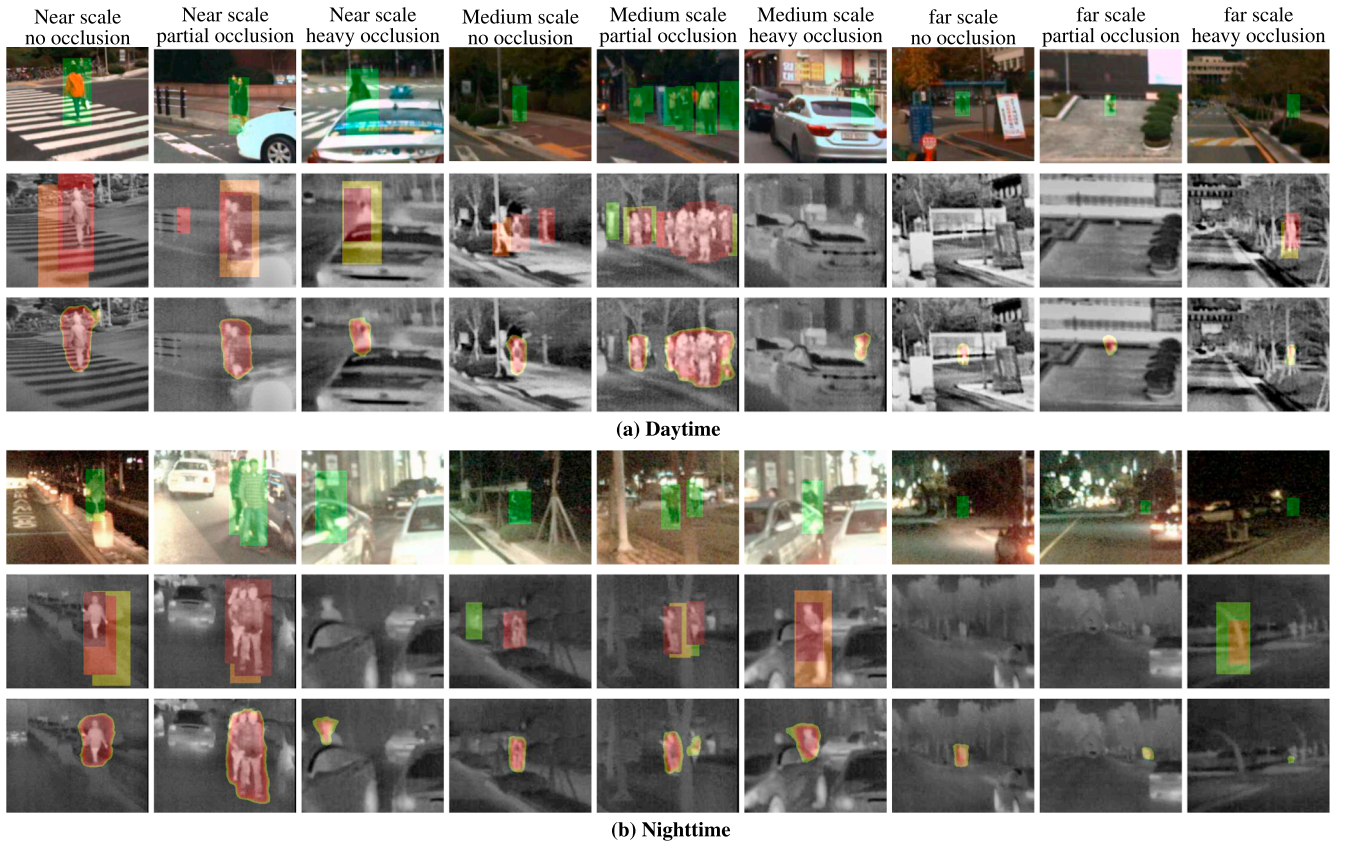
Fig. 6. Qualitative comparison of multispectral pedestrian detection results of RPN-HMFFN-640 and HMFFN-640 in the KAIST testing dataset. First row shows the ground truth (displaying using the visible channel) and the others show detection results of RPN-HMFFN-640 and HMFFN-640 respectively (displaying using the infrared channel). Note that the green regions represent ground-truth annotation masks which are generated based on manually labeled bounding boxes, and the detected pedestrian targets are visualized using the heat map representation with a 0.5 threshold. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

proposed HMFFN takes the easily obtained bounding box annotation as input and produces an unambiguous box-level segmentation mask for learning to distinguish target objects from the background. Overall, our experimental results demonstrate that box-level approximate segmentation masks provide better supervision information than anchored boxes for the training of two-stream deep neural networks to learn human-relative characteristic features.

4.5. Comparison with the state-of-the-art

We compare the proposed HMFFN-640 and HMFFN-320 models with a number of state-of-the-art multispectral pedestrian detectors including Halfway Fusion (Jingjing et al., 2016b), Fusion RPN + BDT (König et al., 2017), IATDNN + IAMSS (Guan et al., 2018b), FRPN-Sum + TSS (Guan et al., 2018a), and MSDS-RCNN (Li et al., 2018a). The Fusion RPN + BDT (König et al., 2017) model is re-implemented and trained according to the original papers, and the detection results of Halfway Fusion (Jingjing et al., 2016b), IATDNN + IAMSS (Guan et al., 2018b), FRPN-Sum + TSS (Guan et al., 2018a), and MSDS-RCNN (Li et al., 2018a) are kindly provided by the authors.

The quantitative evaluation results of different multispectral pedestrian detectors are shown in Table 3. Our proposed HMFFN-640 and HMFFN-320 models both achieve higher AP values in all reasonable, scale, and occlusion subset of the KAIST testing dataset. These comparative results indicate that our propose multispectral pedestrian detector achieves more robust performances under various surveillance situations. We qualitatively compare different multispectral pedestrian detectors by visualizing some sample detection results in Fig. 7. The

output of our method is a full-size prediction heat map in which human target regions yields high confident scores, while the bounding box detection results with different prediction scores are transformed to the heat map representation, utilizing different colors to show prediction scores of bounding boxes. Note we only show regions with confident scores larger than 0.5. Different from the existing multispectral pedestrian detection methods which generate a number of bounding boxes, our method estimates a full-size prediction heat map to highlight the existence of pedestrians in a scene. It is observed that our approach is capable of generating accurate detection results even for small human targets and using small-size input images.

We also compare the computational efficiency of HMFFN-640 and HMFFN-320 with state-of-the-art methods. A single Titan X GPU is utilized to evaluate the computation efficiency. Please note that the current state-of-the-art multispectral pedestrian detectors (König et al., 2017; Jingjing et al., 2016a; Guan et al., 2018b,a; Li et al., 2018a) typically perform image up-scaling to achieve their optimal detection performances. For instance, input sizes of (Jingjing et al., 2016b), Fusion RPN + BDT (König et al., 2017), IATDNN + IAMSS (Guan et al., 2018b), FRPN-Sum + TSS (Guan et al., 2018a), and MSDS-RCNN (Li et al., 2018a) models are 750×600, 960×768, 960×768, 960×768, and 750×600, respectively. In comparison, HMFFN-640 directly takes 640×512 multispectral data as input without image up-scaling thus run much faster (10.8 fps vs. 4.4 fps). Moreover, our HMFFN-320 model takes small-size 320×256 images as input and achieves 38.3 fps which is sufficient for real-time autonomous driving applications. Please note HMFFN-320 achieves more accurate detection results than the current state-of-the-art multispectral pedestrian detection methods.

Table 3

Quantitative comparison of HMFFN-640 and HMFFN-320 with the current state-of-the-art methods (König et al., 2017; Jingjing et al., 2016a; Guan et al., 2018b,a; Li et al., 2018a). Input sizes of different models are Halfway Fusion – 750 × 600, Fusion RPN + BDT – 960 × 768, IATDNN + IAMSS – 960 × 768, FRPN-Sum + TSS – 960 × 768, MSDS-RCNN – 750 × 600, HMFFN-640 – 640 × 512, and HMFFN-320 – 320 × 256. The top three results are highlighted in red, green, and blue, respectively.

Model	Reasonable all	Reasonable day	Reasonable night	Near scale	Medium scale	Far scale	No occlusion	Partial occlusion	Heavy occlusion	Inference speed (fps)
Halfway Fusion (Jingjing et al., 2016a)	0.702	0.708	0.691	0.623	0.583	0.062	0.695	0.128	0.037	2.5
Fusion FRPN+BDT (Konig et al., 2017)	0.755	0.767	0.731	0.663	0.681	0.027	0.700	0.165	0.030	1.3
IATDNN+IAMSS (Guan et al., 2018b)	0.766	0.772	0.756	0.614	0.643	0.043	0.715	0.263	0.106	4.0
FRPN-Sum+TSS (Guan et al., 2018a)	0.765	0.767	0.750	0.626	0.638	0.045	0.714	0.277	0.116	4.4
MSDS-RCNN (Li et al. 2018a)	0.744	0.750	0.721	0.670	0.673	0.068	0.712	0.206	0.070	4.4
HMFFN-640 (ours)	0.854	0.865	0.836	0.797	0.785	0.166	0.832	0.391	0.171	10.8
HMFFN-320 (ours)	0.817	0.825	0.808	0.779	0.696	0.111	0.779	0.345	0.140	38.3

(a) Daytime
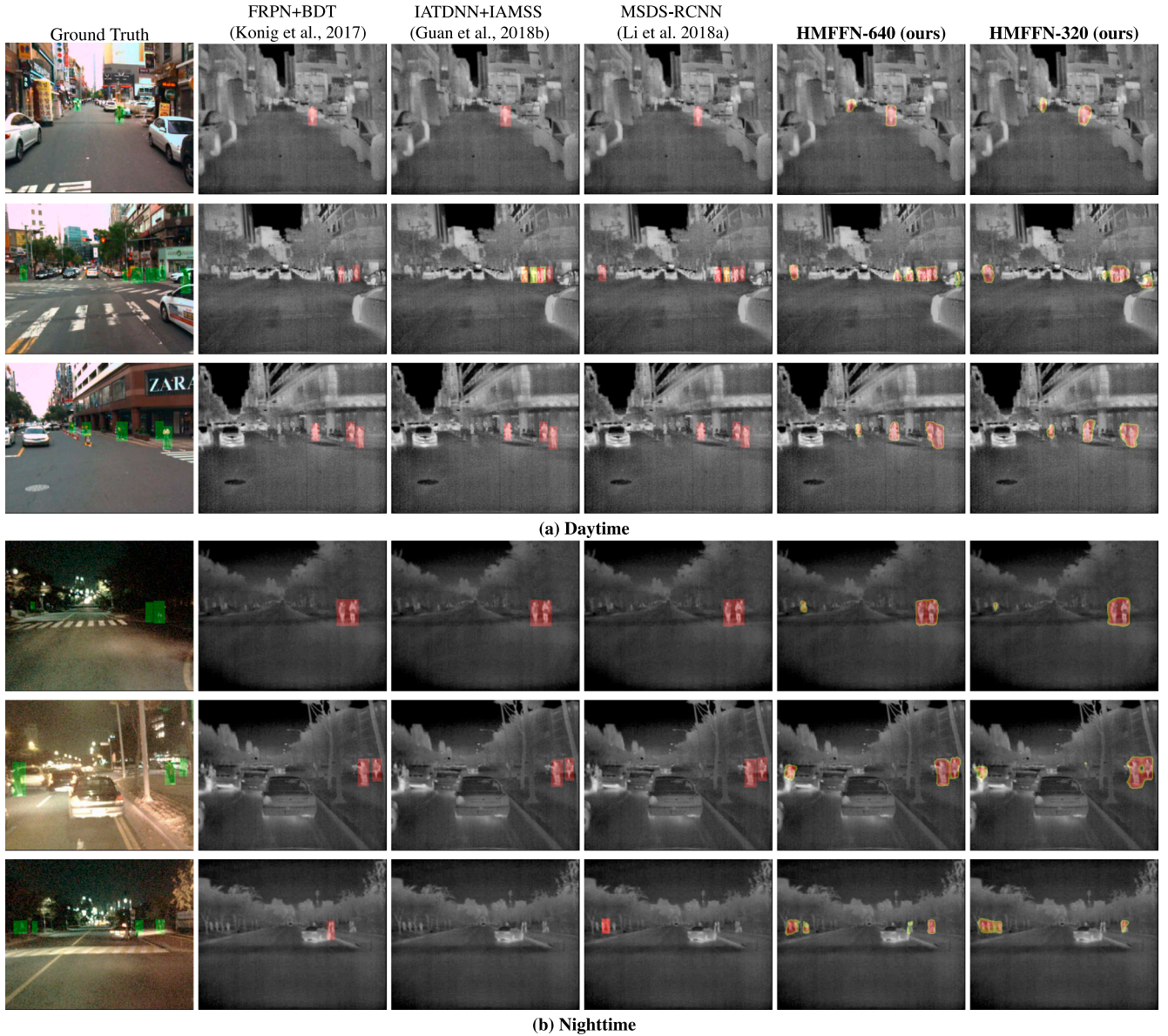
(b) Nighttime

Fig. 7. Qualitative comparison of multispectral pedestrian detection results in the KAIST testing dataset with other state-of-the-art approaches. First column shows the ground truth (displaying using the visible channel) and the others show detection results of Fusion RPN + BDT (König et al., 2017),IATDNN + IAMSS (Guan et al., 2018b), MSDS-RCNN (Li et al., 2018a) and our proposed HMFFN-640 and HMFFN-320 respectively (displaying using the infrared channel). Note that the green regions represent ground-truth annotation masks which are generated based on manually labeled bounding boxes, and the detected pedestrian targets are visualized using the heat map representation with a 0.5 threshold. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Conclusions

In this paper, we propose a powerful box-level segmentation supervised learning framework for accurate and real-time multispectral pedestrian detection. To the best of our knowledge, this represents the first attempt to train multispectral pedestrian detectors without using anchor boxes. Extensive experimental results verify that box-level approximate segmentation masks provide useful information for distinguishing human targets from the background. Also, we design a hierarchical multispectral feature fusion scheme in which the middle-level feature maps (small-scale image characteristics) and the high-level ones (semantic information) are incorporated to achieve more accurate detection results, particularly for far-scale human targets. Experimental results on KAIST benchmark show that our proposed method achieves higher detection accuracy compared with the state-of-the-art multispectral pedestrian detectors. Moreover, this efficient framework achieves real-time processing speed and processes more than 30 images per second on a single NVIDIA Geforce Titan X GPU. The proposed methods can be generalized to other object detection task with multispectral input and facilitate potential applications (e.g., path planning, collision avoidance, and target tracking) in autonomous vehicles.

Acknowledgment

References

Angelova, A., Krizhevsky, A., Vanhoucke, V., 2015. Pedestrian detection with a large-field-of-view deep network. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 704–711.

Balloch, J.C., Agrawal, V., Essa, I., Chernova, S., 2018. Unbiasing semantic segmentation for robot perception using synthetic data feature transfer. arXiv preprint arXiv: 1809.03676.

Brazil, G., Yin, X., Liu, X., 2017. Illuminating pedestrians via simultaneous detection and segmentation. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 4960–4969.

Bu, F., Chan, C.-Y., 2005. Pedestrian detection in transit bus application: sensing technologies and safety solutions. In: 2005. Proceedings. IEEE Intelligent Vehicles Symposium. IEEE, pp. 100–105.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3213–3223.

Dai, J., He, K., Sun, J., 2015. Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1635–1643.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1. IEEE, pp. 886–893.

Dollár, P., Tu, Z., Perona, P., Belongie, S., 2009. Integral channel features. In: British Machine Vision Conference, pp. 91.

Dollár, P., Wojek, C., Schiele, B., Perona, P., 2012. Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. 34 (4), 743–761.

Dollár, P., Appel, R., Belongie, S., Perona, P., 2014. Fast feature pyramids for object detection. IEEE Trans. Pattern Anal. Mach. Intell. 36 (8), 1532–1545.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3354–3361.

Girshick, R., 2015. Fast r-cnn. In: IEEE International Conference on Computer Vision. IEEE, pp. 1440–1448.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. J. Mach. Learn. Res.

González, A., Fang, Z., Socarras, Y., Serrat, J., Vázquez, D., Xu, J., López, A.M., 2016. Pedestrian detection at day/night time with visible and fir cameras: a comparison. Sensors 16 (6), 820.

Guan, D., Cao, Y., Yang, J., Cao, Y., Tisse, C.L., 2018a. Exploiting fusion architectures for multispectral pedestrian detection and segmentation. Appl. Opt. 57 (18), D108.

Guan, D., Cao, Y., Yang, J., Cao, Y., Yang, M.Y., 2018b. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Inform. Fusion.

Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T., 2017. Mfnet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 5108–5115.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask r-cnn. In: IEEE International Conference on Computer Vision. IEEE, pp. 2980–2988.

Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., Torr, P., 2017. Deeply supervised salient object detection with short connections. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 5300–5309.

Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I., 2015. Multispectral pedestrian detection: benchmark dataset and baseline. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1037–1045.

Jafari, O.H., Yang, M.Y., 2016. Real-time rgb-d based template matching pedestrian detection. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 5520–5527.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 1175–1183.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. ACM, pp. 675–678.

Jingjing, L., Shaoting, Z., Shu, W., Dimitris, M., 2016a. Multispectral deep neural networks for pedestrian detection. In: British Machine Vision Conference, pp. 73.1.

Jingjing, L., Shaoting, Z., Shu, W., Dimitris, M., 2016b. Multispectral deep neural networks for pedestrian detection. In: British Machine Vision Conference, pp. 73.1.

Klinger, T., Rottensteiner, F., Heipke, C., 2017. Probabilistic multi-person localisation and tracking in image sequences. ISPRS J. Photogram. Remote Sens. 127, 73–88.

König, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M., 2017. Fully convolutional region proposal networks for multispectral person detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 243–250.

Krotosky, S.J., Trivedi, M.M., 2008. Person surveillance using visual and infrared imagery. IEEE Trans. Circ. Syst. Video Technol. 18 (8), 1096–1105.

Law, H., Deng, J., 2018. Cornernet: Detecting objects as paired keypoints. arXiv preprint arXiv: 1808.01244.

Leykin, A., Ran, Y., Hammoud, R., 2007. Thermal-visible video fusion for moving target tracking and pedestrian classification. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.

Li, X., Li, L., Flohr, F., Wang, J., Xiong, H., Bernhard, M., Pan, S., Gavrila, D.M., Li, K., 2017a. A unified framework for concurrent pedestrian and cyclist detection. IEEE Trans. Intell. Transp. Syst. 18 (2), 269–281.

Li, X., Ye, M., Liu, Y., Zhang, F., Liu, D., Tang, S., 2017b. Accurate object detection using memory-based models in surveillance scenes. Pattern Recogn. 67, 73–84.

Li, C., Song, D., Tong, R., Tang, M., 2018a. Multispectral pedestrian detection via simultaneous detection and segmentation. arXiv preprint arXiv: 1808.04818.

Li, J., Liang, X., Shen, S., Xu, T., Feng, J., Yan, S., 2018b. Scale-aware fast r-cnn for pedestrian detection. IEEE Trans. Multimedia 20 (4), 985–996.

Li, C., Song, D., Tong, R., Tang, M., 2019. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. Pattern Recogn. 85, 161–171.

Lin, T., Dollar, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J., 2017. Feature pyramid networks for object detection. Comput. Vis. Pattern Recogn. 936–944.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2018. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell.

Liu, J., Zhang, S., Wang, S., Metaxas, D., 2018. Improved annotations of test set of kaist. < http://paul.rutgers.edu/~jl1322/multispectral.htm/ > .

Mao, J., Xiao, T., Jiang, Y., Cao, Z., 2017. What can help pedestrian detection? In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 6034–6043.

Nam, W., Dollár, P., Han, J.H., 2014. Local decorrelation for improved pedestrian detection. In: Advances in Neural Information Processing Systems, pp. 424–432.

Oliveira, M., Santos, V., Sappa, A.D., 2015. Multimodal inverse perspective mapping. Inform. Fusion 24, 108–121.

Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., Poggio, T., 1997. Pedestrian detection using wavelet templates. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 193–199.

Pascanu, R., Mikolov, T., Bengio, Y., 2013. On the difficulty of training recurrent neural networks. In: International Conference on Machine Learning, pp. 1310–1318.

Pinheiro, P.H.O., Lin, T., Collobert, R., Dollar, P., 2016. Learning to refine object segments. Eur. Conf. Comput. Vis. 9905, 75–91.

Rajchl, M., Lee, M.C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., et al., 2017. Deepcut: object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imag. 36 (2), 674–683.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vision 115 (3), 211–252.

Salton, G., McGill, M.J., 1986. Introduction to modern information retrieval.

Schindler, K., Ess, A., Leibe, B., Van Gool, L., 2010. Automatic detection and tracking of pedestrians from a moving stereo rig. ISPRS J. Photogram. Remote Sens. 65 (6), 523–537.

Shirazi, M.S., Morris, B.T., 2017. Looking at intersections: a survey of intersection monitoring, behavior and safety analysis of recent studies. IEEE Trans. Intell. Transp. Syst. 18 (1), 4–24.

Shu, G., Dehghan, A., Oreifej, O., Hand, E., Shah, M., 2012. Part-based multiple-person tracking with partial occlusion handling. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1815–1821.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556.

Torabi, A., Massé, G., Bilodeau, G.-A., 2012. An iterative integrated framework for

thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. Comput. Vis. Image Underst. 116 (2), 210–221.

Wagner, J., Fischer, V., Herman, M., Behnke, S., 2016. Multispectral pedestrian detection using deep fusion convolutional neural networks. In: European Symposium on Artificial Neural Networks, pp. 509–514.

Wang, X., Wang, M., Li, W., 2014. Scene-specific pedestrian detection for static video surveillance. IEEE Trans. Pattern Anal. Mach. Intell. 36 (2), 361–374.

Wu, B., Iandola, F., Jin, P.H., Keutzer, K., 2016. Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. arXiv preprint arXiv: 1612.01051.

Zhang, S., Benenson, R., Schiele, B., 2015. Filtered channel features for pedestrian detection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1751–1760.

Zhang, L., Lin, L., Liang, X., He, K., 2016. Is faster r-cnn doing well for pedestrian detection? In: European Conference on Computer Vision. Springer, pp. 443–457.

Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B., 2017a. Towards reaching human performance in pedestrian detection. IEEE Trans. Pattern Anal. Mach. Intell. (99), 1.

Zhang, S., Benenson, R., Schiele, B., 2017b. Citypersons: a diverse dataset for pedestrian detection. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3.

Zinkevich, M., Weimer, M., Li, L., Smola, A.J., 2010. Parallelized stochastic gradient descent. In: Advances in Neural Information Processing Systems, pp. 2595–2603.