Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation

Chengyang Li
licy_cs@zju.edu.cn
Dan Song
songdan1992@zju.edu.cn
Ruofeng Tong
trf@zju.edu.cn
Min Tang
tang_m@zju.edu.cn

State Key Lab of CAD&CG
Zhejiang University
Hangzhou, China

Abstract

Multispectral pedestrian detection has attracted increasing attention from the research community due to its crucial competence for many around-the-clock applications (e.g., video surveillance and autonomous driving), especially under insufficient illumination conditions. We create a human baseline over the KAIST dataset and reveal that there is still a large gap between current top detectors and human performance. To narrow this gap, we propose a network fusion architecture, which consists of a multispectral proposal network to generate pedestrian proposals, and a subsequent multispectral classification network to distinguish pedestrian instances from hard negatives. The unified network is learned by jointly optimizing pedestrian detection and semantic segmentation tasks. The final detections are obtained by integrating the outputs from different modalities as well as the two stages. The approach significantly outperforms state-of-the-art methods on the KAIST dataset while remain fast. Additionally, we contribute a sanitized version of training annotations for the KAIST dataset, and examine the effects caused by different kinds of annotation errors. Future research of this problem will benefit from the sanitized version which eliminates the interference of annotation errors.

1 Introduction

Pedestrian detection is a vigorously studied topic in the field of computer vision over the past few decades, with diversified potential applications such as video surveillance, autonomous driving and robotics. Nevertheless, the majority of existing detectors focus on color images only, and they probably fail to work under insufficient illumination conditions, e.g., nighttime.

Long-wavelength infrared (thermal) images provide an alternative choice to address this challenge. Thermal cameras capture the radiated heat of objects, thus they can present clear human silhouettes even in absence of natural light, yet they lose detailed visual characteristics (e.g., color and texture) that often presented by color images. This makes color images and thermal images complementary with each other by nature. With the introduction of the

KAIST Multispectral Pedestrian Benchmark [16], multispectral pedestrian detection has attracted increasing attention from the computer vision community [7, 14, 17, 18, 19, 26, 29]. Effectively fusing multispectral data for pedestrian detection is a non-trivial task. We create a human baseline and the results indicate that even the current state-of-the-art detectors [14, 18, 19] lag behind human performance by a wide gap. Therefore, there is still a large potential to improve the detection performance by better leveraging multispectral images.

In this work, we investigate how to effectively and efficiently detect pedestrians by leveraging RGB-thermal pairs with convolutional neural networks (convnets). We propose a network fusion architecture for multispectral pedestrian detection, which is denoted as the Multispectral Simultaneous Detection and Segmentation R-CNN (MSDS-RCNN). Specifically, MSDS-RCNN consists of two multispectral fusion networks, among where the former network is responsible for generating candidate proposals and the latter network focuses on handling hard examples. Recent work [9, 12, 24] has shown that semantic segmentation is beneficial for RGB based pedestrian detection. We not only confirm their conclusions on multispectral pedestrian dataset but also reveal that incorporating semantic segmentation task in proposal stage is credited for the majority of the performance improvement.

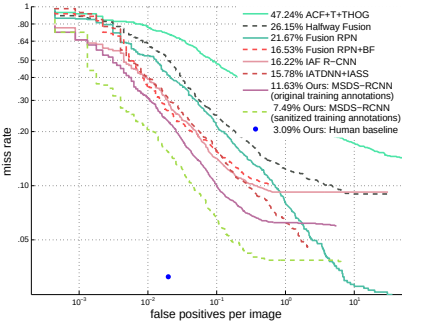

Figure 1: We have made efforts to narrow the gap between detection methods and human baseline (on the KAIST test set). Our method surpasses existing state-of-the-arts by a large margin (26% relative). About one third of the error is attributed to the annotation noise, which can be further eliminated using our sanitized training annotations.

The noise of training annotation is a nonnegligible factor that could lead to performance degeneration. We manually sanitize the KAIST training annotations. Taking MSDS-RCNN as a baseline, we examine the effects of different kinds of training annotation errors, including imprecise localization, misclassification and misaligned regions.

Our major contributions are fourfold: First, we introduce an effective and efficient architecture, called MSDS-RCNN, for multispectral pedestrian detection. Second, we create a human baseline for the KAIST dataset to reveal the gap between current detectors and human performance. Third, we provide a sanitized version of training annotations for the KAIST dataset and based on which, the effects of training annotation quality are evaluated. Last but not least, our MSDS-RCNN pushes the state-of-the-art performance on KAIST dataset from 15.78% to 11.63% in terms of log-average miss rate (26% relative reduction). Using the sanitized training annotations, the detection performance can be further boosted to 7.49%.

2 Related Work

Color Image based Pedestrian Detection: As a canonical case of general object detection, pedestrian detection is one of the hot topics in computer vision [3, 25]. The majority of past work for pedestrian detection is based on color image and recent top performing detectors are typically variants of Fast/Faster R-CNN [13, 27]. MS-CNN [6] and SAF-RCNN [20] handled the scale-variance problem via specifically designed multi-scale sub-networks. Zhang

et al. [30] showed that the under-performance of Faster R-CNN in pedestrian detection task is attributed to the Fast R-CNN classifier due to insufficient input resolution and lack of bootstrapping strategy. Competitive results can be achieved by cascading boosted forest [2] on top of the high-resolution RPN feature maps, denoted as RPN+BF. Zhang et al. [31] revealed that after proper adaptations such as pedestrian-specific RPN scales and input up-scaling, a plain Faster R-CNN gained substantial improvement and almost matched a state-of-the-art detector. F-DNN [12] and SDS-RCNN [4] used separate downstream classifiers that do not share weights with the proposal network, so that they can better handle hard examples. In this paper we will explore this insight in the scope of multispectral pedestrian detection.

Multispectral Pedestrian Detection: Since the release of the KAIST Multispectral Pedestrian Benchmark [16], there is a growing interest in pedestrian detection leveraging aligned color and thermal images. The initial baseline ACF+T+THOG was extended from the aggregated channel features (ACF) [11] and augmented with thermal channels. Wagner et al. [29] adopted ACF+T+THOG to generate region proposals, which were then re-scored by a convnet. Choi et al. [7] first used separate RPNs to generate proposals on color and thermal images and then evaluated them with support vector regression (SVR). A later extension [26] reformulated shallow modules as network architectures so that it can be trained end-to-end. Liu et al. [17] explored different network fusion architectures developed from Faster R-CNN and discovered that halfway fusion produced best performance. König et al. [18] extended RPN+BF to multispectral pedestrian detection and proposed Fusion RPN+BF. Almost at the same time, [14, 19] proposed illumination-aware fusion architectures that fused the outputs from color/thermal sub-networks or day/night sub-networks by a illumination-aware weighted function. In this work, we do not incorporate such illumination-aware weighting mechanism, yet our detection performance already surpasses theirs remarkably. Additional improvement can be expected if we adopt such mechanism in our approach.

Segmentation for Pedestrian Detection: Object detection and semantic segmentation are two highly correlated tasks and recently researchers have explored utilizing semantic segmentation for pedestrian detection. Since many pedestrian datasets do not provide segmentation masks, initial attempts [9, 12, 15, 24, 31] obtained segmentation using models pretrained on segmentation datasets such as Cityscapes [8], MS-COCO [22] and CamVid [5], and then took the generated masks as additional cue for inference. Recent work [4] resorted bounding box annotations of pedestrians as weak segmentation mask supervision, thus segmentation and detection tasks can be simultaneously trained by optimizing a joint loss function.

3 Preliminaries

3.1 Pedestrian Benchmark

In this work we focus on the KAIST dataset [16], which contains 95,328 aligned color-thermal image pairs, with manual annotations amount to a total of 103,128 bounding boxes covering 1,182 unique pedestrians. Following the method presented in [19], we sample images every 2 frames from training videos, exclude heavily occluded, truncated and small (< 50 pixels) pedestrian instances, and finally obtain 7,601 training images. The test set contains 2,252 images sampled every 20th frame from videos, among which 1,455 images are captured during daytime and the other 797 images are during nighttime. For evaluation, we strictly follow the reasonable setting provided by the KAIST benchmark, and measure the log-average miss rate (MR) over the range of $[10^{-2}, 10^{0}]$ false positives per image (FPPI).

Since the original annotations of the test set contain many problematic bounding boxes, we use the improved annotations provided by Liu et al. [23] to enable a reliable comparison.

3.2 Human Baseline

Before delving into our methodology, we try to figure out how much potential is a detector expected to improve. To this end, we construct a human baseline by asking human annotators to 'detect' on the KAIST test set, which can be viewed as a perfect detector. Considering that existing detectors are based on single image (color-thermal pair), we present frames in random order to the human annotators so that surrounding or temporal information is inaccessible. Since pedestrian instances might be invisible in one modality, we ask human annotators to double check both color and thermal images before drawing their detections. As expected, human performance widely surpasses existing state-of-the-arts. At ~0.02 FPPI, the current top performing detectors [14, 18, 19] produce 8× miss rate than human baseline (see Figure 1), indicating the automatic detector still has a large potential to improve. The superior of human performance owes to their priori knowledge, for example, a human annotator can easily distinguish human figure sculptures from real persons. Detection algorithms are expected to at least approach human performance.

4 Proposed Method

The proposed network architecture consists of two components: a multispectral proposal network (MPN) and a multispectral classification network (MCN). The overview of the proposed MSDS-RCNN is illustrated in Figure 2 and the details are explained below.

4.1 Multispectral Proposal Network

The MPN aims to generate candidate bounding boxes covering the majority of ground-truth pedestrian instances, by leveraging the information from both color and thermal modalities. Consequently, the generated proposals inevitably contain a large portion of false positives, which will be addressed by the subsequent MCN.

As shown in Figure 2, the MPN starts from two networks separately taking the color image or thermal image as input, which are based on the VGG-16 [28] architecture with fully connected layers removed. We fuse the two networks halfway, immediately after their third convolutional blocks, obtaining a merged stream with a balance between fine visual details and semantic information. Network fusion is conducted by first concatenating the feature maps and then reducing dimension using Network-in-Network (NIN) [21], so that the subsequent layers in the VGG-16 architecture can be reused. We do not truncate the original color stream and thermal stream during training phase, since they can be used to provide more diversified proposals for training the subsequent MCN. We further remove the fourth pooling layer to provide a finer feature stride of 8, which is shown beneficial for handling small instances [17, 31]. For each of color stream, thermal stream and merged stream, we build a standard proposal module on the top of each conv5_3 layer in VGG-16 architecture, which consists of a 3×3 intermediate convolutional layer followed by two sibling 1×1 convolutional layers for bounding box regression and classification respectively [27]. The anchors are tailed for pedestrian detection as follows. We split the full scale range of training data into 8 quantile bins and use the resulting 9 endpoints as RPN scales. Besides,

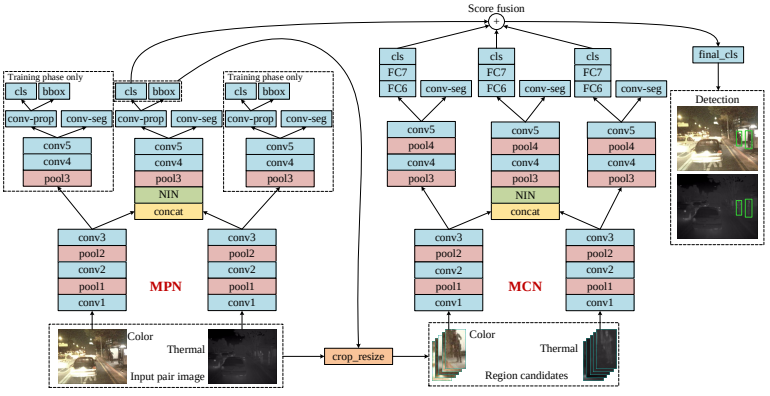

Figure 2: Overview of the proposed MSDS-RCNN.

we use a fixed aspect ratio of 0.41 following [30]. An anchor is assigned a positive label if it has an Intersection-over-Union (IoU) higher than 0.5 with any ground-truth box. Otherwise, we assign it a negative label. Additionally, a segmentation module is also added to the top of each conv5_3 layer, which is a single 1×1 convolutional layer.

The MPN is thus trained by minimizing the following joint loss function with nine terms:

$$\begin{aligned}\mathcal{L} =& \lambda_1 \mathcal{L}_{MPNcls}^{color} + \lambda_2 \mathcal{L}_{MPNcls}^{thermal} + \lambda_3 \mathcal{L}_{MPNcls}^{merged} \\ &+ \lambda_4 \mathcal{L}_{MPNbbox}^{color} + \lambda_5 \mathcal{L}_{MPNbbox}^{thermal} + \lambda_6 \mathcal{L}_{MPNbbox}^{merged} \\ &+ \lambda_7 \mathcal{L}_{MPNseg}^{color} + \lambda_8 \mathcal{L}_{MPNseg}^{thermal} + \lambda_9 \mathcal{L}_{MPNseg}^{merged}\end{aligned} \tag{1}$$

where the first six components remain the same as the PPN loss defined in Faster R-CNN [27], and the last three components are the pixel-level loss introduced by [24]. Let $G_{x,y}$, $P_{x,y}$ respectively represent the ground-truth and predicted segmentation masks, the segmentation loss is computed as: $\mathcal{L}_{seg} = \frac{1}{H \times W} \sum_{(x,y)} l(G_{x,y}, P_{x,y})$, where H and W denote the size of the feature map and l is the cross-entropy loss function. In our experiments, we set all $\lambda_i = 1$.

During inference, we only use the fusion stream to generate pedestrian candidates, as it remarkably speeds up the testing process without obvious performance degradation (see Section 5.3 for details).

4.2 Multispectral Classification Network

As a subsequent stage of the MPN, the MCN is designed to re-score the proposals generated by the MPN and it particularly focuses on handling hard examples.

Pedestrian candidates generated by the MPN with confidence score greater than 0.01 are passed to the MCN and those lower than 0.01 are filtered, for both training and inference phases. Following [4], we pad each candidate proposal by a factor of 0.2 on all sides to incorporate contextual information and avoid partial cropping. For each proposal, we scale it to a fixed size before taken as input for the MCN.

To construct the MCN, we start with two separate networks based on VGG-16, each takes the cropped candidate regions of color image or thermal image as input. Then we fuse the two networks halfway, as we performed in the MPN. On top of each FC7 layer in

color stream, thermal stream and merged stream, an output layer is built for binary proposal classification. A proposal is assigned a positive label if it has an IoU higher than 0.7 with any ground-truth box. Otherwise, we assign it a negative label. Also, a segmentation module is added to each conv5_3 layer, as we did in the MPN. The final loss for the MCN is thus computed as:

$$\begin{aligned}\mathcal{L} =& \lambda_1 \mathcal{L}_{MCNcls}^{color} + \lambda_2 \mathcal{L}_{MCNcls}^{thermal} + \lambda_3 \mathcal{L}_{MCNcls}^{merged} \\ &+ \lambda_4 \mathcal{L}_{MCNseg}^{color} + \lambda_5 \mathcal{L}_{MCNseg}^{thermal} + \lambda_6 \mathcal{L}_{MCNseg}^{merged}\end{aligned} \tag{2}$$

where the first three components are the classification loss and the last three components are the pixel-level segmentation loss averaged on batch instances. We set all $\lambda_i = 1$ in our experiments.

For efficiency purpose, we remove the fifth pooling layer from the VGG-16 architecture, modify the filter size of the fourth pooling layer to 2×1 and then adjust the input size to 112×56. During inference, we take top K proposals as input to further reduce computational cost. We should mention that if no more than K proposals generated from the MPN exist after filtering by the confidence threshold of 0.01, we take the remaining proposals as input.

Since color and thermal modalities exhibit different visual features, it is expected that the classification characteristics from color, thermal and merged streams to be complementary when fused. Moreover, as the MPN and the MCN is designed for handling general cases and hard examples respectively, the classification results from the MPN and the MCN are also complementary. Therefore, we fuse the classification scores from different stages and modalities. Given the predicted 2-class scores from the three streams of MCN: $S_{MCN}^{color} = \{c_0^c, c_1^c\}$, $S_{MCN}^{thermal} = \{c_0^t, c_1^t\}$, $S_{MCN}^{merged} = \{c_0^m, c_1^m\}$, as well as the ones from the MPN: $S_{MPN} = \{c_0^p, c_1^p\}$, the final classification score is obtained via the softmax function:

$$c_1^f = \frac{e^{(c_1^p + c_1^c + c_1^t + c_1^m))}}{e^{(c_0^p + c_0^c + c_0^t + c_0^m))} + e^{(c_1^p + c_1^c + c_1^t + c_1^m))}} \tag{3}$$

5 Experiments

5.1 Implementation Details

The proposed MSDS-RCNN is implemented in the Tensorflow [1] framework. The training process contains two main stages and we adopt the image-centric training scheme. In the first stage, we train the MPN using SGD with a momentum of 0.9 and a weight decay of 0.0001. For each image, we randomly sample 120 anchors with the ratio of positive and negative ones as 1:5. The MPN model is initialized with a VGG-16 model pretrained on the ImageNet dataset [10]. We start training with a learning rate of 0.001, divide it by 10 after 4 epochs, and terminate training after 6 epochs. In the second stage, we train the MCN using almost the same setting as the MPN. The MCN model is initialized with the MPN model generated in the first stage. For each image, we randomly sample 60 proposals generated from the MPN with the ratio of positive and negative ones as 1:2. During inference, we set input image scale $S = 600$ pixels for the MPN and the number of proposals $K = 50$ for the MCN, considering the speed/accuracy trade-off (see Section 5.3 for explanations). Since semantic segmentation masks are unavailable in the KAIST dataset, we use pedestrian bounding box annotations as

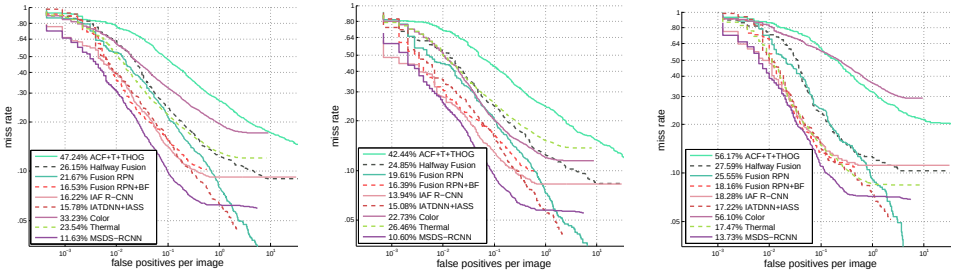

Figure 3: Comparisons of detection results reported on the test set of KAIST dataset, in terms of Reasonable-all (left), Reasonable-day (middle), and Reasonable-night (right).

weak segmentation ground-truth masks following [4]. We consider the 'person', 'person?' and 'people' categories in the KAIST dataset as foreground, and the remaining classes as background. We report the averaged performance after repeating the experiments for 5 times.

5.2 Comparisons with State-of-the-arts

We evaluate the proposed MSDS-RCNN on the test set of KAIST, compared with Halfway Fusion [17], ACF+T+THOG [16], Fusion RPN [18], Fusion RPN+BF [18], IAF R-CNN [19] and IATDNN+IASS [14]. We also implement two single modality baselines for comparison, denoted as Color and Thermal. For implementation, we simply remove layers in the MSDS-RCNN model and corresponding components in the loss function that involve the other modality, then train a single-modality model using the identical procedure.

Figure 3 compares the experimental results, in terms of MR under reasonable setting. It can be observed that MSDS-RCNN outperforms all existing methods and single-modality baselines by a large margin, both on daytime images and nighttime images. IATDNN+IASS is the best among existing detectors, with 15.78% MR. With the proposed method, we obtain 11.63% MR, improving the current state-of-the-art by 26% relative reduction of the error. Moreover, the efficiency of our method also surpasses IATDNN+IASS, with 228 ms/image vs. 250 ms/image on runtime with a single NVIDIA Geforce Titan X GPU.

5.3 Ablation Studies

This subsection is devoted to investigating the effectiveness of different design choices.

Effect of semantic segmentation: Table 1 compares the performance of enabling or disabling the segmentation supervision in the networks. The baseline that does not use the segmentation supervision obtains 13.59% MR. By introducing the segmentation supervision in both the MPN and the MCN, the detection performance improves to 11.63% MR, with 14% relative reduction of the error, indicating that the segmentation supervision is also beneficial for multispectral pedestrian detection. We also compare the effect of introducing the segmentation supervision in the MPN or the MCN. In this case, we enable the segmentation supervision in one network and disable the other. It can be observed that in both cases we obtain performance improvement, but infusing segmentation in the MPN brings considerably greater impact than that in the MCN (12.00% vs. 13.03%). We suppose this can be attributed to the coarse bounding box annotations in the KAIST dataset, which can cause more inconsistence when handling segmentation masks locally in the MCN.

Supervision		Detection performance (MR)		
MPN	MCN	Reasonable-all	Reasonable-day	Reasonable-night
		13.59%	11.95%	16.96%
√		12.00%	11.10%	14.14%
	√	13.03%	11.74%	15.65%
√	√	11.63%	10.60%	13.73%

Table 1: Effectiveness of the segmentation supervision.

Effect of score fusion: As illustrated in Table 2, combining the merged stream with color stream and thermal stream pushes the performance from 14.02% MR to 11.95% MR. This phenomenon reveals that although the merged stream makes use of color and thermal information, the classification characteristics of color stream and thermal stream are still complementary to the merged stream. The scores from the MPN is also complementary, combining it slightly boosts the detection performance to 11.63%.

MPN	MPN			Detection performance (MR)		
	Color	Thermal	Merged	Reasonable-all	Reasonable-day	Reasonable-night
√				18.88%	16.63%	22.89%
	√			24.32%	18.43%	36.28%
		√		22.50%	24.56%	17.76%
			√	14.02%	13.78%	14.54%
	√	√	√	11.95%	10.99%	13.83%
√	√	√	√	11.63%	10.60%	13.73%

Table 2: Effectiveness of the score fusion scheme.

Speed/accuracy trade-off: Finally we evaluate the efficiency of the method. The runtime of the MPN is varied by the scale of input image, while that of the MCN depends on the number of input proposals. Figure 4 compares the performance using different input scales and numbers of proposals. We also compare the effects of using only the merged stream (denoted as 'Merged') and using all three streams (denoted as 'All') in the MPN to generate proposals. Generally, the larger input scale or more proposals bring performance improvement but add more computational cost. Besides, using 'Merge' proposals typically obtains comparable or even better performance than 'All' proposals. Considering the speed/accuracy trade-off, we adopt 600 input image scale for the MPN, 'Merge' proposal mode and 50 proposals for the MCN, which results in a process speed of 228 ms/image.

5.4 Impact of Training Annotation Noise

The annotation noise in the KAIST dataset is a vital factor that could affect the detection performance. The original annotations of KAIST dataset contain many problematic bounding boxes, such as missing annotation and incorrect labeling. Hence, Liu et al. [23] provided improved annotations of KAIST test set to enable a reliable evaluation. As revealed by our previous technical report [19], there is a big difference (>15%) in regard of MR value between using original and improved testing annotations. Similarly, the annotation noise in the training data would lead to error-prone optimizing process.

To study the effects of annotation noise, we create a sanitized version of KAIST training annotations. Since annotating the whole training data is time-consuming (95,328 frames), we first filter the training images using the original annotations and obtain 7,601 valid frames

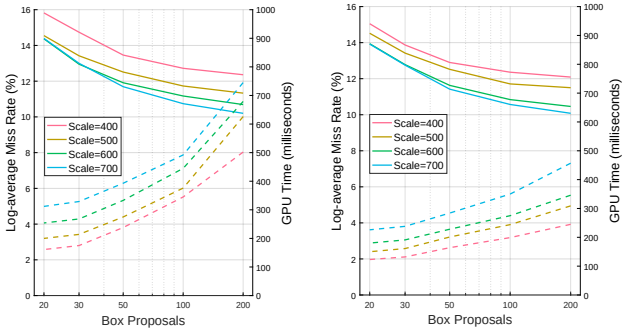

Figure 4: Effect of MPN input scale, number of regions and proposal mode (left: All, right: Merged) on log-average miss rate (solid lines) and GPU inference time (dotted).

(same protocol as described in Section 3.1). Then we carefully re-label all these 7,601 frames to provide a high quality version of ground-truth annotations. The annotation errors we corrected can be divided into three categories as follows:

1) Imprecise localization: In the original annotations, there are many annotated bounding boxes that do not well match the real regions of person instances. The most common case is using a obviously larger box to annotate a small instance. We correct this kind of error so that each instance is tightly bounded by a box (see Figure 5 (left)).

2) Misclassification: The corrected misclassification cases are those assigning an incorrect category or occlusion state to an person. The cases also include missing annotations, i.e., incorrectly labeling a person as background (see Figure 5 (middle)).

3) Misaligned regions: Although efforts have been made to ensure the paired color and thermal images align both spatially and temporally, we find there still exist cases that the multispectral images are not well aligned, particularly when the car is making a turn. For such case, we separately label the bounding boxes in the color image and the thermal image, and then check their IoU value. If they have an IoU lower than 0.5, we use the minimum box that bounds both boxes to represent the instance and label it as '*person?a*' so that it can be ignored during training (see Figure 5 (right)).

Taking MSDS-RCNN as a baseline, we quantitatively study the effects of training an-

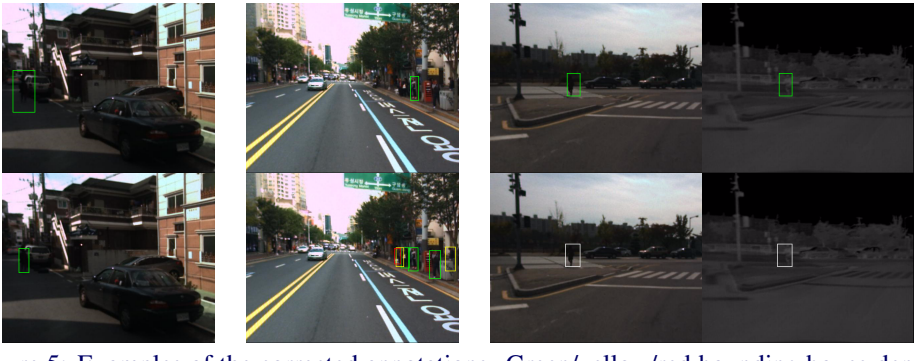

Figure 5: Examples of the corrected annotations. Green/yellow/red bounding boxes denote non/partial/heavy-occluded pedestrians and white boxes denote ignore regions. Top row: the original annotations. Bottom row: the sanitized annotations.

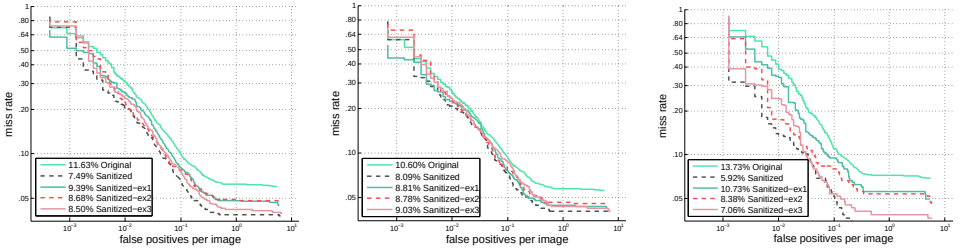

Figure 6: Impact of different training annotations, compared on the test set of KAIST dataset, in terms of Reasonable-all (left), Reasonable-day (middle), and Reasonable-night (right).

notation noise. Apart from comparing the performance using the original and the sanitized annotations, we also examine the results using semi-sanitized annotations that excludes a specific type of error correction (denoted as 'Sanitized-ex1', 'Sanitized-ex2' and 'Sanitized-ex3' respectively). The experimental results are shown in Figure 6. Not surprisingly, using the sanitized training annotations, the detection performance improves significantly from 11.63% MR to 7.45% MR, which indicates that the training annotation noise is responsible for about one third of the inference error. For daytime images, using sanitized annotations gains 24% relative error reduction, and the three types of annotation errors contribute quite similar degrees. For nighttime images, using sanitized annotations obtains an amazing 56% relative error reduction, the most of which is due to the correction of imprecise localization.

6 Conclusion

In this work, we make efforts to narrow the gap between automatic pedestrian detectors and human performance. We present a unified convnet fusion architecture, denoted the MSDS-RCNN, for person detection in multispectral data (color-thermal image pairs). We show that jointly optimizing segmentation and detection tasks as well as effectively fusing the outputs from different branches bring substantial performance improvement, leading to 26% relative reduction of MR compared with existing state-of-the-art detector while remaining faster. Since the original training data contains many problematic annotations, we further study the impact of training annotation noise by carefully creating a sanitized version of ground-truth annotations. We find that the sanitized training annotations benefit the detection performance remarkably, especially for the nighttime images. We hope that future research can benefit from the provided data.

Acknowledgement

The research is supported in part by NSFC (61572424) and the Science and Technology Department of Zhejiang Province (2018C01080).

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow:

Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Ron Appel, Thomas Fuchs, Piotr Dollár, and Pietro Perona. Quickly boosting decision trees–pruning underachieving features early. In *ICML*, pages 594–602, 2013.

[3] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? *arXiv preprint arXiv:1411.4304*, 2014.

[4] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection & segmentation. In *ICCV*, pages 4950–4959, 2017.

[5] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2): 88–97, 2009.

[6] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, pages 354–370, 2016.

[7] Hangil Choi, Seungryong Kim, Kihong Park, and Kwanghoon Sohn. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In *ICPR*, pages 621–626, 2016.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[9] Arthur Daniel Costea and Sergiu Nedevschi. Semantic channels for fast pedestrian detection. In *CVPR*, pages 2360–2368, 2016.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[11] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (8):1532–1545, 2014.

[12] Xianzhi Du, Mostafa El-Khamy, Jungwon Lee, and Larry Davis. Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. In *WACV*, pages 953–961, 2017.

[13] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[14] Dayan Guan, Yanpeng Cao, Jun Liang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *arXiv preprint arXiv:1802.09972*, 2018.

[15] Qichang Hu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Fatih Porikli. Pushing the limits of deep cnns for pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(6):1358–1368, 2018.

[16] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, pages 1037–1045, 2015.

[17] Shu Wang Jingjing Liu, Shaoting Zhang and Dimitris Metaxas. Multispectral deep neural networks for pedestrian detection. In *BMVC*, pages 73.1–73.13, 2016.

[18] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *CVPRW*, pages 243–250, 2017.

[19] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *arXiv preprint arXiv:1803.05347*, 2018.

[20] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 20 (4):985–996, 2018.

[21] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[23] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris Metaxas. Improved annotations of test set of kaist. http://paul.rutgers.edu/~jl1322/multispectral.htm. Accessed February 9, 2018.

[24] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *CVPR*, pages 3127–3136, 2017.

[25] Duc Thanh Nguyen, Wanqing Li, and Philip O Ogunbona. Human detection from images and videos: A survey. *Pattern Recognition*, 51:148–175, 2016.

[26] Kihong Park, Seungryong Kim, and Kwanghoon Sohn. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognition*, 80:143–155, 2018.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[29] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*, pages 509–514, 2016.

[30] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *ECCV*, pages 443–457, 2016.

[31] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, pages 4457–4465, 2017.