Multispectral Deep Neural Networks for Pedestrian Detection

Jingjing Liu[1]
http://paul.rutgers.edu/~jl1322
Shaoting Zhang[2]
szhang16@uncc.edu
Shu Wang[1]
sw498@cs.rutgers.edu
Dimitris N. Metaxas[1]
https://www.cs.rutgers.edu/~dnm

[1] Department of Computer Science
Rutgers University
Piscataway, NJ, USA

[2] Department of Computer Science
UNC Charlotte
Charlotte, NC, USA

Abstract

Multispectral pedestrian detection is essential for around-the-clock applications, *e.g.*, surveillance and autonomous driving. We deeply analyze Faster R-CNN for multispectral pedestrian detection task and then model it into a convolutional network (ConvNet) fusion problem. Further, we discover that ConvNet-based pedestrian detectors trained by color or thermal images separately provide complementary information in discriminating human instances. Thus there is a large potential to improve pedestrian detection by using color and thermal images in DNNs simultaneously. We carefully design four ConvNet fusion architectures that integrate two-branch ConvNets on different DNNs stages, all of which yield better performance compared with the baseline detector. Our experimental results on KAIST pedestrian benchmark show that the Halfway Fusion model that performs fusion on the middle-level convolutional features outperforms the baseline method by 11% and yields a missing rate 3.5% lower than the other proposed architectures.

1 Introduction

As a canonical case of general object detection problem, in the past decades, pedestrian detection has attracted consistent attention from computer vision community [2, 4, 5, 9, 23, 24, 25, 26, 29, 43, 45]. It is the principle technique for various applications, such as surveillance, tracking, autonomous driving, etc. Although numerous efforts have been made for this problem and significant improvement has been achieved in recent years [3], there still exists an insurmountable gap between current machine intelligence and human perception ability on pedestrian detection [48]. Many challenges prevent artificial vision systems from practical applications, including occlusion, low image resolution, and cluttered background. Besides, since most of current pedestrian detectors explored color images of good lighting, they are very likely to be stuck with images captured at night, due to bad visibility of pedestrians. Such defect would cut these approaches off from the around-the-clock applications, *e.g.*, self-driven car and surveillance system.

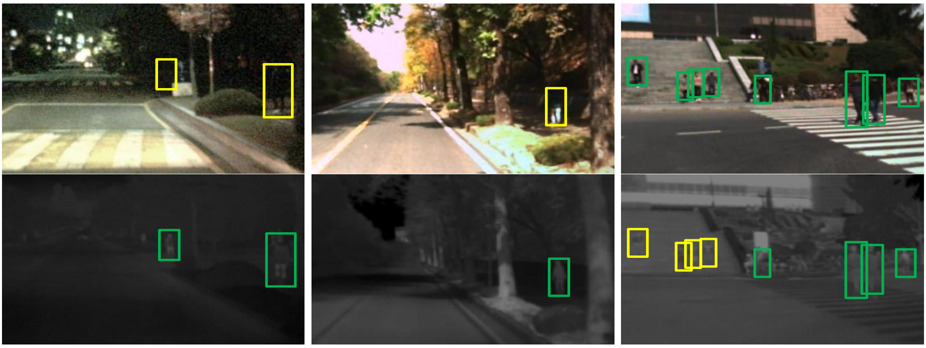

Figure 1: Left and middle: the thermal images capture human shapes even in bad lighting, while the corresponding color images are messed up. Right: with bright background, color image provides more distinctive visual features for the pedestrians (standing on stairs) against background objects; in such scenario, human silhouettes in the thermal image are ambiguous. Yellow bounding boxes indicate detection failures with one image channel.

In fact, aforementioned difficulties not only exist in pedestrian detection, but also in many other vision tasks. To tackle these problems, other types of sensors beyond RGB cameras were developed, such as depth cameras (time-of-flight, or near infrared, *e.g.*, Kinect) and thermal cameras (long-wavelength infrared). Since ambient lighting has less effect on thermal imaging, thermal cameras are widely used in human tracking [41], face recognition [31], and activity recognition [14] for its robustness. With regard to pedestrian detection, thermal images usually present clear silhouettes of human objects [15, 35], but lose fine visual details of human objects (*e.g.*, clothing) which can be captured by RGB cameras (depending on external illumination). As shown in Figure 1, instances of yellow bounding boxes would fail in detection with one image channel (color or thermal), while the other might help.

In some sense, color and thermal image channels provide complementary visual information. Nevertheless, except very recent effort [12, 42], most of previous studies focus on detecting pedestrians with color or thermal images only, rather than leveraging color and thermal images simultaneously. Furthermore, although researchers have applied deep neural networks (DNNs) on vision problems with multimodal data sources, *e.g.*, action recognition [33], image classification [37], etc., it is still unknown how color and thermal image channels can be properly fused in DNNs to achieve the best pedestrian detection synergy.

In this paper, we focus on how to make the most of multispectral images (color and thermal) for pedestrian detection. With the recent success of DNNs [18] on generic object detection [11], it becomes very natural and interesting to exploit the effectiveness of DNNs for multispectral pedestrian detection. We first adapt Faster R-CNN [30] model into our vanilla convolutional network (ConvNet), and train two separate pedestrian detectors on color or thermal images, respectively. It is not surprising to discover that these two ConvNet-based detectors provide complementary detection decisions, and there is a large potential to improve the detection performance by leveraging multispectral images, especially for around-the-clock applications. However, it is not trivial to explore the most effective DNNs architecture that simultaneously utilizes color and thermal images for pedestrian detection. Then the challenge of multispectral pedestrian detection task becomes a ConvNet fusion problem. From the perspective of fusions on different DNNs levels, four ConvNet fusion architectures are carefully designed upon our vanilla ConvNet, and then well investigated with extensive experimental evaluation. Our major contribution is fourfold:

- First, we carefully design four distinct ConvNet fusion architectures that integrate two-branch ConvNets on different DNNs stages, *i.e.*, convolutional stages, fully-connected stages, and decision stage, corresponding to information fusion on low level, middle level, high level, and confidence level. All these models outperform the strong baseline detector Faster R-CNN on KAIST multispectral pedestrian dataset (KAIST) [17].
- Second, we reveal that our Halfway Fusion model – fusion of middle-level convolutional features, provides the best performance on multispectral pedestrian detection. This implies that the best choice of fusion scheme is a balance between fine visual details and semantic information.
- Third, our Halfway Fusion model significantly reduces the missing rate of baseline method Faster R-CNN by 11%, yielding a 37% overall missing rate on KAIST, which is also 3.5% lower than the other proposed fusion models.
- Last but not least, our vanilla ConvNet achieves state-of-art performance (17% missing rate) on Caltech pedestrian benchmark [6]. As far as we know, this is the first time Faster R-CNN [30] has been investigated for pedestrian detection.

2 Related Work

DNNs for Pedestrian Detection: One pioneer work used deep neural network for pedestrian detection was proposed in [32], which combined multi-stage unsupervised feature learning with multi-scale DNNs. In [28], Ouyang *et al.* modeled visibility relationships among pedestrians using DNNs, which improved the visual confidences of human parts. Tian *et al.* [39] trained 45 complementary part detectors with weakly annotated humans, to handle partial occlusions. Tian *et al.* [40] improved pedestrian detection by learning high-level features from DNNs of multiple tasks, including pedestrian attribute prediction. Angelova *et al.* [1] built DNNs cascades that filtered candidates by tiny DNNs and steed up detection to 15 FPS. Li *et al.* [21] proposed scale-aware DNNs with a scale gate function, to capture characteristic features for pedestrians of different image sizes. Recent DNNs-based pipelines focused on incorporating DNNs with some pedestrian detectors [3, 7] that were used to generate class-specific proposals. These proposals were then passed into DNNs for classification. Along this research stream, many DNNs architectures have been investigated, including CifarNet [16], AlexNet [20], and VGG-16 [34]. General speaking, deeper DNNs performed better than shallow DNNs. However, these methods depended largely on qualities of proposals from a pre-trained pedestrian detector. In other words, a good detector should produce qualified hard negatives for training and enough positives for testing as many as possible.

DNNs with Multimodal inputs: It is an essential challenge for many vision problems to integrate information from multimodal data sources. A bunch of DNNs-based multimodal models have been proposed, involving in data sources of different modalities, such as image *vs*. audio [27], image *vs*. text [37, 44], image *vs*. video [19, 33], etc. In [27], Naiam *et al.* learned a hidden layer from deep belief networks (DBNs) as the shared representations of videos and audios. Similar network was also applied in [37] for image classification and retrieval, while missing modality was tolerated. Wang *et al.* [44] imposed structure-preserving constraints into their similarity objective function for images and texts, achieving better phrase localization results in images. For video recognition, Simonyan *et al.* [33]

proposed two-stream Covnets that incorporated spatial and temporal information. In their method, optical flow ConvNet was combined with key frame based ConvNets. Color and depth images have also been exploited simultaneously for 3D object classification [36] and 3D object detection [13]. Most of the aforementioned methods used two-branch network and then fused features from different channels at very end, *i.e.*, last feature layer fusion [13, 36, 44, 46] or confidence fusion [33]. Karpathy *et al.* [19] discussed some fusion schemes for video classification. However, DNNs-based multispectral pedestrian detection has not been studied thoroughly. It is still an open question that how color and thermal image channels could be fused properly in DNNs for pedestrian detection, to obtain 'optimal' synergy.

3 Methodology

3.1 Vanilla ConvNet

There exists many ConvNet architectures that are applicable for multispectral pedestrian detection. Some of them have been discussed in Section 2. Inspired by the recent success on general object detection, we consider starting with Faster R-CNN [30] and verify its performance on Caltech pedestrian benchmark [6].

Faster R-CNN model is consisted of a Region Proposal Network (RPN) and a Fast R-CNN detection network [10]. RPN is a fully convolutional network that shares convolutional features with the detection network. Compared to its methodological ancestries, *i.e.*, R-CNN [11] and Fast R-CNN [10], Faster R-CNN could produce proposals of high-quality by RPN. This is different from other ConvNet-based pedestrian detectors [16, 21] that reply an independent proposal generators. RPN enables nearly cost-free region proposals which generates 300 proposals in less than 0.3s. Besides, Faster R-CNN uses Region of Interest (RoI) Pooling layer that pools the feature map of each proposal into a fixed size, *i.e.*, 7×7, thus could handle pedestrians of arbitrary sizes.

Implementation details: We exploited the original Faster R-CNN model with a few twists and adapt it into our vanilla ConvNet. We removed the fourth max pooling layer of the very deep VGG-16 model [34]. This is encouraged by the observation in [21] that larger feature maps are beneficial for detecting pedestrians of small image sizes. Faster R-CNN uses reference anchors of multi-scale ($\times 3$) and multi-ratios ($\times 3$) to predict locations of region proposals. Given the typical aspect ratio of pedestrians, we discarded the anchor ratio 0.5 to accelerate the training and testing of RPN.

Caltech$\times$10 training set was used for fine-tuning. We excluded occluded, truncated, and small (< 50 pixels) pedestrian instances, resulting in around 7000 training images. Proposals of intersection-over-union (IoU) with any ground truth pedestrians larger than 0.5 were regarded as positives, otherwise they were used as negative samples. Follow-

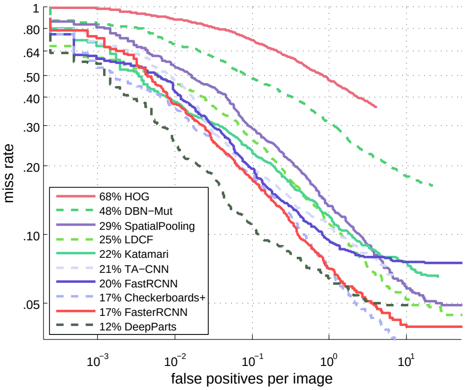

Figure 2: Comparison of detection results reported on the test set of Caltech pedestrian benchmark. Our vanilla ConvNet achieved 17% MR.

ing the alternative training routine of Faster R-CNN, the networks were initialized by the pre-trained VGG-16 model and then fine-tuned with Stochastic Gradient Descent (SGD) for about 6 epochs. The learning rate (LR) was set to 0.001 and reduced to 0.0001 after 4 epochs. Single image scale (600 pixels) was used, without considering feature pyramids.

Comparison of Detections: We compared our vanilla ConvNet (FasterRCNN) with some other methods reported on Caltech test set, including HOG [5], DBN-Mut [28], SpatioPooling [29], LDCF [26], Katamari [3], TA-CNN [40], FastRCNN [10], Checkerboards+ [47], and DeepParts [39]. For FastRCNN, we used ACF pedestrian detector [8] to obtain proposals which were then used to train and test Fast R-CNN detection network. A low threshold (-50) was set for ACF detector, so as to produce enough proposals. We used IoU 0.5 to validate detections. Detection performance was measured by log-average miss-rate over the range of $[10^{-1}, 10^{0}]$ (MR, lower is better) under reasonable configuration [7]. As shown in Figure 2, with completely data-driven and end-to-end training, our vanilla ConvNet beat most of state-of-art approaches that depend on sophisticated features or network design. We achieved 17% MR, which is lower than some DNNs-based methods, such as DBN-Mut(48%), TA-CNN(21%), and FastRCNN (20%). With a single ConvNet, the performance of our vanilla ConvNet is approaching DeepParts (12%) that is an assembly of 45 part ConvNets.

Given its state-of-art performance and some merits, *e.g.*, end-to-end training and capability of handling pedestrians of arbitrary sizes, in our paper, the vanilla ConvNet is used as the fundamental DNNs architecture in designing multispectral detectors.

3.2 Multispectral ConvNets

Intuitively, color and thermal image channels provide auxiliary visual information to each other in depicting pedestrian objects. If one image channel fails in detection, *i.e.*, missing true detections or recalling false alarms, the other could still make the correct decision. In this section, first of all, we are trying to answer the following questions: when strong ConvNet-based detectors are involved, does color and thermal images still provide complementary information? To what extend the improvement should be expected by fusing them together? Next, we model the multispectral pedestrian detection task as a ConvNet fusion problem. We carefully design four distinct ConvNet fusion models that integrate two-branch ConvNets at different DNNs stages. Each model represent one multispectral pedestrian detector.

3.2.1 Are They Really Complementary?

To study the complementary potential between color and thermal images, we first trained two separate pedestrian detectors with color or thermal images only, based on our vanilla ConvNet, namely FasterRCNN-C and FasterRCNN-T. The training set of KAIST dataset [17] (more details will be given in Section 4) was used in fine-tuning of neural networks, while the results were validated on the test images. In our following analyses, only detections of more than 0.5 confidence scores were considered. We regarded detections of IoUs with any ground truth (GT) larger than 0.5 as true positives (TPs), otherwise as false positives (FPs). Multiple detections on the same GT were treated as FPs. In Table 1, we enumerate the numbers of GTs, TPs, and FPs of FasterRCNN-C and FasterRCNN-T, in terms of all-day, daytime, and nighttime images. $\mathrm{TP}_{(C,T)}$ denotes pedestrians detected by both of the two detectors. $\mathrm{TP}_{(C)}$ and $\mathrm{TP}_{(T)}$ represent instances exclusively detected by FasterRCNN-C or FasterRCNN-T. Analogously, we have $\mathrm{FP}_{(C,T)}$, $\mathrm{FP}_{(C)}$, and $\mathrm{FP}_{(T)}$ for false alarms.

	GT	$TP_{(C,T)}$	$TP_{(C)}$	$TP_{(T)}$	$FP_{(C,T)}$	$FP_{(C)}$	$FP_{(T)}$
All	2,757	924	390	397	345	1,169	1,158
Day	2,003	720	346	176	303	745	827
Night	754	204	44	221	42	424	331

Table 1: Numbers of ground truths, true detections, and false alarms reported on the test images of KAIST pedestrian dataset, in terms of all-day, daytime, and nighttime, respectively.

Obviously, FasterRCNN-C and FasterRCNN-T have consensuses on pedestrian detections to some extent, but not alway. They have overall 924 common TPs, while 390 pedestrians captured by FasterRCNN-C were regarded as background by FasterRCNN-T. In daytime, FasterRCNN-C gets more TPs than FasterRCNN-T (1,066 *vs*. 896), while the trend is opposite on nighttime images (248 *vs*. 425). It is reasonable that during daytime most pedestrians are in good lighting conditions, except some corner cases (standing in shadow), when thermal images are apt to be affected by sunlight. In contrast, thermal images could capture better visual features of pedestrians at night. Besides, FasterRCNN-C and FasterRCNN-T share relatively fewer FPs (345), while they get FPs 2,672 in total. It is not hard to infer that there is a large potential in excluding FPs by using two image channels.

Without doubt, color and thermal images provide complementary information on pedestrian detection. Based on the 2,252 test images, if we make an extreme assumption that all true detections from either FasterRCNN-C or FasterRCNN-T were kept and only shared false alarms were retained, the detection rate can be increased from 47.9% to 62.1%, with the FPPI (false positives per image) reduced from 0.549 to 0.125. Hence, we should pay serious attention on the potential improvement that would be raised from multispectral detection.

3.2.2 ConvNet Fusion Models

The question now is how a good multispectral pedestrian detector that explores color-thermal image pairs could be achieved. As shown in Figure 3, a ConvNet-based detector is composed of three stages: the convolutional stage, the fully-connected stage, and the decision stage. Features at different stages corresponding to various levels of semantic meanings and fine visual details. We think fusion at different stages would lead to different detection results. Therefore, the multispectral pedestrian detection task comes to be a ConvNet fusion problem, *i.e.*, what architecture of the fusion model could get best detection synergy. To this end, we make thorough inquiries on four fusion models designed upon our vanilla ConvNet. Basi-

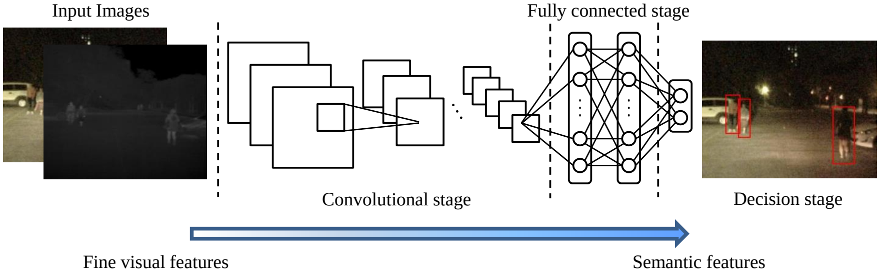

Figure 3: Different stages in a ConNet. Features at different stages correspond to various levels of semantic meanings and fine visual details.

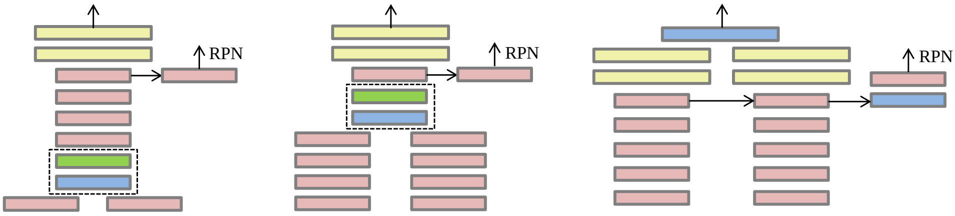

Figure 5: Explored approaches to fuse color and thermal images for multispectral pedestrian detection. These approaches implement feature fusions. From left to right are feature fusions at low level (Early Fusion), middle level (Halfway Fusion), and high level (Late Fusion), respectively. Red and yellow boxes represent convolutional and fully-connected layers. Blue boxes represent concatenate layer. Green boxes denote Network-in-Network (NIN) used for dimension reduction. For the sake of conciseness, ReLU layers, pooling layers, and dropout layers are hidden from view in this figure. (Best viewed in color.)

cally, they are two-branch ConvNet models that perform fusions at different stages, denoted as Early Fusion, Halfway Fusion, Late Fusion, and Score Fusion. Three of them implement feature fusions, as shown in Figure 5, while the other one combines confidence scores from color and thermal ConvNet branches at decision stage, as shown in Figure 4.

Early Fusion concatenates the feature maps from color and thermal branches immediately after the first convolutional layers (C1). Afterwards, we introduce Network-in-Network (NIN) [22, 38] after feature concatenation, which is actually a 1 × 1 convolutional layer. NIN reduces the dimension of concatenate layer to 128, such that other filters from the pre-trained VGG-16 model can be reused. Besides, NIN outputs linear combinations of local features from color and thermal branches. Followed by ReLU, it can enhance the discriminability of local patches. Since C1 captures low-level visual features, such as corners, line segments, etc., Early Fusion model fuses features at low-level.

Halfway Fusion also implements fusion at convolutional stage. Different from Early fusion, it puts the fusion module after the fourth convolutional layers (C4). NIN is also used after the concatenate layer, for the same reasons as discussed before. Features from C4 layers contain more semantic meanings than C1 features, while retaining some fine visual details.

Late Fusion concatenates the last fully-connected layers (F7), which performs feature fusion at fully-connected stage. Conventionally, F7 features are used as new representations of objects. Late Fusion executes high-level feature fusion. To be noticed, RPN here exploits C5 features from the two branches to predict human proposals.

Score Fusion can be regarded as a cascade of two ConvNets (Figure 4). We first get detections from color ConvNet which are then sent into the other ConvNet to obtain detection scores based on thermal image, and vice verse. In practice, this can be accomplished by using RoI Pooling layer. Detection scores from the two branches are merged with equal weights (*i.e.*, 0.5).

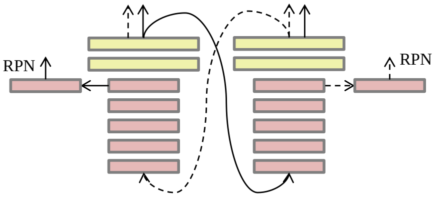

Figure 4: Explored approach for combining scores of two ConvNets (Score Fusion).

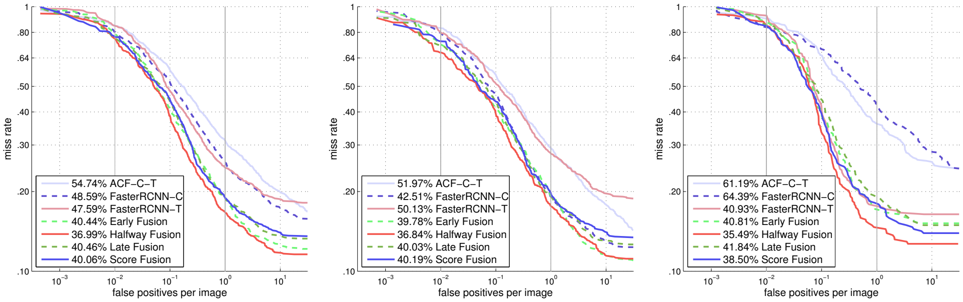

Figure 6: Comparison of detection results reported on the test set of KAIST multispectral pedestrian dataset, in terms of all-day (left), daytime (middle), and nighttime (right).

4 Experiments

Dataset: KAIST multispectral pedestrian dataset (KAIST) [17] contains 95,328 aligned color-thermal frame pairs, with 103,128 dense annotations on 1,182 unique pedestrians. We sampled images from training videos with 2-frame skips, and finally obtained 7,095 training images binded with qualified pedestrians (the same criteria as discussed in Section 3.1). The test set of KAIST contains 2,252 images sampled from test videos with 30-frame skips, among which 1,455 images were captured during daytime and 797 others for nighttime.

Implementation Notes: Most of filters in the ConvNet fusion models were initialized by the corresponding parameters of the pre-trained VGG-16 model, except new introduced layers. For instance, in Early Fusion and Halfway Fusion, the weights of NINs were initialized by a Gaussian distribution. Parallel branches in the four fusion models did not share weights. The two branches in Early Fusion, Halfway Fusion, and Late Fusion were trained simultaneously, while the two ConvNets in Score Fusion were trained individually. All the models were fine-tuned with SGD for 4 epochs with LR 0.001 and 2 more epochs with LR 0.0001. Besides, non-maximum suppression (NMS) was applied to the detections of Score Fusion model, in order to avoid double detections from color and thermal channels.

4.1 Evaluation of Detections:

We evaluated the proposed four ConvNet fusion models on the test set of KAIST, compared to FasterRCNN-C and FasterRCNN-T, as well as ACF-C-T detector reported in [17]. The ACF-C-T detector used 10-channel aggregated features to fuse color and thermal images. Comparisons of detection results are presented in Figure 6, in terms of MR under reasonable configuration [7].

Generally speaking, detectors with single image modality obtain inferior results than fusion models. FasterRCNN-C obtains 42.5% MR on daytime images, while working worse than the ACF-C-T detector (64.4% *vs*. 61.2%) on nighttime images. On the other side, FasterRCNN-T suffers on daytime images, although it gets similar MR as some fusion models on nighttime images. Consequently, both FasterRCNN-C and FasterRCNN-T are not applicable for around-the-clock applications. Compared to FasterRCNN-C and FasterRCNN-T, ConvNet fusion models produce significantly better results, which reduce the overall MR from 48% to around 40%.

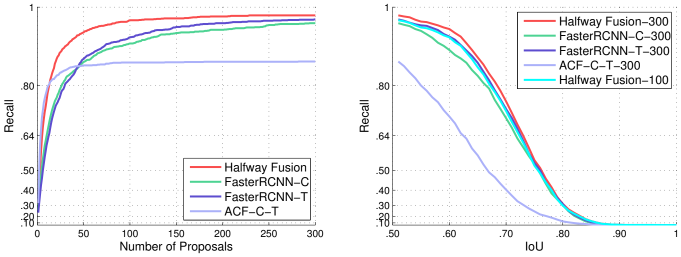

Figure 7: Comparison of pedestrian proposals reported on test set of KAIST multispectral pedestrian dataset. Left: Recall *vs*. Number of Proposals; Right: Recall *vs*. IoU.

Among the four ConvNet fusion models, Halfway Fusion achieves the lowest overall MR (36.99%), which is 3.5% lower than other fusion models, showing the most effective multispectral synergy for pedestrian detection. Since the four DNNs architectures correspond to information fusion on different ConvNet levels, we speculate that middle-level convolutional features from color and thermal branches are more compatible in fusion: they contain some semantic meanings and meanwhile do not completely throw all fine visual details. However, Early Fusion combines low-level features, such that some task irrelevant low-level features would be fused, which could undermine the fusion power. Late Fusion executes fusions on high-level semantic features and Score Fusion combines confidence scores. In some cases, it would be difficult for these two models to eliminate semantic noise or to adjust decision mistake of one image channel. Compared to FasterRCNN-C or FasterRCNN-T, given the whole test set, Halfway Fusion reduces the overall MR by around 11%,

Some detection samples are illustrated in Figure 8. Detections with confidences large than 0.5 are presented. Obviously, compared to the color image based detector, our multispectral pedestrian detector achieves more true detections, especially when some pedestrians are of bad external illumination. Meanwhile, some false alarms are also removed.

4.2 Evaluation of Proposals:

We also assessed the proposals generated by RPN in the Halfway Fusion, with regard of recalls. RPNs of FasterRCNN-C and FasterRCNN-T were considered in comparison, along with the ACF-C-T pedestrian detector. The comparison results of pedestrian proposals performed on the test set of KAIST are shown in Figure 7.

Recall *vs*. Number of Proposals: Given IoU 0.5, Halfway Fusion model obtains the highest recall with the same number of proposals. This model achieves 94% recall with 50 proposals, compared to other approaches of around 87% recall. In other words, Halfway Fusion could reach the same recall with fewer proposals. This is very useful in practice, since fewer proposals could make DNNs save time in classification. In particular, Halfway Fusion gets 90% recall with 30 proposals, while FasterRCNN-C and FasterRCNN-T require around 80 proposals to achieve competitive recall.

Recall *vs*. IoU: Given 300 proposals from RPN, Halfway Fusion obtains 93.9% recall at IoU 0.6, which is better than other approaches. With 100 proposals, Halfway Fusion model accomplishes comparative recalls against other methods with 300 proposals. Apparently, Halfway Fusion model produces proposals with better overlaps on true detections. In this scenario, we conclude that Halfway Fusion model generates proposals of higher quality.

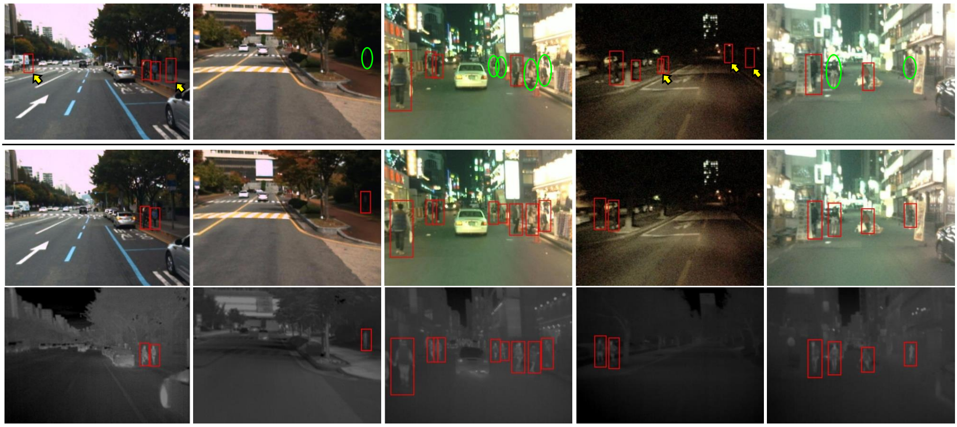

Figure 8: Detection samples. Red bounding boxes denote detections. Yellow arrows indicate false positives and green ellipses represent miss detections. First row: detections by FasterRCNN-C (color images only). Bottom two rows: detections by Halfway Fusion model (multispectral images), illustrated in both color and thermal images.

5 Conclusion

In this paper, we focused on leveraging DNNs for multispectral (color and thermal images) pedestrian detection. Our multispectral detectors were built upon Faster R-CNN detection framework, which archived the state-of-art performance on Caltech pedestrian benchmark. Four ConvNet fusion architectures were proposed, which fused channel features at different ConvNet stages, corresponding to low-level, middle-level, high-level feature fusion, and confidence fusion, respectively. All of them yielded better performance compared with the baseline detector based on Faster R-CNN. Extensive empirical results revealed that our Halfway Fusion model – the fusion of middle-level convolutional features, achieved the best detection synergy and the state-of-the-art performance. It significantly reduced the missing rate of baseline method Faster R-CNN by 11%, and performed a 37% overall missing rate (3.5% lower than other proposed architectures) on KAIST datasets.

References

[1] Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson. Real-time pedestrian detection with deep network cascades. In *BMVC*, 2015.

[2] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc Van Gool. Pedestrian detection at 100 frames per second. In *CVPR*, pages 2903–2910, 2012.

[3] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *ECCVW*, pages 613–627. Springer, 2014.

[4] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*, pages 3361–3369, 2015.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005.

[6] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *CVPR*, pages 304–311, 2009.

[7] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 34(4):743–761, 2012.

[8] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *TPAMI*, 36(8):1532–1545, 2014.

[9] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *T-PAMI*, 32(9):1627–1645, 2010.

[10] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[12] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M López. Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, 16(6):820, 2016.

[13] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, pages 345–360. Springer, 2014.

[14] Ju Han and Bir Bhanu. Human activity recognition in thermal infrared imagery. In *CVPR*, pages 17–17, 2005.

[15] Ju Han and Bir Bhanu. Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 40(6):1771–1784, 2007.

[16] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. In *CVPR*, pages 4073–4082, 2015.

[17] 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.

[18] Yoshua Bengio Ian Goodfellow and Aaron Courville. Deep learning. Book in preparation for MIT Press, 2016. URL http://www.deeplearningbook.org.

[19] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPs*, pages 1097–1105, 2012.

[21] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *arXiv preprint arXiv:1510.08160*, 2015.

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

[23] Jingjing Liu, Quanfu Fan, Sharath Pankanti, and Dimitris N Metaxas. People detection in crowded scenes by context-driven label propagation. In *WACV*, pages 1–9, 2016.

[24] Javier Marin, David Vázquez, Antonio M López, Jaume Amores, and Bastian Leibe. Random forests of local experts for pedestrian detection. In *ICCV*, pages 2592–2599, 2013.

[25] Markus Mathias, Rodrigo Benenson, Radu Timofte, and Luc Gool. Handling occlusions with franken-classifiers. In *ICCV*, pages 1505–1512, 2013.

[26] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. Local decorrelation for improved pedestrian detection. In *NIPs*, pages 424–432, 2014.

[27] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[28] Wanli Ouyang, Xingyu Zeng, and Xiaogang Wang. Modeling mutual visibility relationship in pedestrian detection. In *CVPR*, pages 3222–3229, 2013.

[29] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *ECCV*, pages 546–561. 2014.

[30] 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.

[31] M Saquib Sarfraz and Rainer Stiefelhagen. Deep perceptual mapping for thermal to visible face recognition. *BMVC*, 2015.

[32] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*, pages 3626–3633, 2013.

[33] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPs*, pages 568–576, 2014.

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

[35] Yainuvis Socarrás, Sebastian Ramos, David Vázquez, Antonio M López, and Theo Gevers. Adapting pedestrian detection from synthetic to far infrared images. In *ICCVW*, volume 7, 2011.

[36] Richard Socher, Brody Huval, Bharath Bath, Christopher D Manning, and Andrew Y Ng. Convolutional-recursive deep learning for 3d object classification. In *NIPs*, pages 665–673, 2012.

[37] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPs*, pages 2222–2230, 2012.

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[39] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *ICCV*, pages 1904–1912, 2015.

[40] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks. In *CVPR*, pages 5079–5087, 2015.

[41] Atousa Torabi, Guillaume Massé, and Guillaume-Alexandre Bilodeau. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *CVIU*, 116(2):210–221, 2012.

[42] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. 2016.

[43] Jinqiao Wang, Wei Fu, Jingjing Liu, and Hanqing Lu. Spatiotemporal group context for pedestrian counting. *TCSVT*, 24(9):1620–1630, 2014.

[44] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. *arXiv preprint arXiv:1511.06078*, 2015.

[45] Xiaoyu Wang, Tony X Han, and Shuicheng Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, pages 32–39, 2009.

[46] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, pages 3441–3450, 2015.

[47] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Filtered channel features for pedestrian detection. In *CVPR*, pages 1751–1760, 2015.

[48] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? *arXiv preprint arXiv:1602.01237*, 2016.