Fully Convolutional Region Proposal Networks for Multispectral Person Detection

Daniel König[1,2], Michael Adam[1], Christian Jarvers[2], Georg Layher[2], Heiko Neumann[2], and Michael Teutsch[1]

[1] Hensoldt Optronics GmbH, Oberkochen, Germany
{daniel.koenig, michael.adam, michael.teutsch}@hensoldt.net
[2] Institute of Neural Information Processing, Ulm University, Ulm, Germany
{christian.jarvers, georg.layher, heiko.neumann}@uni-ulm.de

Abstract

Multispectral images that combine visual-optical (VIS) and infrared (IR) image information are a promising source of data for automatic person detection. Especially in automotive or surveillance applications, challenging conditions such as insufficient illumination or large distances between camera and object occur regularly and can affect image quality. This leads to weak image contrast or low object resolution. In order to detect persons under such conditions, we apply deep learning for effectively fusing the VIS and IR information in multispectral images. We present a novel multispectral Region Proposal Network (RPN) that is built up on the pre-trained very deep convolutional network VGG-16. The proposals of this network are further evaluated using a Boosted Decision Trees classifier in order to reduce potential false positive detections. With a log-average miss rate of $29.83\,\%$ *on the reasonable test set of the KAIST Multispectral Pedestrian Detection Benchmark, we improve the current state-of-the-art by about 18 %.*

1. Introduction

The detection of pedestrians and persons in general is a crucial task in many applications that rely on visual perception such as surveillance, autonomous driving, or search and rescue. Despite the fact that significant progress was made in developing computer vision algorithms for automatic person detection in recent years [1, 11, 13, 15, 19], the problem is still far from being solved [41] especially under challenging conditions such as partial occlusions, nighttime without sufficient illumination, or large distance between camera and persons resulting in weak contrast or low resolution. Such conditions, however, are highly relevant for the already mentioned applications and thus it becomes apparent that even after many years of research there is still both potential and need for improvement.

Unlike human eyes, cameras are not limited to the visual-optical (VIS) spectrum. Especially in absence of natural light, infrared (IR) cameras are able to reach higher detection performance compared to VIS [16, 24, 28, 34]. Since we want to avoid active illumination that is necessary for near infrared (NIR) cameras, we focus on video data acquired by purely passive sensors such as VIS and long-wave thermal infrared (LWIR) cameras in the remainder of this paper. By combining VIS and LWIR image information, even more improvement in person detection performance can be achieved [20, 25].

In this paper, we aim at detecting persons in multispectral videos that consist of three VIS channels (RGB) and one thermal IR channel. Inspired by the success of deep learning based proposal generation for object detection in the Faster Region-based Convolutional Neural Network (Faster R-CNN) architecture [30], we propose a novel fusion Region Proposal Network (RPN) that is built up on the pre-trained very deep Convolutional Neural Network (CNN) VGG-16 [32]. Starting with individual CNNs for VIS and IR, we fuse these CNNs *halfway* [27] in order to generate multispectral deep features for the RPN. Figure 1 shows that the multispectral fusion RPN is able to provide more promising proposals compared to single VIS and IR RPNs especially in the challenging conditions mentioned before. In contrast to [27], we omit the classification network of the Faster R-CNN architecture and demonstrate that the RPN alone already outperforms Faster R-CNN for the special task of person detection. This was initially discovered by Zhang *et al.* [39]. They further proposed to use a Boosted Decision Trees (BDT) classifier (in their paper they call it *Boosted Forest*) to verify the proposals achieving state-of-the-art results on several public datasets such as the Caltech

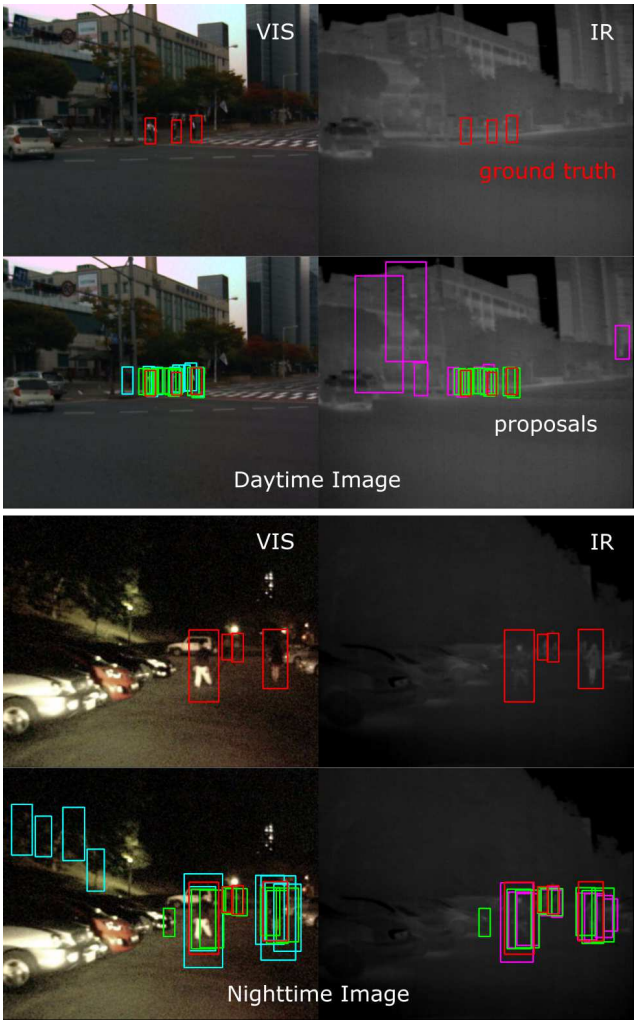

Figure 1. Best 10 proposals generated by a VIS RPN (cyan boxes), an IR RPN (magenta boxes), and the proposed fusion RPN (green boxes). Compared to the ground truth (red boxes), we see less false positives and better localization by the fusion RPN.

Pedestrian Detection Benchmark [13]. We not only confirm their conclusions on the KAIST Multispectral Pedestrian Detection Benchmark [20] but also improve the current state-of-the-art on this dataset by about 18 % using our multispectral fusion RPN and BDT.

This paper is organized as follows: related work is discussed in Section 2, the person detection approach is described in Section 3, experimental results are presented in Section 4, and concluding remarks are given in Section 5.

2. Related Work

Person detection is a combination of localization and classification. In many previous approaches, the localization problem was solved using exhaustive search based methods such as sliding windows at different image scales [7, 11, 12, 38]. At each window position, features are calculated and evaluated using a pre-trained classifier. Among the most popular methods are Histograms of Oriented Gradients (HOG) related features in combination with a Support Vector Machine (SVM) [7, 38] or feature pools consisting of either Haar features [35] or different kinds of channel features [12, 11, 42, 5] evaluated by BDT. In order to accelerate the exhaustive search, proposal generators [18] were introduced that apply fast screening with sliding windows using simple features and objectness measures that are prone to produce many false positives. However, in this way much less proposals have to be evaluated compared to exhaustive search and thus more sophisticated machine learning approaches such as CNNs can be applied. With the introduction of deep learning for proposal generation [29, 30], it was discovered that for the rather simple classification problem in person detection with only two classes (person and non-person), it can be more beneficial to use CNNs to solve the localization problem [39] rather than the classification problem [19].

Since most of the thermal IR datasets are acquired by stationary cameras, background subtraction is a popular approach for generating proposals [2, 6, 8, 9, 14, 37]. Other methods that are applicable with moving cameras are based on keypoint detection [22], sliding windows [28, 40], or hot spot detection using thresholding techniques [34]. The methods that are applied to evaluate these proposals are mainly based on gradient related features and classifiers such as SVM and BDT. Hence, they are comparable to the ones used in the visible spectrum.

An early work in multispectral person detection is given by Leykin *et al*. [25]: on the OSU Color-Thermal Database [9], proposals are generated using background subtraction and evaluated using periodic gait analysis. With the introduction of the KAIST Multispectral Pedestrian Detection Benchmark [20], research on multispectral person detection was revived. In order to detect pedestrians, Aggregated Channel Features (ACF) [11] are used in combination with BDT. Wagner *et al*. [36] apply ACF and BDT just for proposal generation and classify these proposals with a CNN, which fuses the VIS and IR information. Liu *et al*. [27] use the Faster R-CNN architecture and analyze different stages of fusing the VIS and IR information inside the CNN. Finally, Choi *et al*. [4] use separate individual RPNs for both VIS and IR images and evaluate the proposals generated by both networks with Support Vector Regression (SVR) on fused deep features.

3. Multispectral Person Detection

In this section, we first present the architecture of the RPN that is used to fuse the VIS and IR image information and to generate proposals. Then, we describe the BDT classifier and how we extract the deep features that are utilized for person classification.

3.1. Region Proposal Network

Ren *et al*. [30] introduced RPNs within their Faster R-CNN architecture for object detection in general. The idea is to train a fully convolutional network to perform bounding box regression and to determine objectness scores simultaneously. This RPN shares its convolutional layers with a second CNN which is used for classification on the proposed regions. The RPN starts with proposals of different scales at *anchor* locations on a regular grid in the image. Each initial proposal is then regressed to the most likely object position in a limited image region surrounding the anchor location. In this way, a fixed number of proposals is generated that can be ranked using their scores. This approach was adopted by Liu *et al*. [27] and tailored to the task of multispectral person detection: the width to height ratio of the proposed bounding boxes is fixed to 0.5 and the convolutional layers of two separately pre-trained VIS and IR CNNs are connected at a certain layer to generate fused deep features. These features are then evaluated by both the RPN and the classification network of the Faster R-CNN.

Our approach is inspired by the just mentioned Faster R-CNN but deviates in several aspects: (1) we experimentally identify the best convolutional layer to fuse the VIS and IR CNNs, which gives us the most accurate proposals and provides deep features of higher resolution for the BDT classifier, (2) we use different training data, which slightly improves the quality of the proposals, (3) motivated by [39], we do not use a classification network but show that the RPN alone achieves higher detection performance already.

We start with two separate CNNs that are based on the VGG-16 architecture [32]. The fully connected layers are removed and the convolutional layers are initialized with the weights pre-trained on the ImageNet dataset [31]. These two networks are the foundation to individually train a VIS RPN and an IR RPN tailored to person detection. Each RPN is built on top of the conv5_3 layer of the VGG-16 network followed by an intermediate 3×3 convolutional layer and two sibling 1×1 convolutional layers for bounding box regression and classification [30, 39]. Hence, the RPN is a fully convolutional network. While the regression layer provides the positions of the proposals, the classification layer gives us the score. Each input image is resized to 960×768 pixels and 60×48 output feature maps of the conv5_3 layer represent anchor positions for bounding box regression with a 16 pixel stride inside the resized original image. At each anchor position we consider 9 different scales. Hence, each RPN generates 25,920 proposals per image. We perform the fine-tuning (training) of the two individual networks in two stages: the VIS RPN is first fine-tuned using the Caltech training dataset and then the KAIST VIS training dataset. For the IR RPN we first use the CVC-09 dataset [33] and then the KAIST IR training dataset. The two-stage fine-tuning is inspired by [36]. In that work is has been recommend to use the red channel of the Caltech dataset for fine-tuning the IR CNN. Here, we achieved slightly better results using the CVC-09 dataset, which is a real IR dataset. An overview of the number of positive training samples is given in Table 1. We only use *reasonable* training samples, i.e. non-occluded samples with a bounding box height of 50 pixels or larger. The standard Caltech training dataset only contains about 2,000 positive ground truth (GT) labels that are sampled from every 30th training image. However, the performance of CNNs is usually dependent on the volume of the training data. In order to generate a ten times larger training dataset, it is recommended to sample GT labels from every 3rd image instead [19, 39]. Since we want to have a similar number of positive training samples for the KAIST and the CVC-09 dataset, too, we collect them from every 2nd image on both datasets.

Table 1. Datasets and number of GT labels for RPN fine-tuning.

dataset	Caltech [13]	CVC-09 [33]	KAIST [20]
GT labels	16,376	15,058	13,853
images	42,782	8,418	50,172

Having those two separately fine-tuned RPNs available for a fusion architecture, we analyze five options where this fusion can take place, namely after each max pooling sublayer of the conv1, conv2, conv3, conv4, or conv5 layer. An early fusion at the input level by stacking the VIS and IR images and a late fusion at the score level is possible, too, but does not perform promising enough [27, 36] to be considered here. The resulting architectures for each of these five options are shown in Fig. 2. The convolutional layers are depicted in red color and the RPN layers in yellow color. The ticks inside each convolutional layer visualize the number of convolutional sublayers, i.e. conv1 consists of the sublayers conv1_1, relu1_1, conv1_2, relu1_2 and pool1. *conv-prop* represents the intermediate 3×3 convolutional layer and *cls-score* and *bbox-pred* the two sibling layers for classification and bounding box regression. According to [27], the fusion of VIS and IR is done by concatenating the feature maps (blue layer) of the previous convolutional layer. This leads to doubling the number of feature maps. However, since we want to use the pre-trained VGG-16 weights for the convolutional layers after the fusion, we need to reduce this number to the original number of feature maps. Therefore, an additional 1×1 convolutional layer called Network-in-Network (NIN) [26] is introduced (green color) and used for dimension reduction. For the fusion after conv5, the NIN can be omitted since we do not reuse VGG-16 layers here and train the RPN layers from scratch with random initialization. Compared to the fusion approaches described in [27], the architecture in Fig. 2 (a) corresponds to the *Early Fusion*, Fig. 2 (d) corresponds to the *Halfway Fusion*, and Fig. 2 (e) corresponds to the *Late Fusion*.

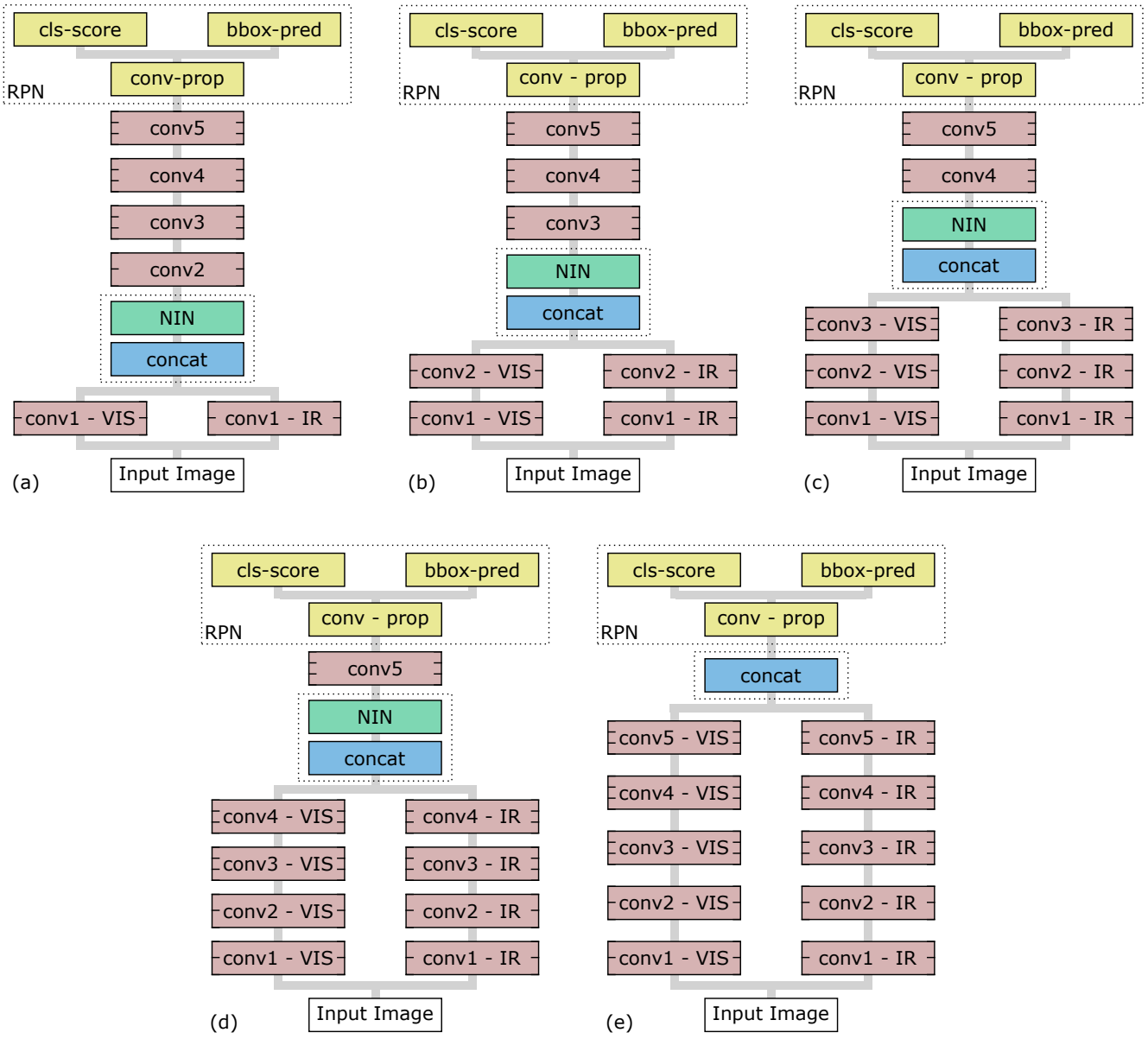

Figure 2. In order to find the best convolutional layer in our RPN for fusing VIS and IR deep features, we identify and analyze five options. The fusion itself uses feature map concatenation and Network-in-Network (NIN) based feature map reduction as proposed in [27]. Please refer to the text for more details.

3.2. Classification using Boosted Decision Trees

For the binary classification problem in person detection, Zhang *et al*. [39] have not only shown that the RPN alone already reaches similar performance compared to the entire Faster R-CNN architecture. They have also shown that reusing the deep features with a smaller classifier model such as BDT can improve detection performance significantly by reducing the number of false positives. This motivates us to apply a similar BDT classifier to our proposals.

Figure 3 demonstrates how we extract our deep features from the proposals (ROIs) provided by the RPN. The RPN architecture here is taken from Fig. 2 (c), which gave us the best results for proposal generation (see Section 4). In this way, we can take features not only from the separate layers conv3 VIS and conv3_3 IR but also from the fused layer conv4_3. This is different compared to [30], where features are picked after the conv5_3 layer. However, Zhang *et al*. [39] experimentally demonstrated that conv3_3 and conv4_3 are the most promising sources of deep features for the BDT. Following [39], we extract the features right before max pooling and apply the à trous trick [3] to layer conv4_3 to generate features of higher density and higher resolution. For each ROI, ROI pooling [17] is ap-

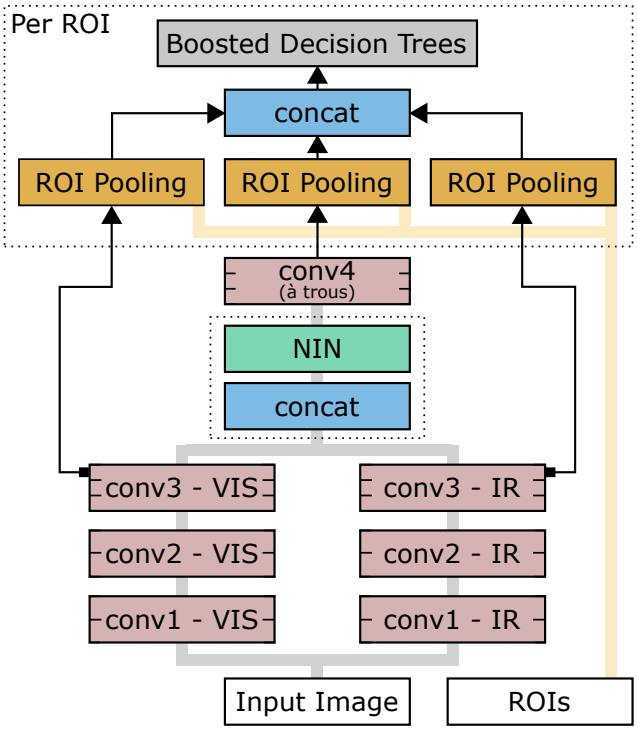

Figure 3. Instead of the Faster R-CNN classification network [30], we use BDT [39]. Deep features are extracted at two different convolutional layers right before max pooling. ROI pooling is used to get a fixed number of features independent of the ROI scale.

plied in order to get a fixed number of features that is independent of the ROI scale. The resolution of features inside each ROI is fixed to 7×7. Our feature pool then consists of 12,544 (conv3 VIS) + 12,544 (conv3 IR) + 25,088 (conv4 Fusion) = 50,176 features compared to 19,350 that are generated by ACF-T-THOG [20]. The training samples for the training of the BDT are taken directly from the 1,000 top-ranked proposals of the RPN per image on the KAIST training dataset. The training is bootstrapped six times with an increasing level of hard negative sample mining. 20,594 positive and between 30,000 and 50,000 negative samples are used. Further information on the BDT classifier itself and its training can be found in [39].

3.3. Implementation Details

Our implementation is based on the MATLAB code provided by Zhang *et al*. [39]. We re-trained their models for RPN and BDT and applied them to the Caltech test dataset achieving similar results compared to the ones presented in their paper. CNN architecture design and training are performed using the Caffe framework [21]. As the VGG-16 network expects three planes per input image (RGB), we simply clone the single plane of the thermal IR images and thus generate three plane IR images for the IR RPN. The anchor aspect ratio between bounding box width and height within the RPN is fixed to 0.41. For proposals with an overlap of 0.7 or more w.r.t. the Intersection over Union (IoU) criterion, only the one with the highest score is kept (non-maximum suppression). The BDT classifier model is trained with the Real AdaBoost algorithm taken from the Piotr Toolbox [10]. More details about the training procedure of RPN and BDT can be found in [23].

4. Experimental Results

Our experiments are conducted using the publicly available KAIST Multispectral Pedestrian Detection Benchmark [20]. As already mentioned, we use every second image of the training images subset to generate a training dataset consisting of 50,172 images with 13,853 annotated persons. The *reasonable* test subset contains 2,252 images (every 20th test image) with 1,356 annotated persons. Furthermore, we use standard evaluation measures such as log-average miss rate, proposals vs. recall, and IoU vs. recall.

In the first experiment, we analyze the influence of the training data to the log-average miss rate of the different fusion RPN architectures shown in Fig. 2 (a-e). We focus on the separate fine-tuning of the VIS RPN and the IR RPN before the actual fusion is performed. As shown in Table 2, three different options are explored: (1) fine-tuning using the KAIST VIS dataset for the VIS RPN and the KAIST IR dataset for the IR RPN [27], (2) using the Caltech dataset and KAIST VIS for the VIS RPN and the Caltech red channel and KAIST IR for the IR RPN [36], and (3) using Caltech and KAIST VIS for the VIS RPN and the CVC-09 dataset and KAIST IR for the IR RPN. The fusion training data only consists of the KAIST multispectral (= VIS + IR) dataset since there is no other multispectral person detection dataset available at this time. As test data, we use the *reasonable* subset of the KAIST test dataset that consists of GT labels that are both not heavily occluded and not smaller than 50 pixels in height. All other labels are declared as *ignore* regions, i.e. any detections in these regions are not considered for the quantitative evaluation at all.

We see that an early fusion after the first and second convolutional layer conv1 and conv2 is not recommendable independent of the training data. For the other three fusion approaches there is no significant difference in performance. The best result, however, is achieved with a fusion after the conv3 layer and by using the CVC-09 dataset for training the IR RPN. Hence, we use this RPN for the follow-up experiments. With a log-average miss rate of 35.50 %, this RPN outperforms the Faster R-CNN approach proposed by Liu *et al*. [27] with a reported log-average miss rate of 36.99 %. This confirms the conclusions of Zhang *et al*. [39] that the classification network does not improve the performance of Faster R-CNN for person detection.

In the second experiment, we evaluate the RPN's region proposal performance by plotting the number of proposals

Table 2. Influence of the training data to the log-average miss rate of the five different fusion RPN architectures (see Fig. 2 (a-e)): The third training data option combined with a fusion after the third convolutional layer conv3 achieves the best result (underlined).

training option	separate training data		fusion training data	fusion after layer				
	VIS RPN	IR RPN		conv1	conv2	conv3	conv4	conv5
(1)	KAIST VIS	KAIST IR	KAIST VIS + IR	45.21 %	40.99 %	36.46 %	36.11 %	36.23 %
(2)	Caltech + KAIST VIS	Caltech R + KAIST IR		47.94 %	39.87 %	36.86 %	35.81 %	36.18 %
(3)	Caltech + KAIST VIS	CVC-09 + KAIST IR		44.71 %	41.49 %	<u>35.50 %</u>	36.42 %	35.52 %

and the IoU against recall. In addition, we compare the fusion RPN with the separately trained pure VIS and pure IR RPN. The results are visualized in Fig. 4 and Fig. 5. The proposed fusion RPN achieves a recall of more than 0.8 for 5 proposals per image and a recall of more than 0.9 for 10 proposals. With 40 proposals per image we can even reach a recall of about 0.98, i.e. we miss only two percent of the annotated true positive (TP) persons. The gap between the fusion RPN and the IR RPN, which performs second best, is about 0.05 recall. A similar conclusion can be made for the second plot: with increasing IoU, the recall of the fusion RPN is consistently higher compared to the VIS and the IR RPN. For an IoU between 0.5 and 0.7, the gap is similar with about 0.05 recall.

The third and last experiment is the comparison with other approaches. We evaluate our proposed fusion RPN and the fusion RPN + BDT with four other methods: (1) the KAIST baseline approach consisting of ACF+T+THOG features in combination with BDT [20], (2) this baseline approach applied as proposal generator in combination with a *LateFusion* CNN for classification [36], (3) the Checkerboards based detector [42] extended for processing multispectral images (Checkerboards+T+THOG), and (4) the *Halfway Fusion* Faster R-CNN approach proposed in [27]. Figure 6 shows the results using the log-average miss rate as evaluation measure. We report two versions of the ACF+T+THOG baseline approach [20]: one with the original parameters achieving 54.40 % log-average miss rate and one with our own optimization reaching 42.57 %. Here, we can see that there is still potential to improve channel features based detectors. The fusion RPN alone already outperforms the other three methods. By applying the BDT classifier in addition, the log-average miss rate can be even further reduced to 29.83 %, which is the best reported result up to now. Liu *et al*. [27] reported 36.99 % in their paper. However, the authors sent us their detections and by using the evaluations scripts of the Piotr toolbox [10], we got the 36.22 % that we added to our plot in Fig. 6. With the proposed method, we are able to improve the current state-of-the-art by about 18 %.

However, the log-average miss rate can probably be further reduced by improving the annotations. In Fig. 7, we show the qualitative evaluation. GT bounding boxes are depicted in red color and detection bounding boxes in green color. Ignore regions are visualized with orange rectangles. Red arrows point at image regions, in which the annotations are either imprecisely located or missing. Especially in Fig. 7 (a) and (e), there are several missing annotations. In image (e), our approach is able to detect some of the not annotated persons. This causes additional false positive (FP) detections in the quantitative evaluation. If such imprecise or missing annotations appear in the training data, this could even lead to an RPN or BDT classifier model of weaker discriminative power. In the past, similar observations were made for the Caltech dataset and the annotations were improved [41]. In this way, we do not want to diminish the authors' great work behind the KAIST dataset. Instead, we want to encourage prospective authors to improve the annotations in the future.

5. Conclusions

Inspired by recent literature [39], we proposed a fully convolutional fusion RPN in combination with a BDT classifier for person detection in multispectral video data. This data consists of the three VIS channels RGB and an additional thermal IR channel. Within the fusion RPN architecture, we experimentally identified the best convolutional layer after which to fuse the multispectral image information. The fusion RPN alone already outperforms the current state-of-the-art w.r.t. log-average miss rate on the KAIST Multispectral Pedestrian Detection Benchmark. Furthermore, the fusion RPN is able to achieve a recall larger than 0.9 with only 10 proposals per image. Even more improvement in detection performance is achieved with the additional application of the BDT classifier pushing the log-average miss rate to 29.83 %. To the best of our knowledge this is currently the best reported result on this benchmark.

References

[1] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV Workshops*, 2014. 1

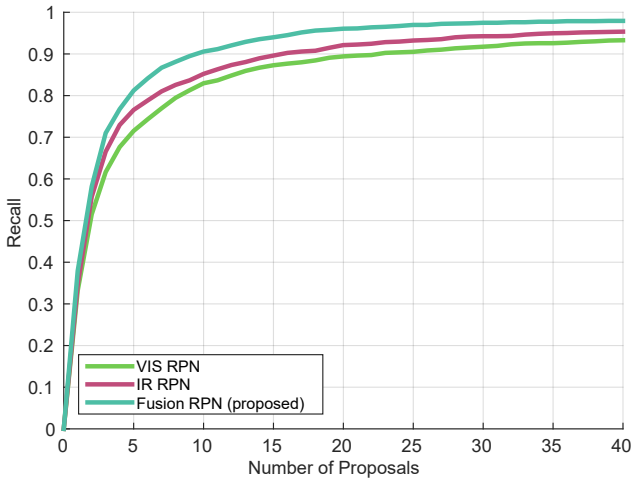

Figure 4. By plotting the number of proposals per image against recall, we see that the proposed fusion RPN clearly outperforms both the VIS RPN and the IR RPN.

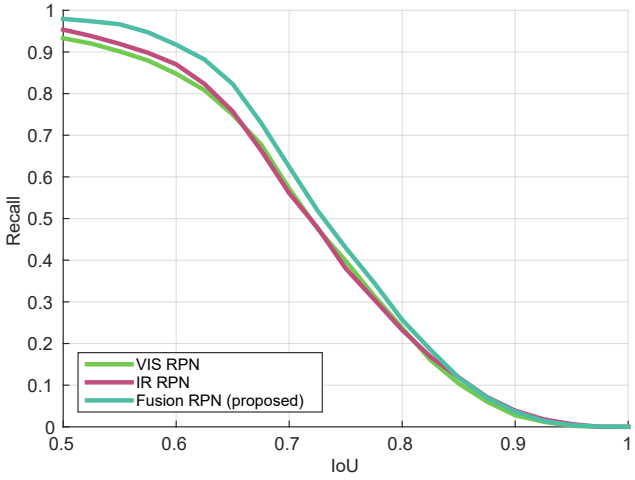

Figure 5. With increasing Intersection over Union (IoU), the recall of our proposed fusion RPN is consistently at least 0.05 higher compared to the VIS RPN and the IR RPN.

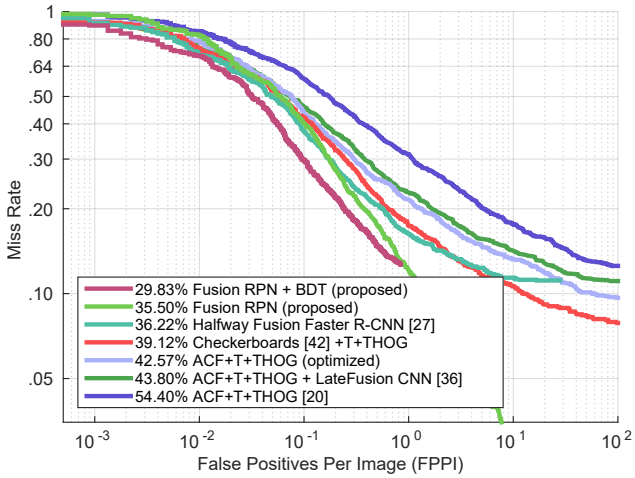

Figure 6. On the KAIST reasonable test dataset, our proposed method achieves a log-average miss rate of 29.83 % and outperforms the current state-of-the-art by about 18 %.

[2] B. Chen, W. Wang, and Q. Qin. Robust multi-stage approach for the detection of moving target from infrared imagery. *SPIE Optical Engineering*, 51(6), June 2012. 2

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *ICLR*, 2015. 4

[4] H. Choi, S. Kim, K. Park, and K. Sohn. Multi-spectral Pedestrian Detection Based on Accumulated Object Proposal with Fully Convolution Network. In *ICPR*, 2016. 2

[5] A. D. Costea and S. Nedevschi. Semantic Channels for Fast Pedestrian Detection. In *CVPR*, 2016. 2

[6] C. Dai, Y. Zheng, and X. Li. Layered Representation for Pedestrian Detection and Tracking in Infrared Imagery. In *CVPR Workshops*, 2005. 2

[7] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005. 2

[8] J. W. Davis and M. A. Keck. A two-stage approach to person detection in thermal imagery. In *WACV*, 2005. 2

[9] J. W. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *CVIU*, 106, 2007. 2

[10] P. Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). https://github.com/pdollar/toolbox. 5, 6

[11] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast Feature Pyramids for Object Detection. *TPAMI*, 36(8), 2014. 1, 2

[12] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral Channel Features. In *BMVC*, 2009. 2

[13] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian Detection: An Evaluation of the State of the Art. *TPAMI*, 34(4), 2012. 1, 2, 3

[14] T. Elguebaly and N. Bouguila. A Nonparametric Bayesian Approach for Enhanced Pedestrian Detection and Foreground Segmentation. In *CVPR Workshops*, 2011. 2

[15] M. Enzweiler and Dariu M. Gavrila. Monocular pedestrian detection: Survey and experiments. *TPAMI*, 31(12), 2009. 1

[16] J. Ge, Y. Luo, and G. Tei. Real-Time Pedestrian Detection and Tracking at Nighttime for Driver-Assistance Systems. *IEEE Transactions on ITS*, 10(2), 2009. 1

[17] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 4

[18] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? *TPAMI*, 2015. 2

[19] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *CVPR*, 2015. 1, 2, 3

[20] S. Hwang, J. Park, N. Kim, Y. Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, 2015. 1, 2, 3, 5, 6

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5

[22] K. Jüngling and M. Arens. Feature based person detection beyond the visible spectrum. In *CVPR Workshops*, 2009. 2

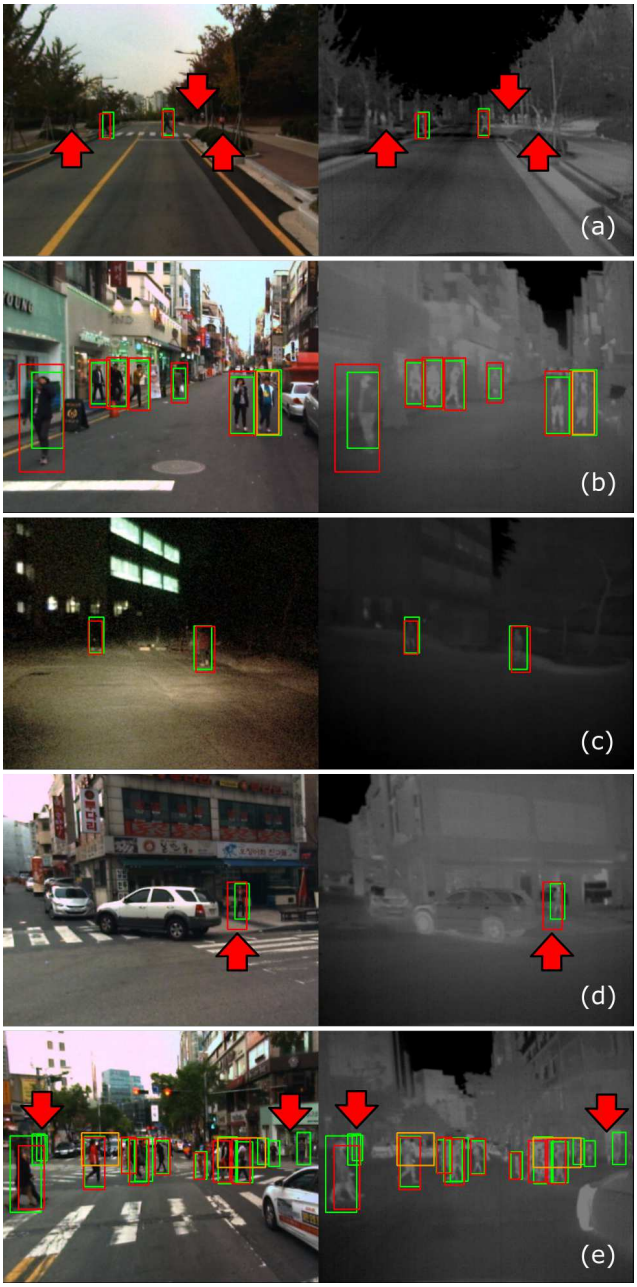

Figure 7. Examples taken from the KAIST Multispectral Pedestrian Detection Benchmark. GT bounding boxes are depicted in red, detections in green, and ignore regions in orange color. Red arrows indicate missing or imprecise GT annotations.

[23] D. König. Deep Learning for Person Detection in Multi-Spectral Videos. Master's thesis, Ulm University, Germany, 2017. 5

[24] Y. Lee, Y. Chan, L. Fu, and P. Hsiao. Near-Infrared-Based Nighttime Pedestrian Detection Using Grouped Part Models. *IEEE Transactions on ITS*, 16(4), 2015. 1

[25] A. Leykin, Y. Ran, and R. Hammoud. Thermal-visible video fusion for moving target tracking and pedestrian classification. In *CVPR*, 2007. 1, 2

[26] M. Lin, Q. Chen, and S. Yan. Network In Network. In *ICLR*, 2013. 3

[27] J. Liu, S. Zhang, S. Wang, and Dimitris N. Metaxas. Multispectral Deep Neural Networks for Pedestrian Detection. In *BMVC*, 2016. 1, 2, 3, 4, 5, 6

[28] R. Miezianko and D. Pokrajac. People Detection in Low Resolution Infrared Videos. In *CVPR Workshops*, 2008. 1, 2

[29] P. Pinheiro, R. Collobert, and P. Dollár. Learning to Segment Object Candidates. In *NIPS*, 2015. 2

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015. 1, 2, 3, 4, 5

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3), 2015. 3

[32] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015. 1, 3

[33] Y. Socarras, S. Ramos, D. Vazquez, A. Lopez, and T. Gevers. Adapting Pedestrian Detection from Synthetic to Far Infrared Images. In *ICCV Workshops*, 2013. 3

[34] M. Teutsch, T. Müller, M. Huber, and J. Beyerer. Low resolution person detection with a moving thermal infrared camera by hot spot classification. In *CVPR Workshops*, 2014. 1, 2

[35] P. Viola and M. Jones. Robust Real-time Face Detection. *IJCV*, 57(2), 2004. 2

[36] J. Wagner, V. Fischer, M. Herman, and S. Behnke. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2016. 2, 3, 5, 6

[37] J. Wang, G. Bebis, and R. Miller. Robust Video-Based Surveillance by Integrating Target Detection with Tracking. In *CVPR Workshops*, 2006. 2

[38] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li. Robust Multi-resolution Pedestrian Detection in Traffic Scenes. In *CVPR*, 2013. 2

[39] L. Zhang, L. Lin, X. Liang, and K. He. Is Faster R-CNN Doing Well for Pedestrian Detection? In *ECCV*, 2016. 1, 2, 3, 4, 5, 6

[40] L. Zhang, B. Wu, and R. Nevatia. Pedestrian Detection in Infrared Images based on Local Shape Features. In *CVPR Workshops*, 2007. 2

[41] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How Far are We from Solving Pedestrian Detection? In *CVPR*, 2016. 1, 6

[42] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *CVPR*, 2015. 2, 6