Domain Adaptation for Privacy-Preserving Pedestrian Detection in Thermal Imagery

My Kieu, Andrew D. Bagdanov(✉), Marco Bertini, and Alberto Del Bimbo

Media Integration and Communication Center, University of Florence,
50134 Florence, FI, Italy
{my.kieu,andrew.bagdanov,marco.bertini,alberto.delbimbo}@unifi.it

Abstract. Pedestrian detection is a core problem in computer vision, and is a problem that is gaining prominence due to its importance in assisted and autonomous driving applications. Many state-of-the-art approaches, especially those used for autonomous driving, combine thermal and visible spectrum imagery in order to robustly detect persons independent of time of day or weather conditions. In this paper we investigate two domain adaptation techniques for fine-tuning a YOLOv3 detector to perform accurate and robust pedestrian detection using *thermal* images. Our approaches are motivated by the fact that thermal imagery is *privacy-preserving* in the sense that person identification is difficult or impossible. Results on the KAIST dataset show that our approaches perform comparably to state-of-the-art approaches and outperform the state-of-the-art on nighttime pedestrian detection, even outperforming multimodal techniques that use both thermal and visible spectrum imagery at test time.

Keywords: Pedestrian detection · Thermal imaging · Domain adaptation · Privacy-preservation

1 Introduction

Object detection is a classical problem in computer vision, and person and pedestrian detection is one of the most important topics for safety and security applications such as video surveillance, autonomous driving, person re-identification, and numerous others. The estimate of the total number of installed video surveillance cameras range was already at 240 million worldwide in 2014 [16]. The advent of autonomous driving promises to add many more cameras, all detecting and observing humans in public spaces.

Recent works on pedestrian detection have investigated the use of thermal imaging sensors as a complementary technology for visible spectrum images [21]. Approaches such as these aim to combine thermal and RGB image information in order to obtain the most robust possible pedestrian and person detection and

E. Ricci et al. (Eds.): ICIAP 2019, LNCS 11752, pp. 203–213, 2019.
https://doi.org/10.1007/978-3-030-30645-8_19

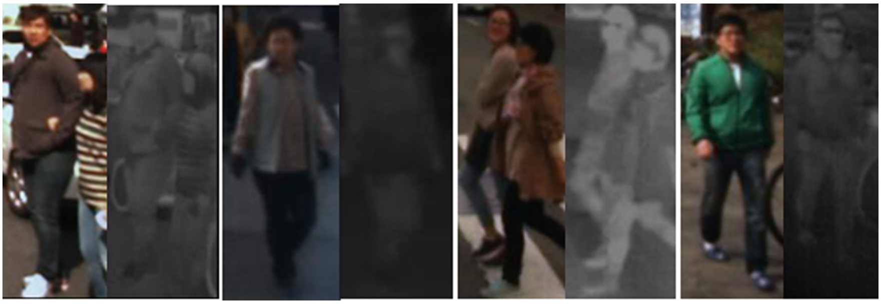


Fig. 1. Thermal imaging and privacy preservation. Shown are three cropped images from the KAIST dataset. On the left of each is the RGB image, to the right the crop from the corresponding thermal image. Note how persons are readily identifiable in visible spectrum images, but not in corresponding thermal images. Although identity is concealed, there is still enough information in thermal imagery for detection. (Color figure online)

any time of the day or night. Such detectors require both visible spectrum and thermal images to function.

Citizens are naturally concerned that being observed violates their right to privacy. In this paper we are interested in investigating the limits of pedestrian detection using thermal imagery alone. Figure 1 gives an example of four matched pairs of color and thermal images from the KAIST dataset [10]. From these examples we see that, even in relatively low resolution color images, persons can be readily identified. Meanwhile, thermal images retain distinctive image features for detection while *preserving privacy*. Our hypothesis is that thermal images can guarantee the balance between security and privacy concerns.

The rest of this paper is organized as follows. In the next section, we briefly review related work from the computer vision literature on domain adaptation, thermal imaging, and pedestrian detection. In Sect. 3 we describe several approaches to domain adaptation that we apply to the problem of privacy-preserving person detection. We report on a range of experiments conducted in Sect. 4, and conclude in Sect. 5 with a discussion of our contribution and future research directions.

2 Related Work

In this section we review some recent work related to pedestrian detection, domain adaptation, and computer vision for thermal imagery.

Person and Pedestrian Detection. The literature, both classical and contemporary, on pedestrian detection is vast [3]. With the advent of deep neural networks in recent years, pedestrian detection is achieving higher and higher accuracy! [1]. However, pedestrian detection remains a challenging task due to occlusion, changing illumination and variation of viewpoint and background [17].

Several CNN-based pedestrian detection methods compete for the state-of-the-art on standard benchmark datasets for pedestrian detection. Examples include Pedestrian Detection aided by Deep Learning Semantic Tasks [24], Scale-Aware Fast RCNN [14], Learning Mutual Visibility Relationship [17]. These state-of-the-art techniques use RGB images as input, while our goal is to investigate the potential of detection in thermal imagery alone.

Domain Adaptation. Domain adaptation has played a main role in both supervised and unsupervised recognition in computer vision. Domain adaptation attempts to exploit learned knowledge from the source domain in the target domain. One of our approaches was inspired by the AdapterNet [8], which proposed adding a new shallow Convolutional Neural Network (CNN) before the original model that transforms the input image the target domain before passing through an unmodified network trained in the source domain. Several works have tried to mitigate the distance between the two domains by applying transformation techniques. For example, the idea from [9] was to transform infrared data (thermal domain) as close as possible to the color domain by using feature transformations: inversion, equalization and histogram stretching. A deep architecture, called Invertible Autoencoder (InvAuto), introduced a method to treat an encoder as an inverted version of a decoder in order to decrease the trainable parameters of image translation processing [20].

Pedestrian Detection Exploiting Thermal Imagery. Several works demonstrate that using thermal images in combination with RGB images can improve object detection results. An example is the work in [23], which suggests a method based on a cross-modality learning framework focusing only on visible images at test time. During training time, they use thermal image features to boost visible detection results. Their method has two main phases: Region Reconstruction Network (RRN), for learning a non-linear feature mapping between visible and thermal image pairs, and a Multi-Scale Detection Network (MDN) which performs pedestrian detection from visible images by exploiting the cross-modal representations learned with RRN.

A variety of recent works leverage two-stage network architectures to investigate the combination of visible and thermal features. In [22] the authors investigated two types of fusion networks. Another approach is the ACF+T+HOG technique [15] which considers four different network fusion approaches (early, halfway, late, and score fusion). The authors of [11] introduced a combination Fully Convolutional Region Proposal Networks (RPN) and Boosted Decision Trees Classifier (BDT) for person detection in multispectral video. Illumination-aware Faster R-CNN (IAF RCNN) [13] and Illuminating Pedestrians via Simultaneous Detection and Segmentation [4] used the Faster R-CNN detector to perform pedestrian detection on paired RGB and thermal imagery. A Fusion architecture network (MSDS-RCNN) including a multispectral proposal network (MPN) and a multispectral classification network (MCN) was proposed by [5]. This fusion network currently yields the best results on both visible and thermal image pairs on the KAIST dataset.

In a slightly different direction, the combination of HOG and SVM in [2] focused on only nighttime detection. Their method uses a Thermal Position Intensity Histogram of oriented gradient (TPIHOG) and the additive kernel SVM (AKSVM) for training and testing.

Differing from most of the above works which used two-stage detectors, some the papers utilize a one-stage detector [12,21]. The authors of [12] used a deconvolutional single shot multi-box detector (DSSD) to exploit correlation between visible and thermal features for person detection. A fast RGB single-pass network architecture (YOLOv2 [18]) was adopted by [21] for fine-tuning for person detection.

3 Domain Adaptation Approaches

In this section we describe the approaches to domain adaptation that we will later evaluate in Sect. 4. All of our approaches use the YOLOv3 detector which is adapted to a target domain through a sequence of domain adaptation steps. We use YOLOv3 pretrained on the ImageNet and subsequently fine-tuned on the MS COCO Person class 3.

3.1 Top-Down Domain Adaptation

We use the term *top-down domain adaptation* to refer to the fine-tuning approach to domain adaptation in which the network is fine-tuned in the new domain to adapt weights to the new input distribution. Thus it is top-down in the sense that adaptation happens only via backpropagation from the detection loss at the end of the network. We investigate three different top-down approaches. In the descriptions below we use a notational convention to refer to each technique that indicates which image modalities are used for training and testing. For example, the technique indicated as TD(VT, T) is Top-Down domain adaptation, with adaptation on Visible spectrum images, followed by adaptation on Thermal images, and finally tested on Thermal images.

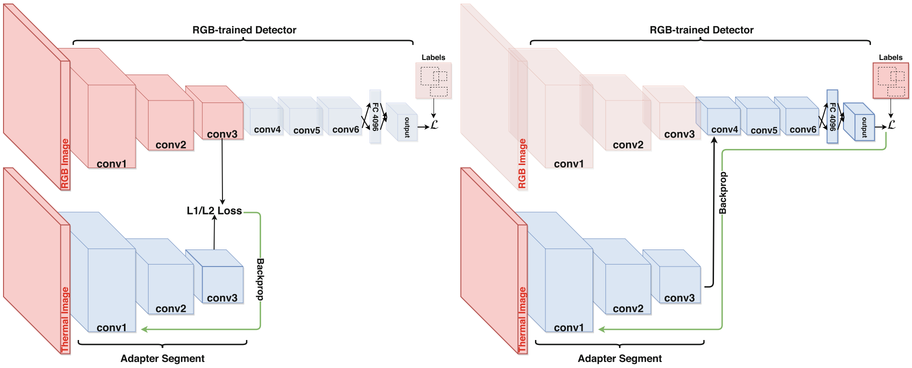
Top-Down Visible: TD(V, V). This domain adaptation approach directly fine-tunes YOLOv3 on visible images in the target domain (pedestrians in the KAIST dataset for all experiments). Testing is performed on visible spectrum images. This baseline adaptation approach serves as a sort of upper bound for performance achievable during daytime (since visible spectrum images should contain most information).

Top-Down Thermal: TD(T, T). This approach directly fine-tunes YOLOv3 on thermal images by duplicating the thermal image three times, once for each input channel of the RGB-trained detector. Testing is performed only on thermal imagery. This baseline adaptation method serves as a sort of upper bound for the performance achievable at nighttime (since thermal images should convey most information).

Top-Down Visible/Thermal: TD(VT, T). This approach is a variant of the two top-down approaches described above. First we adapt YOLOv3 to the visible spectrum pedestrian detection domain, then we fine-tune that detector on thermal imagery. Testing is performed on thermal images. The idea here is to determine if knowledge from the visible spectrum can be retained and exploited after final adaptation to the thermal domain.

3.2 Bottom-Up Domain Adaptation: BU(VAT, T)

A hypothesis of ours is that in top-down domain adaptation, as described in the previous section, early convolutional layers are difficult and slow to adapt to the new input distribution due to their distance from the backpropagated loss. Here we propose a type of *bottom-up* domain adaptation which first trains a bottom-up adapter segment and then proceeds to fine-tune the detector using a top-down loss. A conceptual schema of this approach is given in Fig. 2. The main components of our bottom-up domain adaptation approach are as follows.



(a) Adapter segment training (b) Fine-tuning of entire detector

Fig. 2. Bottom-up domain adaptation. (a) An *adapter segment* is first trained to take thermal images as input, then matches the feature activations of a RGB-pretrained detector on the corresponding RGB image. (b) After training the adapter segment, the RGB input branch is discarded and the entire detector pipeline is fine-tuned on thermal images.

Adapter Segment Training. As illustrated in Fig. 2(a), the main idea of the Adapter Segment is to intervene at some early stage of the RGB-trained detector network and to train a parallel branch that takes only thermal imagery as input and *matches* as best as possible the RGB feature maps at the point of intervention. In our implementation, we decapitate the YOLOv3 network after the first ten convolutional layers and train a ten-layer adapter segment to match the RGB-network using only thermal images as input. We use a simple L2 loss function on the output feature maps from the truncated network and adapter segment.

The starting point for this approach is the TD(V, V) network described above. That is, the detector weights we start from are already adapted to the KAIST domain on visible images. We then train the adapter segment using RGB/thermal image pairs from the KAIST training set. This is the “A” in the “VAT” for training in the mnemonic for this approach: BU(VAT, T).

Final Adaptation. After adapter segment training has converged, we reconnect the newly trained adapter segment to the original network for final fine-tuning of the entire detector on thermal images Fig. 2(b).

4 Experimental Results

In this section we report results of experiments we performed to evaluate the performance of adapted detectors for pedestrian detection in thermal imagery.

4.1 Experimental Setup

To evaluate our proposed approaches to domain adaptation we used a standard benchmark dataset of RGB/thermal image pairs and standard evaluation protocols.

Dataset and Evaluation Metrics. All experiments were performed on the publicly available KAIST Multispectral Pedestrian Detection Benchmark [10], which consists of 95,328 color-thermal pairs images. The KAIST dataset contains 103,128 annotations of 1,182 unique pedestrians with. The originally proposed splits had 50,328 training and 45,000 test images. According to the official sampling method from the some recent papers [10,11,21], we also do a 2-frame sample on the training set and 20-frame sample on the test set. The training set used contains 19,058 RGB/thermal pairs after sampling filtering (e.g occlusion, the bounding box under 50 pixels), and the test set consists of 2,252 images (after 20-frame sample).

To evaluate the performance of our detection results, we use log-average miss rate (miss rate) and precision/recall metrics, which almost all pedestrian detectors use and is described in [6]. The evaluation protocol we followed is the same as reported in [21], which is an updated version of the Matlab code from [6].

Implementation Details. We used the YOLOv3 [19] detector to evaluate our approach on KAIST. Our detectors were implemented using PyTorch, and we trained every domain adaptation strategy for 50 epochs with a learning rate 0.0001 and the Adam optimizer.

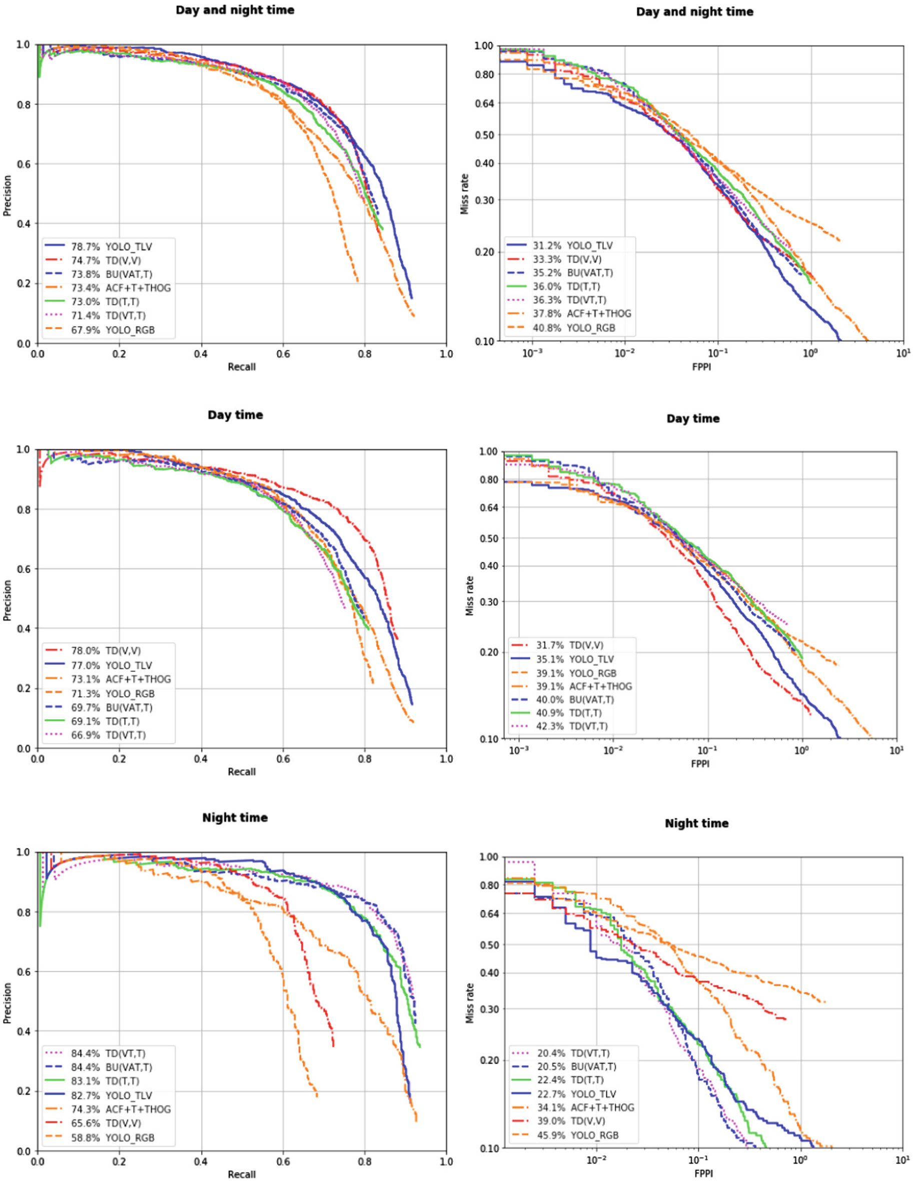


Fig. 3. Comparative performance analysis. Precision/Recall (left, higher is better) and Log-average Miss Rate (right, lower is better) of our method and other state-of-the-art papers are given. See text for detailed analysis.

4.2 Comparative Performance Analysis

The plots in Fig. 3 show detailed results for our approach and those described in [21] in terms of precision/recall (left column) and log-average miss rate (right column). The plots also break down results in terms of time-of-day: first row averaged over all times, second row daytime only, third row nighttime only.

From the results in Fig. 3 we can make several observations. First of all, for combined day and night results (first row) multimodal techniques like YOLO_TLV which exploit both thermal and visible spectrum images at test time are superior to our domain adaptation approaches which use only thermal imagery. Surprisingly, however, the gap between bottom-up domain adaptation BU(VAT, V) and YOLO_TLV is only about 4% in log-average miss rate, which is quite promising.

The reason that multimodal approaches outperform domain adaptation seems to be due to the advantage they have when detecting during the day. In the second row of Fig. 3, in fact, we see that the technique exploiting visible spectrum images during at test time on daytime images outperform all our approaches which only use thermal imagery.

Or two domain adaptation approaches, both top-down and bottom-up, outperform all other techniques when testing at nighttime only (third row of Fig. 3). Though this is not very surprising, of particular note is the fact that performing domain adaptation on to *visible* images before adapting to thermal input only is beneficial. This can be seen in the difference between TD(VT, T), BU(VAT, T) – both of which start by fine-tuning YOLOv3 on KAIST visible images – and TD(T, T), which directly fine-tunes YOLOv3 on thermal images. This seems to indicate that both top-down and bottom-up domain adaptation are able to retain and exploit some domain knowledge acquired when training the detector on visible spectrum imagery.

As a final comment, we note that the BU(VAT, T) approach requires significantly less training time that the others. In only 15 epochs it converged to 84.4% precision, which is the same result for top-down adaptation after 50 epochs. Bottom-up adaptation seems to be an effective way to accelerate top-down adaptation through fine-tuning.

In Table 1 we provide a comparison of our methods and 10 others methods from the state-of-the-art. Our approaches outperform all other single modality techniques (both visible- and thermal-only). Compared to multi-model approaches, we outperform all of them at nighttime, and comparably on all.

Table 1. Log-average miss rate on KAIST dataset (lower is better). The final two columns indicate which image modality is used at *test time*. Our approaches outperform all single-modality techniques from the literature, and outperform all methods at night.

Method	MR all (%)	MR day (%)	MR night (%)	Visible	Thermal
KAIST baseline [10]	64.76	64.17	63.99	✓	✓
Late fusion [22]	43.80	46.15	37.00	✓	✓
Halfway fusion [15]	36.99	36.84	35.49	✓	✓
RPN+BDT [11]	29.83	30.51	27.62	✓	✓
IATDNN+IAMSS [7]	**26.37**	**27.29**	24.41	✓	✓
YOLO_TLV [21]	31.20	35.10	22.70	✓	✓
DSSD-HC [12]	34.32	–	–	✓	✓
RRN+MDN [23]	49.55	47.3	54.78	✓	
TPIHOG [2]	–	–	57.38		✓
SSD300 [9]	69.81	–	–		✓
Ours: TD(V, V)	33.30	31.70	39.00	✓	
Ours: TD(T, T)	36.00	40.90	22.40		✓
Ours: TD(VT, T)	36.30	42.30	**20.40**		✓
Ours: BU(VAT, T)	35.20	40.00	20.50		✓

4.3 Qualitative Evaluation

In Fig. 4 we show some example detection results on the KAIST dataset for our BU(VAT, T) domain adaptation approach in daytime (first row) and nighttime (second row). Note how, even though person identification is impossible in all of the example images, the detector adapted using bottom-up domain adaptation is able to detect pedestrians even in the presence of occlusion, scale variation, and changing illumination conditions.

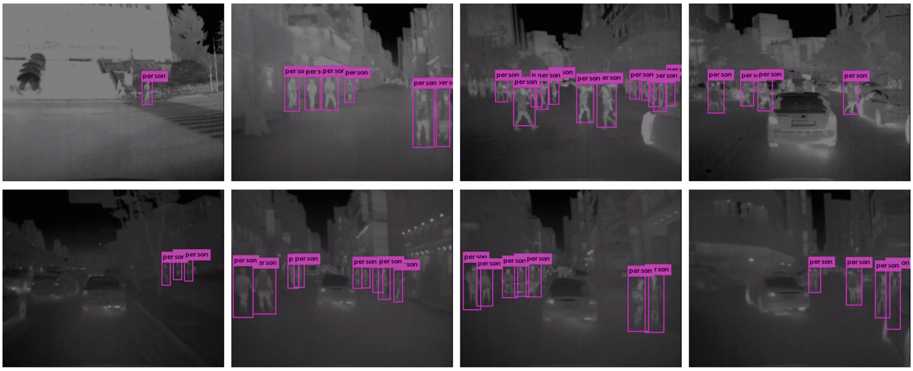


Fig. 4. Qualitative results on the KAIST test set. The first row gives example detections on daytime images from KAIST, and second row on nighttime images. Even in the presence of occlusions and scale variations, thermal imagery retains enough information to effectively perform pedestrian detection – day or night – in a privacy-preserving way without using any visible spectrum imagery at detection time.

5 Conclusions and Future Work

In this paper we investigated the potential of two domain adaptation strategies for adapting pedestrian detectors to work in the thermal domain. The goal of this work is to achieve the best possible person detection performance while relying *solely* on thermal spectrum imagery. This is motivated by the *privacy-preserving* aspects of thermal images, since persons are difficult, if not impossible, to reliably identify in thermal images.

Our results indicate that relatively simple domain adaptation schemes can be effective, and that the resulting detectors can outperform multimodal approaches (i.e. those that use thermal *and* visible images at test time) at nighttime, and can perform comparably when testing on day night images combined. Moreover, results seem to indicate that a first adaptation to visible imagery can be useful to acquire domain knowledge that can then be exploited after final adaptation to thermal spectrum images.

Ongoing work is concentrated on improving daytime and overall performance of adapted detectors. We are investigating techniques to retain more information from visible spectrum adaptation in order to close the gap between privacy-preserving detection in thermal imagery and multimodal techniques which require visible spectrum images.

References

1. Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A., Ferguson, D.: Real-time pedestrian detection with deep network cascades. In: BMVC (2015)
2. Baek, J., Hong, S., Kim, J., Kim, E.: Efficient pedestrian detection at nighttime using a thermal camera. Sensors **17**, 1850 (2017)
3. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8926, pp. 613–627. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16181-5_47
4. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection and segmentation. In: ICCV, pp. 4960–4969 (2017)
5. Li, C., Song, D., Tong, R., Tang, M.: Multispectral pedestrian detection via simultaneous detection and segmentation. In: British Machine Vision Conference (BMVC), September 2018
6. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 743–761 (2012)
7. Guan, D., Cao, Y., Yang, J., Cao, Y., Yang, M.: Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Inf. Fusion **50**, 148–157 (2018)
8. Hazan, A., Shoshan, Y., Khapun, D., Aladjem, R., Ratner, V.: AdapterNet - learning input transformation for domain adaptation. CoRR (2018)
9. Herrmann, C., Ruf, M., Beyerer, J.: CNN-based thermal infrared person detection by domain adaptation. In: Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything (2018)
10. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baselines. In: CVPR (2015)

11. Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M.: Fully convolutional region proposal networks for multispectral person detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (2017)
12. Lee, Y., Bui, T.D., Shin, J.: Pedestrian detection based on deep fusion network using feature correlation. In: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 694–699 (2018)
13. Li, C., Song, D., Tong, R., Tang, M.: Illumination-aware faster R-CNN for robust multispectral pedestrian detection. CoRR (2018)
14. Li, J., Liang, X., Shen, S., Xu, T., Yan, S.: Scale-aware fast R-CNN for pedestrian detection. CoRR (2015). http://arxiv.org/abs/1510.08160
15. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. CoRR (2016)
16. Markit, I.: 245 million video surveillance cameras installed globally in 2014. Web page (2019). https://technology.ihs.com/532501/245-million-video-surveillance-cameras-installed-globally-in-2014. Accessed 5 May 2019
17. Ouyang, W., Zeng, X., Wang, X.: Learning mutual visibility relationship for pedestrian detection with a deep model. Int. J. Comput. Vis. **120**, 14–27 (2016)
18. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
19. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. CoRR (2018). http://arxiv.org/abs/1804.02767
20. Teng, Y., Choromanska, A., Bojarski, M.: Invertible autoencoder for domain adaptation. CoRR (2018)
21. Vandersteegen, M., Van Beeck, K., Goedemé, T.: Real-time multispectral pedestrian detection with a single-pass deep neural network. In: Campilho, A., Karray, F., ter Haar Romeny, B. (eds.) ICIAR 2018. LNCS, vol. 10882, pp. 419–426. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93000-8_47
22. Wagner, J., Fischer, V., Herman, M., Behnke, S.: Multispectral pedestrian detection using deep fusion convolutional neural networks. In: 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 509–514 (2016)
23. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5363–5371 (2017)
24. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5079–5087 (2015)