Unsupervised Domain Adaptation for Multispectral Pedestrian Detection

Dayan Guan[1]   Xing Luo[1]   Yanpeng Cao[1]   Jiangxin Yang[1]   Yanlong Cao[1]

George Vosselman[2]   Michael Ying Yang[2]

[1]Zhejiang University, {11725001, luoxing, caopy, yangjx, sdcaoyl}@zju.edu.cn

[2]University of Twente, {george.vosselman, michael.yang}@utwente.nl

Abstract

Multimodal information (e.g., visible and thermal) can generate robust pedestrian detections to facilitate around-the-clock computer vision applications, such as autonomous driving and video surveillance. However, it still remains a crucial challenge to train a reliable detector working well in different multispectral pedestrian datasets without manual annotations. In this paper, we propose a novel unsupervised domain adaptation framework for multispectral pedestrian detection, by iteratively generating pseudo annotations and updating the parameters of our designed multispectral pedestrian detector on target domain. Pseudo annotations are generated using the detector trained on source domain, and then updated by fixing the parameters of detector and minimizing the cross entropy loss without back-propagation. Training labels are generated using the pseudo annotations by considering the characteristics of similarity and complementarity between well-aligned visible and infrared image pairs. The parameters of detector are updated using the generated labels by minimizing our defined multi-detection loss function with back-propagation. The optimal parameters of detector can be obtained after iteratively updating the pseudo annotations and parameters. Experimental results show that our proposed unsupervised multimodal domain adaptation method achieves significantly higher detection performance than the approach without domain adaptation, and is competitive with the supervised multispectral pedestrian detectors.

1. Introduction

Pedestrian detection has become an important and popular topic within the field of computer vision community over the past few years [6, 1, 24, 32, 34, 33]. Given sensing images captured in complex and changing real-world environment, pedestrian detection solution is required to predict the regions of human. It provides significant information for various human-centric sensing applications. In order to facilitate the around-the-clock robotic applications, such as
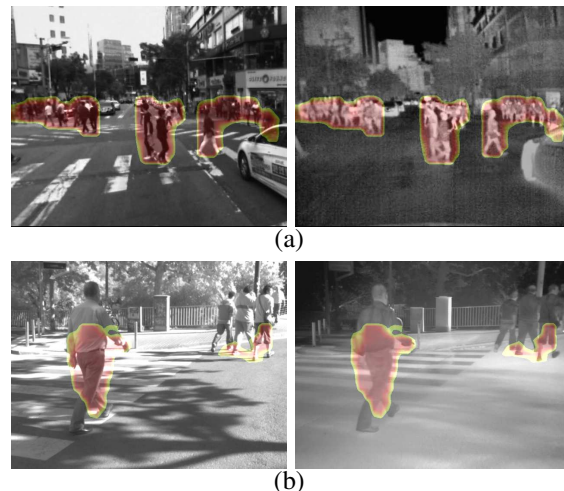

(a)

(b)

Figure 1. Detection results of the current state-of-the-art multispectral pedestrian detector well-trained using visible and thermal image pairs from KAIST [11] dataset following the method presented by Cao *et al.* [4]. (a) Results on the KAIST dataset; (b) Results on the CVC-14 [8] dataset. Please note that the visible images in KAIST dataset are transferred from RGB to gray level in order to decrease domain differences between these two datasets.

autonomous driving and video surveillance, multimodal information (e.g., visible and thermal) have applied to generate more robust and reliable pedestrian detection results in the recent years [15, 11, 13, 29, 14, 10].

Although significant improvements have been accomplished in the research area of multispectral pedestrian detection recently, it still remains a crucial challenge to train a reliable multispectral pedestrian detector working well in different open benchmark datasets simultaneously. The detection performance of a multispectral pedestrian detector well-trained on one benchmark dataset may drop significantly when applying to another one. Specifically, we utilize the KAIST [11] and CVC-14 [8] benchmark datasets to display this phenomenon. Considering that the visible images from CVC-14 dataset are gray scale without much color information, we transfer the visible images in KAIST dataset from RGB to gray level in order to decrease do-
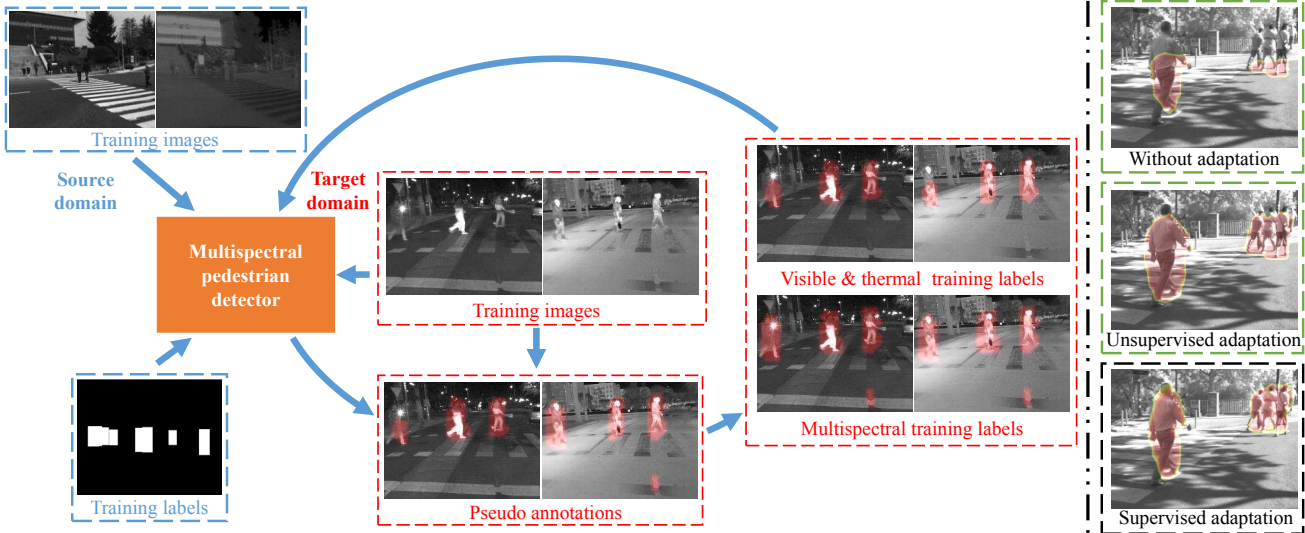
Figure 2. Illustration of our proposed unsupervised domain adaptation framework for multispectral pedestrian detection. Pseudo annotations are generated using the detector trained on source domain, and then updated by fixing the parameters of detector and minimizing the cross entropy loss without back-propagation. Training labels are generated using the pseudo annotations by considering the characteristics of similarity and complementarity between well-aligned visible and infrared image pairs. The parameters of detector are updated using the generated labels by minimizing our defined multi-detection loss function with back-propagation. The optimal parameters of detector can be obtained by iteratively generating pseudo annotations and updating the parameters of our designed multispectral pedestrian detector on target domain.

main differences between these two datasets. For example, as shown in Fig. 1, the current state-of-the-art multispectral pedestrian detector [4] well-trained in the KAIST [11] dataset can't generate reliable detection results on the images from CVC-14 [8] dataset. This is because multispectral pedestrian detectors well-trained on major open dataset tend to overfit the training data, which is usually biased to specific environments [36]. Different benchmark datasets exist domain differences caused by varying conditions of viewpoints, cameras, weather and etc.

To improve multispectral pedestrian detection performance on target domain, multiple-cue information should be generated to update the detector on target data. A nature idea is to annotate data on the target domain. However, densely annotating images is costly and unscalable, since the target domain might change frequently. To overcome this limitation, we propose a novel unsupervised domain adaptation framework for multispectral pedestrian detection, by iteratively generating pseudo annotations and updating the parameters of our designed multispectral pedestrian detector on target domain, as shown in Fig. 2. Pseudo annotations are generated using the detector trained on source domain, and then updated by fixing the parameters of detector and minimizing the cross entropy loss without back-propagation. Training labels are generated using the pseudo annotations by considering the characteristics of similarity and complementarity between well-aligned vis-

ible and infrared image pairs. The parameters of detector are updated using the generated labels by minimizing our defined multi-detection loss function with back-propagation. We show in the experimental part that our designed unsupervised multimodal domain adaptation method achieves significantly higher detection performance than the approach without domain adaptation, which can verify the effectiveness of our proposed approach. Comparing with the supervised multimodal domain adaptation method which need extremely time-consuming manual annotating effort, our proposed unsupervised method barely increase the training time because the additional processing time is caused by the optimization of pseudo annotations without back-propagation.

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

1 we demonstrate the usefulness of visible and thermal data for the task of unsupervised domain adaptation for multispectral pedestrian detection. Characteristics of similarity and complementarity between well-aligned visible and infrared image pairs can be used to adapt the detector trained on an annotated source domain to a target one without manual annotations. To the best of our knowledge, this is the first attempt to explore characteristics of visible and infrared images on the task of unsupervised domain adaptation for multispectral pedestrian detection.

2 We propose a novel unsupervised domain adaptation framework for multispectral pedestrian detection, by iteratively generating training labels and updating the parameters of our designed multispectral pedestrian detector on target domain. Training labels are generated using the pseudo annotations, which are updated by fixing the parameters of detector and minimizing the cross entropy loss without back-propagation. The parameters of detector are updated using the generated labels by minimizing our defined multi-detection loss function with back-propagation.

3 Our proposed unsupervised multimodal domain adaptation method achieves significantly higher detection performance than the approach without domain adaptation, and is competitive with the supervised multispectral pedestrian detectors. Comparing with the supervised approach which need extremely time-consuming manual annotating effort, our proposed unsupervised method barely increase the training time.

2. Related works

Pedestrian detection applications in intelligent robotics, urban surveillance, and self-driving vehicles have been widely spread. Constantly emerging pedestrian detectors and related improvements have accelerated its practical application. Zhang *et al.* [30] adopted a hybrid strategy that extracting the candidate regions utilizing region proposal networks [21] along with boosted classifiers [25]. Mao *et al.* [19] proposed a powerful framework which implements representations of channel features to benefit the detection by additionally learning extra features to assist inference. Brazil *et al.* [2] put detection and segmentation together during the training period, where it suggests that weak box-level annotations could bring benefit to the improvement of detection accuracy. Wang *et al.* [27] designed a novel repulsion loss to restrain the predicted boxes from shifting to surrounding ground truth boxes. The superior visible pedestrian detection performance had achieved with the detectors trained using the repulsion loss.

With the complementary informations given by infrared images, multispectral pedestrian detection expands the research field beyond the traditional visible images and turns to be a potential solution to shrink the gap between machine and human observers. Hwang *et al.* [11] noticed the phenomenon and released the first large-scale multispectral pedestrian dataset (KAIST), containing well-aligned visible and infrared image pairs annotated densely. Liu *et al.* [13] methodically explored the performance of two-stream deep convolutional neural networks where the multi-information feature integrates, showing the architecture that merges two-branch features on the middle-level convolutional layers outperforms any other ones. König *et al.* [14]

adopted the architecture of RPN+BDT [30] in a fusion way, which merges the features generated by two-branch middle-level convolutional layers, in the purpose of multispectral pedestrian detection. Researchers also paid attention to the main difference between visible and infrared images, and proposed illumination-aware weighting mechanism to give extra information to detectors [10, 17]. Guan *et al.* [9] presented a unified multispectral fusion framework, which infuses the multispectral semantic segmentation masks as supervision for learning human-related features, getting more accurate detection results. Li *et al.* [16] designed a cascaded multispectral classification network to distinguish hard negatives sample from pedestrian and human-like instances. Cao *et al.* [4] developed a novel box-level segmentation supervised networks, which can generate more accurate multispectral pedestrian detections on small-size training images. Experimental results showed that their proposed approach achieved the current state-of-the-art pedestrian detection performance using visible and thermal images on both accuracy and speed.

Although significant improvements have been accomplished in the research area of pedestrian detection recently, it still remains a crucial challenge to train a reliable pedestrian detector working well in different open benchmark datasets simultaneously. In the past few years, some researchers have developed different unsupervised domain adaptation schemes in order to avoid the annotation effort. Wang *et al.* [26] presented a new method to achieve unsupervised domain adaptation for a scene-specific pedestrian detector. The approach explores multiple context cues (e.g., structures, locations and sizes) in the static video surveillance to select high-confident training sample on target domain. Liu *et al.* [18] proposed an effect algorithm to iteratively select negative annotations on source domain and annotate positive labels with high score on target domain as the training samples on the task of unsupervised domain adaptation for pedestrian detection in surveillance situations. Wu *et al.* [28] designed a selective ensemble algorithm to adapt the human detector based on Haar-like features [31] and boosted classifier [25] to target domain. The selective ensemble algorithm recombined the useful components that are capable of generating human-related characteristics related to target domain. Cao *et al.* [3] developed a novel unsupervised domain adaptation method to adapt a visible pedestrian detector on source domain to a multispectral pedestrian detector on target domain without using any annotations. An auto-annotation framework was designed to iteratively annotate pedestrian labels.

As far as we know, it has not been solved yet to adapt a multispectral pedestrian detector on source domain to target one without manual annotations. Thus, we propose an unsupervised domain adaptation framework for multispectral pedestrian detection in this paper.
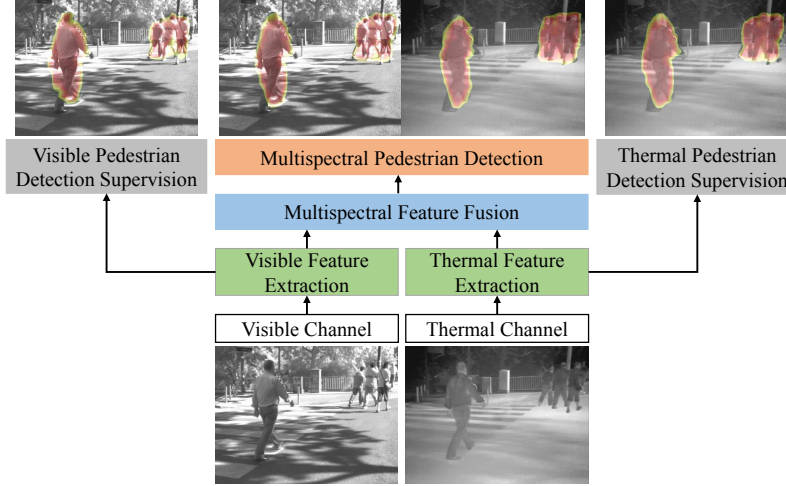
Figure 3. Architecture of our proposed detector for joint training of multispectral pedestrian detections associated with visible and thermal detection supervisions. It contains four major components: feature extraction, feature fusion, pedestrian detection, and detection supervision. The feature extraction module learn features from visible and thermal channels individually; the feature fusion module integrate the visible and thermal features to generate multimodal feature maps; the pedestrian detection module learn the generated multimodal features to produce multispectral pedestrian detections; the detection supervision module learn the individual features to generate visible and thermal pedestrian detections, which provide additional feature information to facilitate the training of multispectral pedestrian detector.

3. Our approach

We first design a new multispectral pedestrian detector and train it on source domain. Based on our designed detector, a novel unsupervised adaptation framework for multispectral pedestrian detection is proposed by iteratively generating training labels and updating the parameters of detector on target domain. Visible and thermal pseudo annotations are generated using our designed detector trained on source domain, and then updated by fixing the parameters of detector and minimizing the cross entropy loss without back-propagation. Training labels in visible, thermal and multispectral channels are generated using the pseudo annotations by considering the characteristics of similarity and complementarity respectively, which are existing in well-aligned visible and infrared image pairs. The parameters of detector are updated using the generated labels by minimizing our defined multi-detection loss function with back-propagation.

3.1. Multispectral pedestrian detector

Inspired by the multi-task framework for joint training of multispectral pedestrian detection and semantic segmentation [9], we combine the visible and thermal pedestrian detection supervision module with the box-level segmentation supervised deep neural networks [3] to build multispectral pedestrian detector, as illustrated in Fig. 3. During the training procedure, the bounding box annotations are used to generate box-level segmentation mask as training labels for the detector on multispectral, visible and thermal chan-

nels simultaneously.

Let $\{x, \overline{y}\}$ denote the training images x with box-level segmentation masks $\overline{y} = \{\overline{y}_i, i = 1, ..., I\}$ (I pixels), where $\overline{y}_i = 1$ represents the foreground pixel and $\overline{y}_i = 0$ denotes the background one. At each iteration step, the parameters θ is updated by minimizing a multi-detection loss function, which is defined as:

$$\theta^{k+1} = \arg\min_{\theta^k}(\mathcal{L}_c(y_M, \overline{y}, I \,|\, x; \theta^k) \\ + \mathcal{L}_c(y_V, \overline{y}, I \,|\, x; \theta^k) + \mathcal{L}_c(y_T, \overline{y}, I \,|\, x; \theta^k)); \tag{1}$$

where y_M, y_V and y_T represent the prediction of pedestrian regions on multispectral, visible and thermal channels; I represents the set of training pixels on the box-level segmentation masks; and $\mathcal{L}_c(y, \overline{y}, I \,|\, x; \theta)$ is the cross entropy loss for classification which is defined as:

$$\mathcal{L}_c(y, \overline{y}, I \,|\, x; \theta) = -\sum_{i \in I}(\overline{y}_i \log(y_i) + (1 - \overline{y}_i)\log(1 - y_i)), \tag{2}$$

where $y_i \in [0, 1]$ represents the confident score which predicts the probability of the corresponding pixel belonging to pedestrian regions, $\overline{y}_i = 1$ presents the foreground pixel and $\overline{y}_i = 0$ denotes the background pixel.

The optimal parameters of detector θ^* can be obtained after iteratively updating the parameters θ. During the testing phase, the output of our designed detector is multispectral pedestrian detections, which are form of full-size heat map predictions.
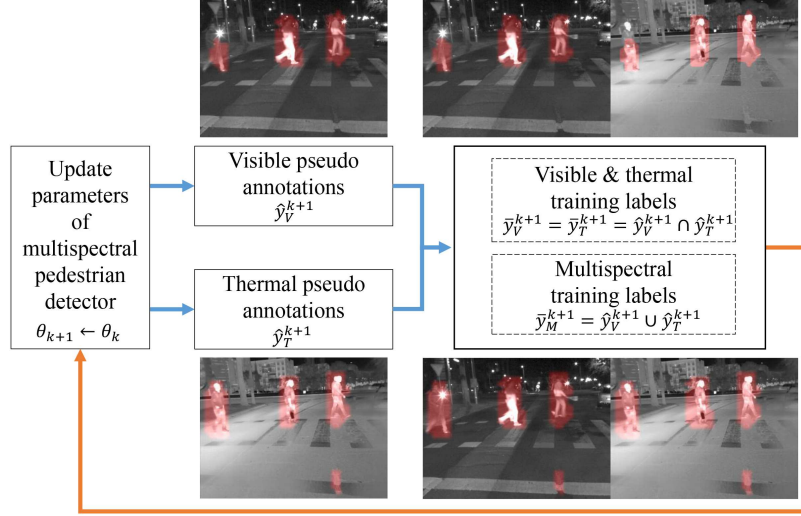
Figure 4. Framework of our proposed unsupervised multimodal domain adaptation model. At each iteration step, visible and thermal pseudo annotations are updated by fixing the parameters of detector and minimizing the cross entropy loss without back-propagation. Training labels in visible, thermal and multispectral channels are generated using the pseudo annotations by considering the characteristics of similarity and complementarity respectively, which are existing in well-aligned visible and infrared image pairs. The parameters of detector are updated using the generated labels by minimizing our defined multi-detection loss function with back-propagation.

3.2. Unsupervised multimodal domain adaptation

Based on our designed multispectral pedestrian detector, we propose an unsupervised domain adaptation framework for multispectral pedestrian detection by iteratively generating training labels and updating the parameters of detector on target domain, as illustrated in Fig. 4. Firstly, the visible and thermal pseudo annotations $\{\hat{y}_V^0, \hat{y}_T^0\}$ are initialized using the visible and thermal pedestrian detection supervision module of our designed detector, which has been trained on source domain. At each iteration step k, the most confident pseudo labels can be selected by fixing the parameters of detector θ^k and minimizing the cross entropy loss function $\mathcal{L}_c$. The visible and thermal pseudo annotations $\{\hat{y}_V^{k+1}, \hat{y}_T^{k+1}\}$ are generated by adding the most confident pseudo labels into the existing set $\{\hat{y}_V^k, \hat{y}_T^k\}$.

Thus, the optimization of visible pseudo annotations $\hat{y}_V^{k+1} \in \{0, 1\}$ is defined as:

$$\hat{y}_V^{k+1} = \arg\min_{\hat{y}_V}(\mathcal{L}_c(y_V, \hat{y}_V, I \,|\, x; \theta^k)) \cup \hat{y}_V^k, \quad (3)$$

and the thermal pseudo annotations $\hat{y}_T^{k+1} \in \{0, 1\}$ is optimized as:

$$\hat{y}_T^{k+1} = \arg\min_{\hat{y}_T}(\mathcal{L}_c(y_T, \hat{y}_T, I \,|\, x; \theta^k)) \cup \hat{y}_T^k. \quad (4)$$

The optimized visible and thermal pseudo annotations are used to generate training labels on visible, thermal and multispectral channels, by considering the characteristics of similarity and complementarity between well-aligned visible and thermal image pairs. The similarity means that there

exists obvious human-related features on both visible and thermal channels, which can be used as a cue to train the visible and thermal pedestrian detection supervision module simultaneously. We consider the intersection of visible and thermal pseudo annotations as the regions that exist obvious human-related features on both channels. Thus, the visible and thermal training labels are generated as:

$$\overline{y}_V^{k+1} = \overline{y}_T^{k+1} = \hat{y}_V^{k+1} \cap \hat{y}_T^{k+1}. \quad (5)$$

Considering that the pseudo annotations should not be considered as negative training labels, we define the set of visible training pixels as:

$$I_V = I - \hat{y}_V^{k+1} + \overline{y}_V^{k+1}, \quad (6)$$

and the set of thermal training pixels as:

$$I_T = I - \hat{y}_V^{k+1} + \overline{y}_T^{k+1}. \quad (7)$$

The complementarity means that visible and thermal data can provide complementary information about objects of interest to improve the detection accuracy. We consider the union of visible and thermal pseudo annotations as the complementary information to update the parameters of multispectral pedestrian detector. Thus, the multispectral training labels are generated as:

$$\overline{y}_M^{k+1} = \hat{y}_V^{k+1} \cup \hat{y}_T^{k+1}. \quad (8)$$

The parameters of detector are updated using the generated training labels by minimizing a multi-detection loss

function with back-propagation, which is defined as:

$$\theta^{k+1} = \arg\min_{\theta^k}(\mathcal{L}_c(y_M, \overline{y}_M^{k+1}, I \mid x; \theta^k)$$
$$+ \mathcal{L}_c(y_V, \overline{y}_V^{k+1}, I_V \mid x; \theta^k) \qquad (9)$$
$$+ \mathcal{L}_c(y_T, \overline{y}_T^{k+1}, I_T \mid x; \theta^k)).$$

The optimal parameters of detector θ^* can be obtained after iteratively updating the visible and thermal pseudo annotations $\{\hat{y}_V, \hat{y}_T\}$ and the parameters θ.

4. Experiments

4.1. Datasets

In order to conduct our experiments on multimodal domain adaptation, we utilize the KAIST [11] and CVC-14 [8] multispectral pedestrian benchmarks as the source and target domain datasets respectively.

The KAIST training dataset contains 50172 well-aligned color visible and thermal infrared sequential image pairs with 13853 dense pedestrian annotations. The images on KAIST dataset were captured in various traffic environments with a resolution of 640×512. Considering that the visible images from CVC-14 dataset are gray scale without much color information, we transfer the visible images in KAIST dataset from RGB to gray level in order to decrease domain differences between these two datasets. According to the current state-of-the-art multispectral pedestrian detector designed by Cao *et al.* [4], the training images are downscaled to the resolution of 320×256 through bilinear interpolation and the bounding box annotations are transfered to box-level segmentation masks to train the detectors on source domain.

The CVC-14 training dataset consists of 7085 aligned gray visible and thermal infrared sequential image pairs with 8105 dense pedestrian annotations. It should be noted that the manual annotations on the CVC-14 training dataset are abandoned in our designed unsupervised multimodal domain adaptation method. The CVC-14 testing dataset contains 1433 aligned image pairs in which 706 pairs were captured during daytime and others in nighttime. All the images on CVC-14 dataset were captured in city traffic environments with a resolution of 640×480. For a fair comparison with the current state-of-the-art multispectral pedestrian detector designed by Cao *et al.* [4], we downscale the images to the resolution of 320×240 through bilinear interpolation during training and testing phase on target domain. The annotations of CVC-14 test set under the reasonable setting (pedestrians larger than 50 pixels [8]) are used to evaluate detection performance.

4.2. Implementation Details

All the detectors are trained and tested using the Caffe [12] deep learning framework with the image-centric strategy to generate mini-batches. We set the batch size to one. For the supervised multimodal domain adaptation, the box-level segmentation masks are generated as training labels using the bounding box annotations, following the multispectral pedestrian detection method designed by Cao *et al.* [4]. Each stream in multispectral deep neural networks is initialized using the parameters in VGG-16 [23] pre-trained on the ImageNet dataset [22] and the other convolutional layers are initialized according to Xavier initialization following [7]. The multispectral pedestrian detector on source domain is trained with stochastic gradient descent (SGD) algorithm [35] for the first 2 epochs with learning rate (LR) 0.001 and 1 more epoch with LR 0.0001 following [4]. The multispectral pedestrian detector on target domain is finetuned with SGD algorithm for 4 epochs with a low LR of 0.00005. In order to avoid gradient exploding, we utilize the adjustable gradient clipping method [20] in the training procedure to suppress exploding gradients.

4.3. Evaluation Metric

The final output of our approach contains human regions and background regions classified by the confident scores according to our frameworks prediction in a heat map style, following the method presented by Cao *et al.* [4]. Considering the difference of results depicted in heat map and traditional bounding box style, to compare impartially, we turn detection results in bounding box into heat map representation depending on the prediction scores. As its utilized diffusely, we also adopt the average precision (AP) [5, 4] in pixel-level as metric to quantify the comparison result between our method and others. More specifically, the average precision (AP) refers to 4 concepts including true positive (TP), true negative (TN), false positive (FP), false negative (FN). Provided with the human-target labels, some bounding box in this case, we call the pixels in bounding box containing human-target foreground pixels, while those surrounding ones is treated as background pixels. After we get the final heat map, true positive (TP) counts those pixels belonging to human-targets inferred correctly, true negative (TN) counts those pixels that is not belonging to human-targets but gets inferred, false positive (FP) counts those background pixels inferred incorrectly, false negative (FN) counts those background pixels inferred correctly. The precision is the ratio TP / (TP + FP), while the recall is the ratio TP / (TP + FN). The AP depicts the shape of the precision/recall curve, and is defined as the mean precision at each recalls by varying the threshold on detection scores.

4.4. Evaluation of UMDA

In order to verify the effectiveness of our proposed approach, we evaluate the detection performance of our proposed unsupervised multimodal domain adaptation (UMDA) model with the detection model without multi-

modal domain adaptation (WMDA). The quantitative performance (pixel-wise AP [4]) of UMDA and WMDA are compared in Tab. 1. It is observed that our proposed unsupervised multimodal domain adaptation method achieves multispectral pedestrian detection performance significantly higher than the approach without domain adaptation, pixel-level AP [4] of UMDA is 31.37% higher than the results of WMDA.

Table 1. Comparing the quantitative performance (pixel-wise AP [4]) of UMDA and WMDA.

Model	All-day	Daytime	Nighttime
WMDA	0.4886	0.3986	0.5584
UMDA	**0.8023**	**0.7688**	**0.8503**

In addition, the qualitative performance of multispectral pedestrian detection results of UMDA and WMDA are compared in Fig. 5. We can observe that the UMDA can generate accurate detections in the case that the performance of WMDA is not satisfactory, as shown in Fig. 5 (a). Even in the situation when WMDA can't detect any pedestrian regions, the UMDA is capable of generating accurate detections.

The quantitative and qualitative comparison of UMDA and WMDA can prove that our proposed unsupervised multimodal domain adaptation framework is able to improve the multispectral pedestrian detection performance on target domain with a large margin.

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

We define the supervised multimodal domain adaptation (SMDA) model as our proposed multispectral pedestrian detector trained on target domain with manual annotations. The proposed SMDA and unsupervised multimodal domain adaptation (UMDA) models are compared with the current state-of-the-art multispectral pedestrian detectors, such as ACF+T+THOG [11], Fusion RPN+BDT [14], and HMFFN320 [4] on target domain. Considering that these detectors were trained on the KAIST multispectral pedestrian benchmark, we fine-tune these detectors on the CVC-14 multispectral pedestrian dataset with supervised multimodal domain adaptation method.

The quantitative performance (pixel-wise AP [4]) of different multispectral pedestrian detectors are compared in Tab. 2. It is observed that our proposed SMDA model outperforms the current state-of-the-art supervised multispectral pedestrian detectors, pixel-level AP [4] of SMDA is 0.91% higher than the results of HMFFN320 [4] and 5.18% higher than the ones of Fusion RPN+BDT [14]. Considering that our proposed SMDA model incorporates the visible and thermal pedestrian detection supervision module into the HMFFN320 [4] model, we can prove that the two-stream detection supervision module is able to provide ad-
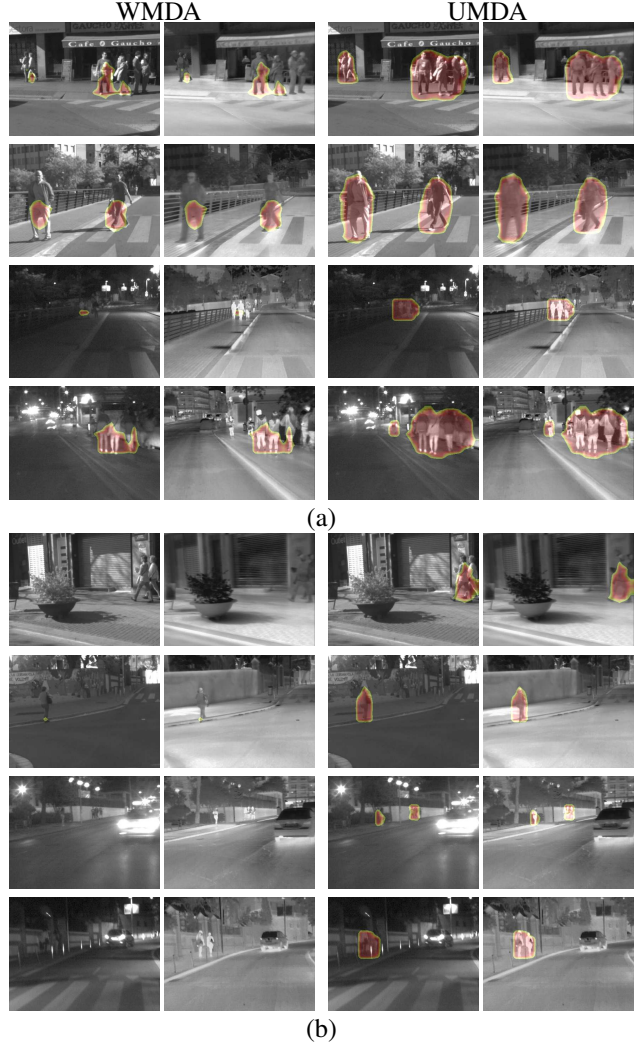

Figure 5. Comparing the qualitative performance of multispectral pedestrian detection results of UMDA and WMDA. (a) The performance of WMDA is not satisfactory while the UMDA can generate accurate detections; (b) WMDA can't detect any pedestrian regions while the UMDA is capable of generating accurate detections.

ditional feature information to facilitate the training of multispectral pedestrian detectors.

In Tab. 2, we observe that our proposed unsupervised multimodal domain adaptation (UMDA) model achieves competitive detection accuracy comparing with the supervised multispectral pedestrian detectors, pixel-level AP [4] of UMDA is 6.67% lower than the results of SMDA and 8.84% higher than the ones of ACF+T+THOG [11]. In order to investigate the gap between unsupervised and supervised multimodal domain adaptation models, we also compare the qualitative performance of multispectral pedestrian detection results of UMDA and SMDA in Fig. 6. When human-related characteristics are distinct in either visible or

Table 2. Comparing the quantitative performance (pixel-wise AP [4]) of UMDA and SMDA with the current state-of-the-art methods. Please note that ACF+T+THOG [11], Fusion RPN+BDT [14], HMFFN320 [4], and SMDA are supervised multispectral pedestrian detectors; UMDA is unsupervised domain adaptation model for multispectral pedestrian detection.

Model	All-day	Daytime	Nighttime
ACF+T+THOG [11]	0.7139	0.6926	0.7334
Fusion RPN+BDT [14]	0.8172	0.8103	0.8241
HMFFN320 [4]	0.8599	0.8355	0.8942
SMDA (ours)	**0.8690**	**0.8485**	**0.8944**
UMDA (ours)	**0.8023**	**0.7688**	**0.8503**

thermal image as illustrated in Fig. 6 (a), the multispectral pedestrian detection results of UMDA are comparable with the ones of SMDA. However, the UMDA may generate unsatisfactory results comparing with the SMDA when either the pedestrian samples appear indistinct or the background is clutter. Our future research will focus on enhancing the multispectral pedestrian detection methods to separate the human-related features with background ones.

It should be mentioned that manual annotating of large-scale multispectral pedestrian dataset is extremely time-consuming. As mentioned in [3], it takes more than 80 hours to annotate the visible and infrared image pairs on the KAIST training dataset, which contains 50172 aligned visible and infrared sequential image pairs. Considering that the CVC-14 training dataset consist of 7085 aligned multispectral sequential image pairs, we consider that the annotating time is more than 11 hours. In comparison, our proposed unsupervised multimodal domain adaptation (UMAD) framework can be used to train the multispectral pedestrian detector without manual annotating effort. It is worth mentioning that comparing with the training procedure of supervised multispectral pedestrian detection approach, the additional processing time of our proposed UMAD method is the optimization of visible and thermal pseudo annotations without back-propagation, which barely increases the training time.

5. Conclusion

In this paper, we present an unsupervised multimodal domain adaptation (UMAD) framework for multispectral pedestrian detection, by iteratively generating pseudo annotations and updating the parameters of our designed multispectral pedestrian detector on target domain without manual annotating effort. Our proposed UMAD method achieves multispectral pedestrian detection performance significantly higher than the approach without multimodal domain adaptation (pixel-level AP [4] of UMDA is 31.37% higher than the results of WMDA) , and is competitive with the supervised multispectral pedestrian detectors (pixel-level AP [4] of UMDA is 6.67% lower than
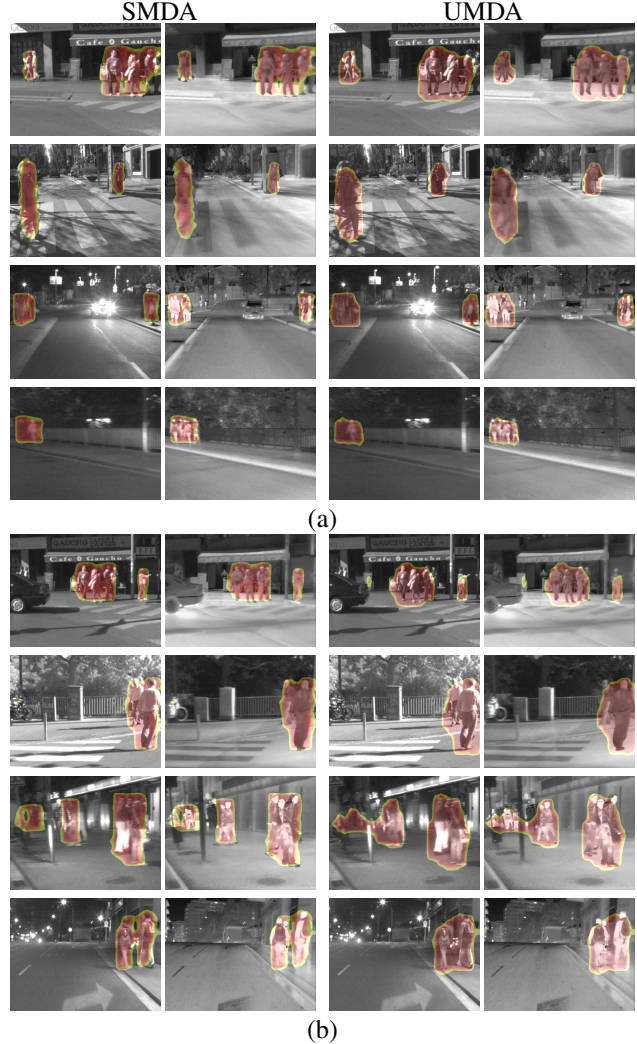

(a)

(b)

Figure 6. Comparing the qualitative performance of multispectral pedestrian detections of UMDA and SMDA. (a) The results of UMDA is comparable with SMDA; (b) The results of UMDA is not satisfactory comparing with SMDA.

the results of SMDA and 8.84% higher than the ones of ACF+T+THOG [11]). It is worth mentioning that the training time of our proposed UMAD framework is barely the same as the training time of supervised approach. Our proposed method can be adapted to other multimodal computer vision tasks on unsupervised domain adaptation without manual annotating effort.

Acknowledgment

References

[1] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2903–2910. IEEE, 2012. 1

[2] G. Brazil, X. Yin, and X. Liu. Illuminating pedestrians via simultaneous detection and segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4960–4969. IEEE, 2017. 3

[3] Y. Cao, D. Guan, W. Huang, J. Yang, Y. Cao, and Y. Qiao. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. *Information Fusion*, 46:206–217, 2019. 3, 4, 8

[4] Y. Cao, D. Guan, Y. Wu, J. Yang, Y. Cao, and M. Y. Yang. Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150:70–79, 2019. 1, 2, 3, 6, 7, 8

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990. 6

[6] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311. IEEE, 2009. 1

[7] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, 2010. 6

[8] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, and A. López. Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, 16(6):820, 2016. 1, 2, 6

[9] D. Guan, Y. Cao, J. Yang, Y. Cao, and C.-L. Tisse. Exploiting fusion architectures for multispectral pedestrian detection and segmentation. *Applied optics*, 57(18):D108–D116, 2018. 3, 4

[10] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 2018. 1, 3

[11] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 1, 2, 3, 6, 7, 8

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 6

[13] L. Jingjing, Z. Shaoting, W. Shu, and M. Dimitris. Multispectral deep neural networks for pedestrian detection. In *British Machine Vision Conference*, pages 73.1–73.13, 2016. 1, 3

[14] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 49–56, 2017. 1, 3, 7, 8

[15] S. J. Krotosky and M. M. Trivedi. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 8(4):619–629, 2007. 1

[16] C. Li, D. Song, R. Tong, and M. Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference (BMVC)*, 2018. 3

[17] C. Li, D. Song, R. Tong, and M. Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019. 3

[18] L. Liu, W. Lin, L. Wu, Y. Yu, and M. Y. Yang. Unsupervised deep domain adaptation for pedestrian detection. In *European Conference on Computer Vision*, pages 676–691. Springer, 2016. 3

[19] J. Mao, T. Xiao, Y. Jiang, and Z. Cao. What can help pedestrian detection? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6043. IEEE, July 2017. 3

[20] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013. 6

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 6

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 6

[24] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5079–5087, 2015. 1

[25] P. Viola, M. Jones, et al. Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, 1:511–518, 2001. 3

[26] X. Wang, M. Wang, and W. Li. Scene-specific pedestrian detection for static video surveillance. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):361–374, 2014. 3

[27] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen. Repulsion loss: detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018. 3

[28] S. Wu, S. Wang, R. Laganiere, C. Liu, H.-S. Wong, and Y. Xu. Exploiting target data to learn deep convolutional networks for scene-adapted human detection. *IEEE Transactions on Image Processing*, 27(3):1418–1432, 2018. 3

[29] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe. Learning cross-modal deep representations for robust pedestrian de-

tection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5363–5371, 2017. 1

[30] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *European Conference on Computer Vision*, pages 443–457. Springer, 2016. 3

[31] S. Zhang, C. Bauckhage, and A. B. Cremers. Informed haar-like features improve pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 947–954, 2014. 3

[32] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1259–1267, 2016. 1

[33] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. Towards reaching human performance in pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):973–986, 2018. 1

[34] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017. 1

[35] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010. 6

[36] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 2