Visible-Thermal Pedestrian Detection via Unsupervised Transfer Learning

CHENGJIN LYU, TELIN-IPI, Ghent University - imec, Belgium
PATRICK HEYER, TELIN-IPI, Ghent University - imec, Belgium
ASAD MUNIR, Dept. of Computer Science, University of Udine, Italy
LJILJANA PLATISA, TELIN-IPI, Ghent University - imec, Belgium
CHRISTIAN MICHELONI, Dept. of Computer Science, University of Udine, Italy
BART GOOSSENS, TELIN-IPI, Ghent University - imec, Belgium
WILFRIED PHILIPS, TELIN-IPI, Ghent University - imec, Belgium

Recently, pedestrian detection using visible-thermal pairs plays a key role in around-the-clock applications, such as public surveillance and autonomous driving. However, the performance of a well-trained pedestrian detector may drop significantly when it is applied to a new scenario. Normally, to achieve a good performance on the new scenario, manual annotation of the dataset is necessary, while it is costly and unscalable. In this work, an unsupervised transfer learning framework is proposed for visible-thermal pedestrian detection tasks. Given well-trained detectors from a source dataset, the proposed framework utilizes an iterative process to generate and fuse training labels automatically, with the help of two auxiliary single-modality detectors (visible and thermal). To achieve label fusion, the knowledge of daytime and nighttime is adopted to assign priorities to labels according to their illumination, which improves the quality of generated training labels. After each iteration, the existing detectors are updated using new training labels. Experimental results demonstrate that the proposed method obtains state-of-the-art performance without any manual training labels on the target dataset.

CCS Concepts: • **Computing methodologies** → **Object detection**; • **Computer systems organization** → *Neural networks.*

Additional Key Words and Phrases: Pedestrian detection, Unsupervised transfer learning, Domain adaption, Deep neural networks

ACM Reference Format:
Chengjin Lyu, Patrick Heyer, Asad Munir, Ljiljana Platisa, Christian Micheloni, Bart Goossens, and Wilfried Philips. 2021. Visible-Thermal Pedestrian Detection via Unsupervised Transfer Learning. In *2021 the 5th International Conference on Innovation in Artificial Intelligence (ICIAI 2021), March 5–8, 2021, Xia men, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3461353.3461369

1 INTRODUCTION

Pedestrian detection, known as a sub-problem of general object detection, has attracted great attention in recent years. Meanwhile, it has been widely used in various applications, such as video surveillance [20], care for the elderly [19] and autonomous driving [22]. Despite the achievements of deep convolutional neural networks (DCNNs) [1, 8], a pedestrian detector built on a single camera

ICIAI 2021, March 5–8, 2021, Xia men, China

ACM ISBN 978-1-4503-8863-4/21/03…$15.00
https://doi.org/10.1145/3461353.3461369

might fail to work in some challenging real-world scenarios, where there are various illumination situations, shadows and weather conditions.

To overcome the limitation of a single camera, it is useful to adopt dual-camera systems where a visible and thermal camera pair could supply multimodal complementary information for better detection performance [10]. In particular, a visible camera could capture the detailed visual appearances of pedestrians under good illumination situations. Nevertheless, a detector built for visible images is easy to fail when the illumination level is low. Different from visible cameras which are similar to human eyes, thermal cameras are specially designed to sense infrared radiation. The nature of thermal imaging makes it possible to measure the temperature differences in the sensing range. For pedestrian detection, thermal images could provide robust shape information of a pedestrian against various illumination conditions, by distinguishing it from the background's thermal emission. However, compared to visible cameras, thermal cameras usually provide fewer details and are more sensitive to the environment's temperature. By combining the advantages of both visible and thermal cameras, visible-thermal fused pedestrian detectors are able to achieve better around-the-clock applications, under various illumination and weather conditions.

Although well-trained detectors based on DCNNs have already obtained significant performance [11, 12, 14], it is still very challenging to deploy an existing detector directly to new scenarios. Technically speaking, an ideal detector is supposed to have a good generalization ability and perform well across different datasets. However, well-generalizing detectors are not always available in real practice, where the camera sensors and locations of deployment might affect the properties of acquired data. There is still a performance gap between detectors trained on the target dataset and those trained on the source without transfer [4]. Therefore, human annotation of a dataset on the new scenario is always necessary for visible-thermal pedestrian detection, while it is both time-consuming and labor-intensive.

To tackle this issue, we propose an unsupervised transfer learning framework for visible-thermal pedestrian detection. The framework is designed to transfer well-trained detectors from a source dataset to the target without any manual annotations, without much performance loss. Thus, an existing detector can be easily inserted into the framework and realize a quick deployment in a new scenario. Our main contributions in this work are as follows:

(1) An iterative process is proposed to transfer an existing visible-thermal detector automatically with the help of two single-modality (visible and thermal) auxiliary detectors. During the iteration, pseudo training labels from different modalities are fused and the detectors are updated based on the generated labels.
(2) A label fusion strategy is presented, where the knowledge of daytime and nighttime is utilized to determine the fusion priorities of the labels from different modalities under different illumination situations.
(3) Experimental evaluation is conducted on the large-scale KAIST dataset [10]. The results show that our proposed method outperforms the existing state-of-the-art method and does not drop too much compared with detectors trained on manual annotations.

2 RELATED WORK

In this section, we review the related researches in the areas of visible-thermal pedestrian detection and unsupervised transfer learning.

Visible-thermal pedestrian detection. Similar to many tasks in computer vision, the research of visible-thermal pedestrian detection is data-driven. An early dataset in this area is OSU Color-Thermal Database [3], where background subtraction is used to generate region proposals for moving pedestrians. With the release of large-scale datasets (e.g., KAIST [10] and CVC-14 [5]) and

a great success of DCNNs, research on visible-thermal pedestrian detection dramatically increases. Faster R-CNN [16] is one of the most popular DCNN-based object detection methods and has become the widely used basis of visible-thermal pedestrian detectors. Liu et al. [14] first adapted Faster R-CNN and designed four architectures to fuse visible and thermal modalities in different stages. Among the four methods, Halfway Fusion provides the best performance. König et al. [11] replaced the classification network in Faster R-CNN with a Boosted Decision Trees (BDT) classifier to reduce the number of false positives. Illumination-aware weighting mechanism and semantic segmentation for visible-thermal pedestrian detection are explored in [6, 12], which could supervise the training progress in a more explicit way. Recently, to achieve a more fine-grained fusion, attention mechanisms that build connections between two modalities at multiple layers are investigated in [24, 25]. Most recently, Zhang et al. [23] proposed a new fusion method that could cyclically assign features from each modality as residuals for refinement. It is worth mentioning that the method is implemented on the single stage detector FSSD [13] rather than Faster R-CNN and trained with an auxiliary semantic segmentation layer.

Unsupervised transfer learning. Similar to traditional machine learning approaches, DCNN-based methods are sensitive to the difference between the source and target data. Generally, the need for transfer learning occurs when there is a limited supply of labeled target training data, due to the data being expensive to label or easy to leak privacy [21]. In this paper, unsupervised transfer learning is defined as the case of having no labeled training target data. Although there are numerous applications of unsupervised transfer learning in image classification [15], it is still quite challenging to perform unsupervised transfer learning on object detection, which is a more complicated task consisting of both localization and classification [9]. Specially, in the area of visible-thermal pedestrian detection, there are usually significant gaps between the source and target domain, owing to illumination, weather and camera sensor differences. Cao et al. [2] proposed an auto-annotation framework to generate pseudo training labels iteratively. Their framework could adapt a generic pedestrian detector to visible-thermal scenes without human annotation. Recently, Guan et al. [7] designed an unsupervised method to transfer a visible-thermal detector trained on a source dataset to the target. The method applies a special design of joint training of pedestrian detection and semantic segmentation, and the final outputs are full-size heat maps instead of regular bounding boxes. Motivated by these two works, we propose a unified framework to perform unsupervised transfer learning for visible-thermal pedestrian detection in this paper. With the help of single-modality auxiliary detectors and daytime/nighttime information, an existing multimodal detector could be easily adapted to a new scenario. A detailed description of the framework could be found in Section 3.

3 METHODOLOGY

3.1 Framework Overview

Let D_s denote a source domain with data and annotations (X_s, Y_s) and D_t denote the target domain with only data X_t, where manual annotations Y_t on X_t are unavailable. Specifically, source data from visible and thermal cameras are denoted as X_s^V and X_s^T, respectively. Our task in this work is to transfer a fusion detector Θ^F well-trained on a source domain D_s to the target D_t without Y_t, leveraged by single-modality auxiliary (i.e., visible Θ^V and thermal Θ^T) detectors and the knowledge of daytime and nighttime.

The overview of our proposed framework is shown in Fig. 1. Firstly, visible and thermal detectors are trained on source visible data X_s^V and thermal data X_s^T, respectively, while visible-thermal fusion detector is trained on multimodal data $X_s = \{X_s^V, X_s^T\}$. Then, the initial pseudo training labels Y_t^0 on D_t are generated by directly applying Θ_s^V and Θ_s^T on target data X_t^V and X_t^T, where

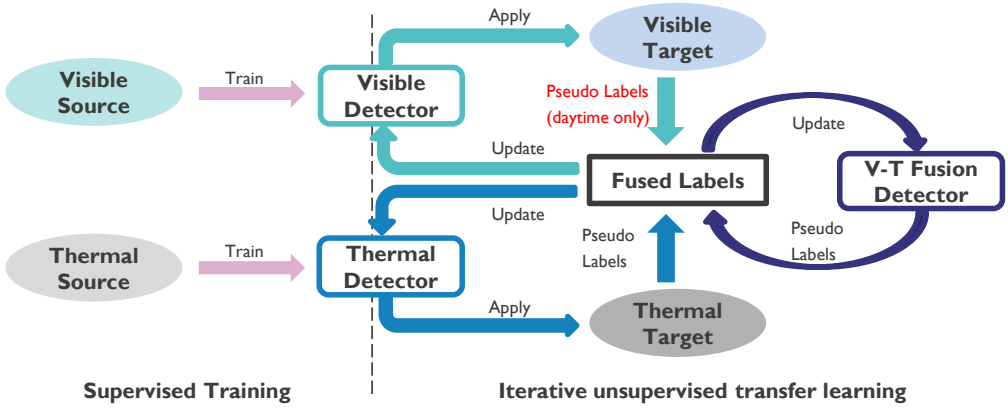

Fig. 1. Illustration of the proposed transfer learning framework. Two single-modality auxiliary detectors, i.e., visible and thermal detectors, are firstly used to generate pseudo training labels on the target dataset. With the help of daytime and nighttime information, the framework disables the visible detector when the illumination level is low (nighttime). Then fused labels are utilized to update all the existing detectors (fusion and auxiliary detectors), and the new training labels are generated iteratively.

the visible detector Θ_s^V only contributes pseudo labels on images captured in daytime where the illumination is enough. Based on the initial labels Y_t^0 on data X_t, detectors $\{\Theta_s^V, \Theta_s^T, \Theta_s^F\}$ are updated to $\{\Theta_t^V, \Theta_t^T, \Theta_t^F\}$ to fit into the new scenario. Afterward, all these three detectors are utilized to generate new training labels and themselves are updated iteratively. During the iterative process, illumination information is also used to determine the fusion priorities of labels from different modalities.

In this work, two visible-thermal fusion detectors are adopted, which are illustrated in Section 3.2. The label fusion operation is described in Section 3.3. The detailed information about how the iterative process works is given in Section 3.4.

3.2 Fusion Detectors

Here, two visible-thermal fusion detectors are adopted to fuse complementary information from two modalities, which could obtain better pedestrian detection performances compared to single-modality methods. Specifically, the widely used Faster R-CNN [16] based on the VGG16 architecture [18] is used to construct fusion detectors. An illustration of the detectors is given in Fig. 2.

Similar to Halfway Fusion [14], Feature-Map Fusion concatenates feature maps from visible and thermal modalities, where a 1×1 convolutional layer named Network-in-Network (NIN) is applied to reduce the dimension. Different from Halfway Fusion which fuses features after conv4 layers, Feature-Map Fusion is built on top of conv5 layers. It is straightforward that the fusion of independent feature maps of two modalities could make the detector focus on a bit more high-level features. The experimental results in Section 4.2 validate its effectiveness.

3.3 Label Fusion

Generally, a visible camera is sensitive to the illumination of the sensed environment and a visible-only detector always fails when the illumination is not good. According to the [12], fusion detectors could generate better results than any single-modality detectors during daytime, while the visible

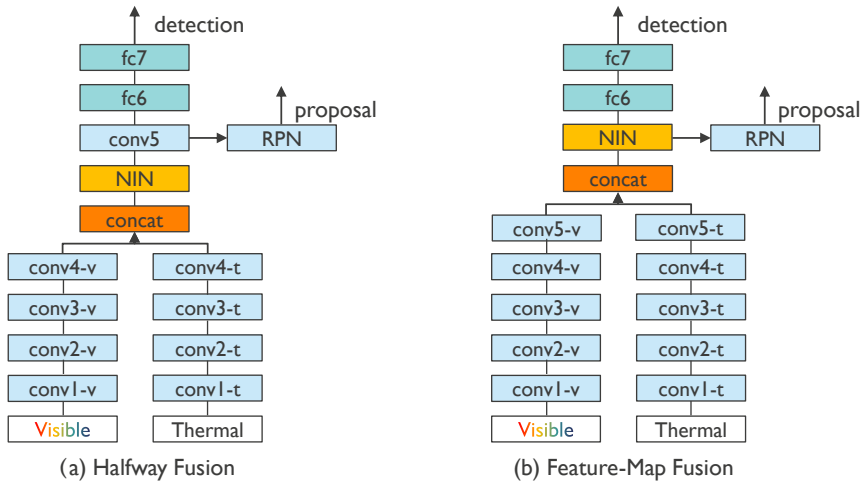

Fig. 2. Architectures of two fusion networks used in this paper. Feature-Map Fusion has a similar fusion strategy to Halfway Fusion [14], where the features from different modalities are concatenated and Network-in-Network (NIN) is used to reduce the dimension. The difference is that the fusion operation in Feature-Map Fusion happens after the extraction of single-modality feature maps.

detector outperforms the thermal slightly. As for the nighttime, they found that the visible detector becomes unreliable and the thermal detector surpasses the tested fusion detectors slightly.

Motivated by the above findings, a label fusion strategy based on the knowledge of illumination is adopted to generate the best fused labels from different modalities. Considering the ambiguous definition of illumination for pedestrian detection, we simply take the *daytime/nighttime* information as the binary classes of illumination, where images captured in daytime are treated as having a good illumination level. When the candidate labels from different modalities have bounding boxes with an Intersection Over Union (IoU) greater than 0.5, their priorities decide which bounding box can be kept as the final one. For images captured in daytime, labels from the fusion detector have the highest priority and those from the thermal detector are assigned with the lowest priority, By contrast, the visible detector gets blocked by the framework and labels from the thermal detector take a higher priority with nighttime images whose illumination level is low.

$$Priority = \begin{cases} F > V > T & \text{in daytime} \\ T > F & \text{in nighttime} \end{cases} \tag{1}$$

Specially, in the initialization phase to be mentioned in Section 3.4, the fusion detector does not take part in the generation of training labels. Thus, labels from the visible detector take a higher priority than those from the thermal detector for daytime images, and only the thermal detector generates labels for nighttime images.

$$Priority = \begin{cases} V > T & \text{in daytime} \\ T & \text{in nighttime} \end{cases} \tag{2}$$

3.4 Iterative Process

In order to transfer from a source visible-thermal dataset $X_s = \{X_s^V, X_s^T\}$ to the target $X_t = \{X_t^V, X_t^T\}$, auxiliary detectors $\{\Theta_s^V, \Theta_s^T\}$ that specialize on single modalities are adopted to generate pseudo

initial training labels, while the adapted fusion detector Θ_t^F also contributes after the initialization. In particular, the source dataset used in this work is CVC-14 [5] and the target is KAIST [10], which are two large-scale visible-thermal pedestrian detection datasets. The reason for not using Θ_s^F in the initialization is that the image pairs in CVC-14 suffer from the weakly alignments of the two modalities which lead to many problematic detection results of a fusion detector. Thus, Θ_s^F is not adopted in the initialization phase to avoid its unexpected false detection.

During the initialization and iteration phases, a detector Θ applied to the target training data generates a set of detection results $\hat{Y}$ with confidence score P. Only results with high confidence scores are selected as candidate training labels:

$$Y = \{\hat{y} \in \hat{Y} : p > p_{thr}\} \tag{3}$$

where p_{thr} is a confidence threshold. For example, in the initialization phase, we get two sets of candidate labels: Y_t^{V0} and Y_t^{T0}. After the label fusion operation based on priorities (see Section 3.3), the existing detectors are updated based on the generated training labels. In this work, we apply the early stopping mechanism to avoid overfitting and save the computation time. The maximum number of iterations is set to 2 empirically.

4 EXPERIMENTS

In this section, two public large-scale datasets KAIST [10] and CVC-14 [5] are used to conduct experiments. Fusion detectors pre-trained on CVC-14 are transferred to KAIST without using any manual annotations.

4.1 Experimental Setup

4.1.1 Dataset. **KAIST** is one of most popular visible-thermal pedestrian detection datasets covering both daytime and nighttime scenarios, which contains well-aligned image pairs with a resolution of 640×512. Following the setting in [12], a training set with 7,601 color-thermal image pairs including both non-occluded and partially-occluded instances is adopted. During unsupervised transfer learning, the manual annotations of training set are abandoned. The test set contains 2,252 images pairs sampled every 20th frame from videos, and its improved annotations provided by [14] are used in this experiment to avoid unfair comparison due to problematic and missing bounding boxes in the original test set. **CVC-14** is another large-scale dataset containing visible-thermal image pairs with a resolution of 640×480. The training set consists of 7,085 frames, while the test set contains 1,433 frames. It is worth mentioning that the annotations are provided separately in visible and thermal modalities, for the camera pair in CVC-14 is not well calibrated. In our experiment, individual annotations are used to train the auxiliary single-modality detectors, and the fusion detectors are trained with the annotations from thermal modality.

4.1.2 Implementation Details. All the detectors are implemented on Faster R-CNN [16] using VGG16 [18] pre-trained on ImageNet dataset [17] to extract features in this experiment. We re-implement Halfway Fusion [14] and insert it into the proposed framework. During the training phase, horizontal flipping is adopted to perform data augmentation. The parameters of detectors are optimized using stochastic gradient descent (SGD). For supervised training of detectors, we train the networks with a learning rate (LR) of 0.001 for 4 epochs and decay it by 0.1 for another 2 epochs. In the unsupervised transfer learning phase, all the models are fine-tuned for the first epoch with LR 0.001 and one more epoch with LR 0.0001. The confidence threshold p_{thr} is set to 0.9 in order to select the most confident labels into candidate training set.

Table 1. Ablation Study of Fusion Detectors

Methods	Miss Rate (lower, better)		
	All	Daytime	Nighttime
Pre-trained on CVC-14 without transfer:			
Halfway Fusion	66.59%	63.76%	61.72%
Feature-Map Fusion	51.94%	53.83%	44.76%
Supervised training with original annotations:			
Halfway Fusion	26.14%	24.08%	29.01%
Feature-Map Fusion	21.27%	18.63%	26.17%

Table 2. Ablation Study of Daytime/Nighttime Knowledge

Methods	Miss Rate (lower, better)		
	All	Daytime	Nighttime
Without daytime/nighttime	34.50%	39.21%	24.49%
With daytime/nighttime	23.09%	24.55%	17.74%

4.1.3 Evaluation Metric. The detection performances are reported using log-average miss rate (MR) over the range of $[10^{-2}, 10^{0}]$ false positive per image (FPPI). All the detection performances demonstrated in this work are tested on KAIST dataset.

4.2 Ablation Study

Here, we first compare the performance of Halfway Fusion and Feature-Map Fusion and the experimental results are given in Table 1. It shows that Feature-Map Fusion outperforms the classic Halfway Fusion method no matter supervised trained on KAIST or pre-trained on CVC-14 without transfer, which validates the effectiveness of fusing individual feature maps from different modalities. Next, an analytic experiment is carried out to investigate the influence of auxiliary illumination information, using Feature-Map Fusion. Firstly, only fusion detector is used to generate pseudo labels and update itself based on the generated annotations, which is the common way of transferring a generic detector. Secondly, ground-truth daytime/nighttime knowledge is given, and the iteration goes as Section 3.4 describes. In Table 2, the reported miss rate drops significantly in both daytime and nighttime scenes.

4.3 Comparison with State-of-the-art

We compare the proposed method to the state-of-the-art unsupervised transfer learning method U-TS-RPN [2], and report the supervised trained results of the widely used baseline Halfway Fusion [14]. For U-TS-RPN, the pseudo training labels are provided by the original authors and the results reported are fine-tuned with these labels using Feature-Map Fusion as the detector. The experimental results in Table 3 show our framework is superior to the existing state-of-the-art. What is more, compared to the detectors trained with manual annotations, the performance of detectors via unsupervised transfer learning does not drop a lot. In particular, for nighttime images, the performances of our method are significantly better (e.g., 17.74% vs 26.17% using Feature-Map Fusion, 18.10% vs 29.01% using Halfway Fusion). Whereas, supervised trained detectors surpass our method in the subset of daytime images. It is natural that human annotators are familiar with

Table 3. Comparison With State-of-the-Art

Methods	Miss Rate (lower, better)		
	All	Daytime	Nighttime
Supervised training with original annotations:			
Halfway Fusion [14]	26.14%	24.08%	29.01%
Feature-Map Fusion	21.27%	18.63%	26.17%
Unsupervised transfer learning:			
U-TS-RPN [2]	30.07%	31.59%	26.78%
Ours(Halfway Fusion)	27.44%	29.28%	18.10%
Ours(Feature-Map Fusion)	23.09%	24.55%	17.74%

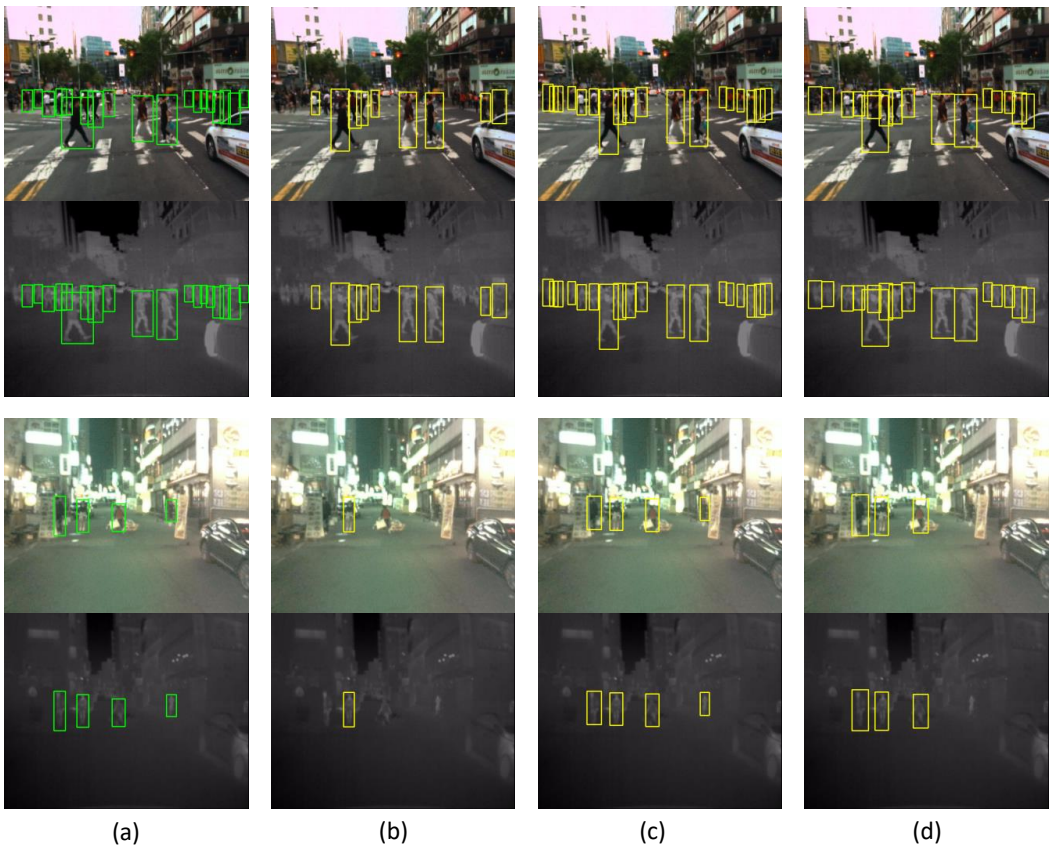

Fig. 3. Examples of pedestrian detection results on KAIST in daytime and nighttime conditions. (a) The ground-truth bounding boxes. Detection results of (b) the detector pre-trained on CVC-14, (c) the detector updated on KAIST using our proposed unsupervised transfer learning method and (d) the supervised training detector on KAIST, where Feature-Map Fusion detector is adopted in all the three experiments.

daytime images and might pay more attention on visible images, leading to high-quality annotations of daytime images. However, the labeling of nighttime scenes needs a knowledge of thermal spectral

which is not inborn. In the improved test set [14] of KAIST used in this experiment, extra efforts are made to solve the inaccurate annotation and missing labels under challenging situations, resulting in a high-quality and fair test set. Thus, the comparison with detectors trained with manual annotations of training test really validates the effectiveness of the proposed framework. Example of some pedestrian detection results are shown in Fig. 3.

5 CONCLUSION

In this paper, a unified framework to perform unsupervised transfer learning for visible-thermal pedestrian detection is proposed. With the help of auxiliary single-modality detectors and the daytime/nighttime information, pseudo training labels from different modalities are fused to generate high-quality training labels and the detectors are updated iteratively. An existing detector could be inserted into this framework without any modifications, which makes it scalable and easy to deploy. Experimental results on KAIST dataset demonstrate the effectiveness of our framework. For the future work, an intermediate domain between the source and target domains generated by adversarial learning methods could be used to reduce the distribution gap.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 765866 - ACHIEVE and the ECSEL joint undertaking grant agreement No 876487 - NextPerception.

REFERENCES

[1] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. 2019. EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 8 (2019), 1844–1861.

[2] Yanpeng Cao, Dayan Guan, Weilin Huang, Jiangxin Yang, Yanlong Cao, and Yu Qiao. 2019. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. *Inf. Fusion* 46 (2019), 206–217.

[3] James W Davis and Vinay Sharma. 2007. Background-Subtraction Using Contour-Based Fusion of Thermal and Visible Imagery. *Comput. Vis. Image. Underst.* 106, 2-3 (2007), 162–182.

[4] Kevin Fritz, Daniel König, Ulrich Klauck, and Michael Teutsch. 2019. Generalization Ability of Region Proposal Networks for Multispectral Person Detection. In *Automatic Target Recognition XXIX*, Vol. 10988. International Society for Optics and Photonics, SPIE, 109880Y.

[5] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M López. 2016. Pedestrian Detection at Day/Night Time with Visible and FIR Cameras: A Comparison. *Sensors* 16, 6 (jun 2016), 820.

[6] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. 2019. Fusion of Multispectral Data Through Illumination-Aware Deep Neural Networks for Pedestrian Detection. *Inf. Fusion* 50 (2019), 148–157.

[7] Dayan Guan, Xing Luo, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, George Vosselman, and Michael Ying Yang. 2019. Unsupervised Domain Adaptation for Multispectral Pedestrian Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 434–443.

[8] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. 2015. Taking a Deeper Look at Pedestrians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4073–4082.

[9] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. 2020. Progressive Domain Adaptation for Object Detection. In *The IEEE Winter Conference on Applications of Computer Vision*. IEEE, 749–757.

[10] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. 2015. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1037–1045.

[11] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. 2017. Fully Convolutional Region Proposal Networks for Multispectral Person Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 49–56.

[12] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. 2019. Illumination-Aware Faster R-CNN for Robust Multispectral Pedestrian Detection. *Pattern Recognit.* 85 (2019), 161–171.

[13] Zuoxin Li and Fuqiang Zhou. 2017. FSSD: Feature Fusion Single Shot Multibox Detector. arXiv:1712.00960

[14] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris Metaxas. 2016. Multispectral Deep Neural Networks for Pedestrian Detection. In *Procedings of the British Machine Vision Conference*. BMVC Press, 73.1–73.13.

[15] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. 2019. Transferrable Prototypical Networks for Unsupervised Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2239–2247.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vision* 115, 3 (2015), 211–252.

[18] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[19] Markus D Solbach and John K Tsotsos. 2017. Vision-Based Fallen Person Detection for the Elderly. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 1433–1442.

[20] Xiaogang Wang, Meng Wang, and Wei Li. 2013. Scene-Specific Pedestrian Detection for Static Video Surveillance. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 2 (2013), 361–374.

[21] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A Survey of Transfer Learning. *J. Big Data* 3, 1 (2016), 9.

[22] Zhiheng Yang, Jun Li, and Huiyun Li. 2018. Real-time Pedestrian and Vehicle Detection for Autonomous Driving. In *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 179–184.

[23] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. 2020. Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks. In *IEEE International Conference on Image Processing*. IEEE, 276–280.

[24] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. 2019. Cross-Modality Interactive Attention Network for Multispectral Pedestrian Detection. *Inf. Fusion* 50 (2019), 20–29.

[25] Yongtao Zhang, Zhishuai Yin, Linzhen Nie, and Song Huang. 2020. Attention Based Multi-Layer Fusion of Multispectral Images for Pedestrian Detection. *IEEE Access* 8 (2020), 165071–165084.