ROBUST THERMAL INFRARED PEDESTRIAN DETECTION BY ASSOCIATING VISIBLE PEDESTRIAN KNOWLEDGE

*Sungjune Park** *Dae Hwi Choi*†* *Jung Uk Kim* *Yong Man Ro*‡

Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

ABSTRACT

Recently, pedestrian detection on thermal infrared images has shown the robust pedestrian detection performance. In this paper, we propose a novel thermal infrared pedestrian detection framework which can associate and utilize the complementary pedestrian knowledge from visible images. Motivated by that humans can associate useful information from other sensors to perform a more reliable decision, we devise a Visible-sensory Pedestrian Associating (VPA) Memory to conduct the robust pedestrian detection by utilizing complementary visible-sensory pedestrian knowledge explicitly. The VPA Memory is trained to store the pedestrian information of visible images and associate it with a given thermal infrared pedestrian knowledge via the memory associating learning. We verify the effectiveness of the proposed framework with extensive experiments, and it achieves state-of-the-art pedestrian detection performance on thermal infrared images.

Index Terms— Thermal infrared pedestrian detection, Complementary visible-sensory pedestrian knowledge, Visible-sensory Pedestrian Associating (VPA) Memory

1. INTRODUCTION

Pedestrian detection is one of the most important topics in computer vision, because it has been applied into various real-world applications such as autonomous vehicle, unmanned surveillance systems, *etc.* [1, 2]. Pedestrian detection usually uses either visible (RGB) or thermal (infrared, IR) image, that is, single-sensory pedestrian detection, and the advent of deep learning has accelerated the performance improvement of the single-sensory pedestrian detection greatly [3–6].

Furthermore, to exploit the plentiful information of visible images (*e.g.,* visual appearances) and illumination-robust thermal images together, many single-sensory methods have been proposed to utilize them together in the training phase [5–7]. For example, to alleviate the illumination problem, a multi-scale detection network was proposed to transfer illumination knowledge of thermal images to visible image features [7]. Also, several thermal infrared pedestrian detection methods have been emerged to prevent privacy problem which is provoked by taking visible images in the inference time, while using thermal and visible images together in the training time [5, 6, 8]. Generative adversarial networks (GANs) were used to generate synthetic thermal images from visible images to augment thermal images [5], and a task-conditioned domain-adaptation network was introduced by conditioning both thermal and visible images on detection heads [6]. However, there is no explicit guidance to complement pedestrian features (*e.g.,* thermal) by exploiting the pedestrian knowledge of another sensor (*e.g.,* visible).

In this paper, we propose a novel thermal infrared pedestrian detection framework which could associate and utilize the complementary pedestrian knowledge from visible images explicitly. The motivation of the paper is that humans can associate useful knowledge of other sensors to conduct a more reliable decision [9]. For example, when humans try to find the pedestrian on thermal images, the corresponding pedestrian in visible images would be helpful to find where pedestrians are located at. Based on the motivation, we devise a Visible-sensory Pedestrian Associating (VPA) Memory, and it is designed to store the pedestrian information of visible images and associate it with a given thermal-sensory pedestrian feature explicitly. Therefore, it could utilize the complementary knowledge of visible-sensory pedestrians, such as visual appearances, even without the visible images during the inference time. Also, in order to guide the VPA memory to address relevant knowledge of visible-sensory pedestrians, we introduce a memory associating learning. We verify the effectiveness of the proposed framework with the VPA Memory through extensive experiments by using two pedestrian detection datasets (*i.e.,* KAIST Multispectral Pedestrian Detection Dataset [10] and CVC-14 [11]). The proposed framework achieves state-of-the-art performance, and the comprehensive visualization results also corroborate the effectiveness of our framework.

2. PROPOSED METHOD

2.1. Overall Architecture

The overall architecture of the proposed framework is shown in Fig. 1. During the training phase, two backbone networks for thermal and visible images are utilized to extract each im-

*Both authors have contributed to this paper equally.

†He is currently working at Samsung Electronics.

‡Corresponding author: Yong Man Ro, ymro@kaist.ac.kr

This work was partly supported by the IITP grant funded by the MSIT (No. 2020-0-00004).

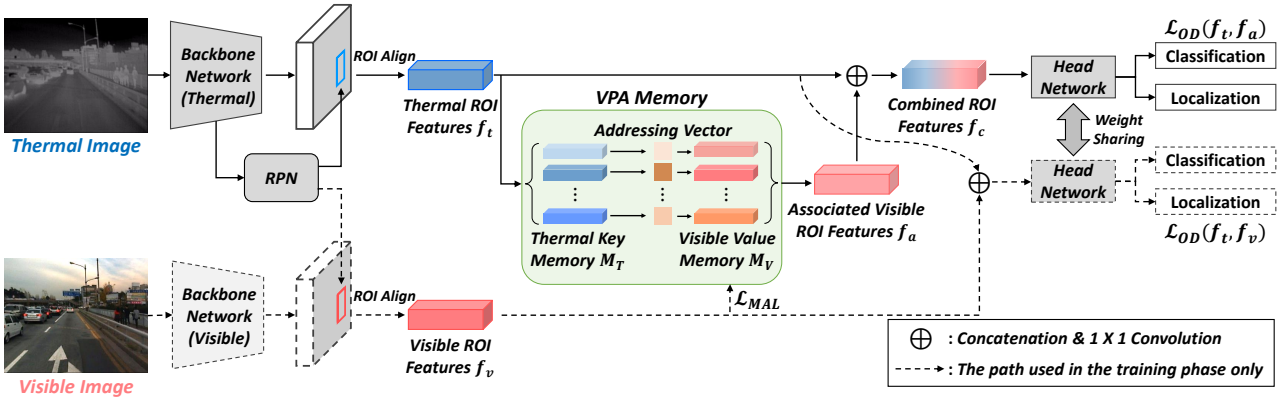

Fig. 1. Overall architecture of the proposed framework. $\boldsymbol{f_t}$ is fed into the VPA Memory to obtain $\boldsymbol{f_a}$, while $\boldsymbol{f_v}$ is used for the memory associating learning. Please note that the dotted lines indicate the path used in the training phase only, and $\mathcal{L}_{OD}(\cdot)$ denotes both classification and localization losses by taking the features combined with two ROI features.

age feature maps. Based on the thermal image feature map, a region proposal network (RPN) proposes region of interests (ROIs). Then, both thermal and visible ROI features, $\boldsymbol{f_t}, \boldsymbol{f_v} \in \mathbb{R}^{h\times w\times c}$, are extracted by ROI Align [12, 13]. $\boldsymbol{f_t}$ is fed into the Visible-sensory Pedestrian Associating (VPA) Memory to associate visible-sensory pedestrian features and obtain associated visible ROI features $\boldsymbol{f_a} \in \mathbb{R}^{h\times w\times c}$. Through the VPA Memory, we aim to obtain $\boldsymbol{f_a}$ which can provide the complementary pedestrian knowledge of visible sensor. To this end, we introduce a memory associating learning (that is, $\mathcal{L}_{MAL}$) to train the VPA Memory to store the information of visible-sensory pedestrians and associate it with a given $\boldsymbol{f_t}$ by taking $\boldsymbol{f_v}$. Finally, $\boldsymbol{f_t}$ and $\boldsymbol{f_a}$ are combined to generate combined ROI features $\boldsymbol{f_c} \in \mathbb{R}^{h\times w\times c}$ which is used for the classification and localization. Note that, during the inference time, the dotted lines (*i.e.,* visible image paths) are not utilized, and only thermal images are used. Even without the visible image input, the VPA Memory could associate the stored knowledge of visible-sensory pedestrians, such as visual appearances, and obtain $\boldsymbol{f_a}$, so that, our framework could conduct the robust pedestrian detection on thermal images.

2.2. VPA Memory

The VPA Memory is designed to store and associate the visible-sensory pedestrian knowledge (*e.g.,* visual appearances) in order to complement $\boldsymbol{f_t}$ with the obtained $\boldsymbol{f_a}$. The details of the VPA Memory are described in Fig. 2. The VPA Memory has a key-value memory structure [14], which consists of thermal key memory $\boldsymbol{M_T} = \{\boldsymbol{m_t^i}\}_{i=1}^N$ and visible value memory $\boldsymbol{M_V} = \{\boldsymbol{m_v^i}\}_{i=1}^N$, where $\boldsymbol{m_t^i}, \boldsymbol{m_v^i} \in \mathbb{R}^{h\times w\times c}$. As shown in the figure, the VPA Memory takes the $\bar{\boldsymbol{f}}_{\boldsymbol{t}} \in \mathbb{R}^{1\times hwc}$ (that is, flattened $\boldsymbol{f_t}$) as an input and measures feature cosine similarity between $\bar{\boldsymbol{f}}_{\boldsymbol{t}}$ and the flattened thermal key memory features $\bar{\boldsymbol{m}}_{\boldsymbol{t}}^{\boldsymbol{i}} \in \mathbb{R}^{1\times hwc}$. Then, thermal similarity vector $\boldsymbol{S_T} = \{s_t^i\}_{i=1}^N$ is obtained as follows:

$$s_t^i = \frac{\bar{\boldsymbol{f}}_{\boldsymbol{t}} \cdot \bar{\boldsymbol{m}}_{\boldsymbol{t}}^{\boldsymbol{i}\top}}{\|\bar{\boldsymbol{f}}_{\boldsymbol{t}}\|_2 \, \|\bar{\boldsymbol{m}}_{\boldsymbol{t}}^{\boldsymbol{i}}\|_2}. \tag{1}$$

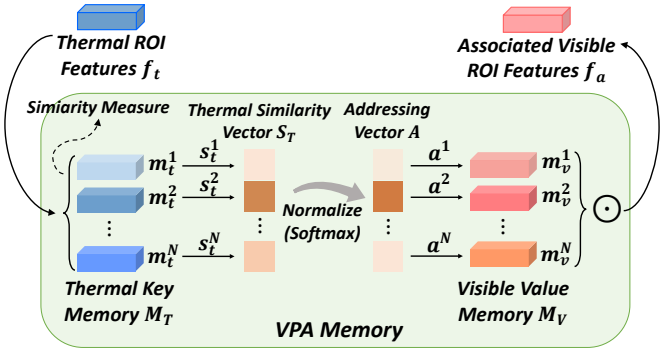

Fig. 2. The details of the VPA Memory. $\odot$ denotes the weighted summation.

Then, $\boldsymbol{S_T}$ is normalized via softmax to generate addressing vector $\boldsymbol{A} = \{a^i\}_{i=1}^N$ which determines the weights to read the relevant slots of $\boldsymbol{M_V}$ to associate the visible-sensory pedestrian knowledge. By using addressing vector $\boldsymbol{A}$, the output of the VPA Memory is obtained by weighted summation of the flattened visible value memory features $\bar{\boldsymbol{m}}_{\boldsymbol{v}}^{\boldsymbol{i}} \in \mathbb{R}^{1\times hwc}$:

$$\bar{\boldsymbol{f}}_{\boldsymbol{a}} = \sum_{i=1}^{N} a^i \cdot \bar{\boldsymbol{m}}_{\boldsymbol{v}}^{\boldsymbol{i}}. \tag{2}$$

After that, $\bar{\boldsymbol{f}}_{\boldsymbol{a}} \in \mathbb{R}^{1\times hwc}$ is reshaped into $\boldsymbol{f_a} \in \mathbb{R}^{h\times w\times c}$. Note that the purpose of the VPA Memory is to associate visible-sensory pedestrian knowledge by taking an input of thermal pedestrian features and complement it to be more discriminative with the associated visible-sensory pedestrian features. In order to achieve the purpose of the VPA Memory, we also design a memory associating learning to store and associate the knowledge of visible-sensory pedestrians.

2.3. Memory Associating Learning

First, we guide the output of the VPA Memory, $\boldsymbol{f_a}$, to contain useful and complementary knowledge of visible-sensory pedestrians. Therefore, we guide $\boldsymbol{f_a}$ to be similar with $\boldsymbol{f_v}$ (which is the original visible ROI features) as follows:

$$\mathcal{L}_{m_1} = \|\boldsymbol{f_a} - \boldsymbol{f_v}\|_2. \tag{3}$$

Method	All	Day	Night
Domain Adaptor [5]	42.65	49.59	26.70
Bottom-up [8]	35.20	40.00	20.50
TC Thermal [6]	28.53	36.59	11.03
TC Det [6]	27.11	34.81	10.31
Kieu *et al.* [16]	25.62	31.86	12.92
Proposed Framework	**16.13**	**20.29**	**7.40**

Table 1. Comparison of detection performances with the existing thermal image based pedestrian detection methods on KAIST dataset.

Second, addressing vector $\boldsymbol{A}$ is required to address the knowledge of $\boldsymbol{f_v}$ properly which is relevant of $\boldsymbol{f_t}$. Therefore, we make $\boldsymbol{S_T}$ (obtained by measuring similarity between $\boldsymbol{f_t}$ and $\boldsymbol{M_T}$) correspond to visible similarity vector $\boldsymbol{S_V} = \{s_v^i\}_{i=1}^N$ (calculated between $\boldsymbol{f_v}$ and $\boldsymbol{M_V}$ as like Eq. (1)). By doing so, we could guide each $\boldsymbol{m_v^i}$ to be corresponding to $\boldsymbol{m_t^i}$ as follows:

$$\mathcal{L}_{m_2} = \|\boldsymbol{S_T} - \boldsymbol{S_V}\|_2. \tag{4}$$

By incorporating $\mathcal{L}_{m_1}$ and $\mathcal{L}_{m_2}$, we design the memory associating learning loss as follows:

$$\mathcal{L}_{MAL} = \mathcal{L}_{m_1} + \mathcal{L}_{m_2}. \tag{5}$$

2.4. Total Training Loss

We incorporate $\mathcal{L}_{MAL}$ with loss functions of RPN ($\mathcal{L}_{RPN}$) and head network ($\mathcal{L}_{OD}$) which are composed of classification and bounding box regression losses [15]. As shown in Fig. 1, first, $\boldsymbol{f_t}$ and $\boldsymbol{f_a}$ are used for the final classification and localization (that is, $\mathcal{L}_{OD}(\boldsymbol{f_t}, \boldsymbol{f_a})$), and $\boldsymbol{f_t}$ and $\boldsymbol{f_v}$ are used to make $\boldsymbol{f_v}$ contain useful information of visible images with $\mathcal{L}_{OD}(\boldsymbol{f_t}, \boldsymbol{f_v})$. The total loss becomes as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_{RPN} + \mathcal{L}_{OD}(\boldsymbol{f_t}, \boldsymbol{f_a}) + \mathcal{L}_{OD}(\boldsymbol{f_t}, \boldsymbol{f_v}) + \lambda\mathcal{L}_{MAL} \tag{6}$$

3. EXPERIMENTS

3.1. Datasets and Implementation Details

To evaluate our framework, we use two public datasets: KAIST Multispectral Pedestrian Dataset [10] (we denote it as KAIST dataset) and CVC-14 [11]. Both datasets include thermal and visible image pairs of various driving scenes. We divide training and test sets and use training annotations from [17] and test annotations from [18] as like [6, 16]. Similar with KAIST dataset, CVC-14 consists of 7,095 training image pairs and 1,433 test image pairs [11], and we follow the dataset protocol in [19, 20].

We build our framework on Faster R-CNN [15] with VGG-16 [21]. We adopt horizontal flip for data augmentation [19]. We train the framework for 3 epochs with the stochastic gradient descent (SGD) optimizer and initial learning rate is 0.008. We set the number of memory slots $N = 100$, and the size of ROI feature as $h = 7$, $w = 7$ and $c = 512$ following [15].

Method	All	Day	Night
Baseline (w/o VPA Memory)	26.52	34.54	17.30
Proposed Framework	**21.70**	**28.60**	**13.80**

Table 2. Detection performances on CVC-14 comparing with the baseline.

N	0	10	50	100
MR ('All')	28.13	19.05	18.47	**16.13**

Table 3. Ablation study on KAIST dataset with varying memory slot size N.

3.2. Detection Performances

To investigate the effectiveness of the proposed framework, we evaluate it on thermal images and compare other thermal image based pedestrian detection methods with KAIST dataset as shown in Table 1. We adopt the Miss Rate (MR) averaged over the false positive per image (FPPI) in the range of $[10^{-2}, 10^0]$ following [16, 19, 20]. As shown in the table, our framework outperforms the existing methods with a large margin on every set ('All', 'Day', and 'Night'). Please note that, 'Day' and 'Night' are the evaluation sets composed of the images captured at day-time and night-time, respectively, and 'All' is the combined evaluation set of 'Day' and 'Night'.

We also evaluate the proposed framework with CVC-14 as shown in Table 2. Compared with the baseline thermal image based detection framework (which is not trained with the VPA Memory), our framework achieves remarkable performance improvement in three sets.

3.3. Ablation Study

As mentioned in Section. 3.1, we set the number of memory slot $N = 100$ as a default. In this subsection, we conduct ablation study on KAIST dataset by varying N with $0, 10, 50$, and 100 as shown in Table 3. When $N = 0$, it represents the baseline thermal image based pedestrian detection framework which has no VPA Memory. As shown in the table, compared to the baseline, the VPA Memory even with small $N = 10$ brings a large improvement of detection performance. When $N = 100$, the proposed framework with the VPA Memory achieves the highest detection performance.

3.4. Visualization Results

3.4.1. ROI Feature Visualization

To demonstrate how the VPA Memory complements $\boldsymbol{f_t}$, we visualize the ROI feature maps, $\boldsymbol{f_t}$, $\boldsymbol{f_a}$, and $\boldsymbol{f_c}$, following [22] in Fig. 3. Also, there are $\boldsymbol{S_T}$ and examples of high- and low-weighted $\boldsymbol{m_v^i}$. As shown in the figure, the estimated $\boldsymbol{f_t}$ could not focus on the whole pedestrian. We analyze that, since the lower-body of pedestrian is similarly activated with the backgrounds as shown in the estimated ROI image, it could not capture the lower-body. Nevertheless, the VPA Memory could associate $\boldsymbol{f_t}$ with visible-sensory pedestrian knowledge. With the help of the high-weighted $\boldsymbol{m_v^5}$, $\boldsymbol{m_v^{25}}$, and $\boldsymbol{m_v^{85}}$ which capture pedestrian properly, $\boldsymbol{f_a}$ could capture the

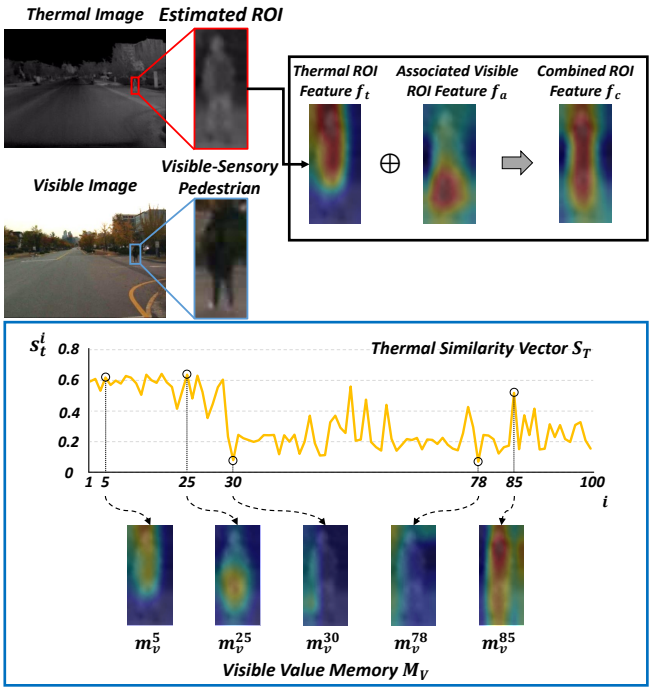

Fig. 3. The visualization of feature maps, f_t, f_a, and f_c. Due to high-activated m_v^i, the obtained f_a complements f_t, so that, f_a can capture the whole pedestrian properly.

lower-body of pedestrian. Then, finally, f_c could focus on the whole pedestrian well. On the other hand, the low-weighted m_v^{30} and m_v^{78} could not focus on the pedestrian. Note that visible image is not used in the inference time.

3.4.2. Detection Result Visualization

By using KAIST dataset, we also visualize the detection results of our framework and compare it with the baseline thermal image based pedestrian detection framework (which does not use the VPA Memory) as shown in Fig. 4. The green and red boxes represent the ground-truth and predicted bounding boxes, respectively. A yellow circle indicates the detection failure case, and a yellow-dotted circle means the false-positive (FP) case. As shown in the day-time images of the figure, the baseline could not distinguish the pedestrian properly, and it usually detects trees (which look similar in the thermal images) as pedestrians. Also, in the night-time, the baseline fails to ignore the the back-light of the vehicle (which is activated in the thermal image). On the other hand, since thermal pedestrian features could be complemented by visible-sensory pedestrian knowledge via the VPA Memory even without the visible image inputs, our framework detects and separates the pedestrians well from the backgrounds.

4. DISCUSSION

We compare the detection performances with the existing multispectral pedestrian detection methods (which take both thermal and visible images) on CVC-14, to show whether the proposed framework could exploit the knowledge of visible-sensory pedestrians properly. As shown in Table 4, our framework achieves the promising performances even without visible inputs. Such experimental results could corroborate that our framework utilizes the knowledge of visible-sensory pedestrians and complements thermal pedestrian features, even without a visible image input in the inference time.

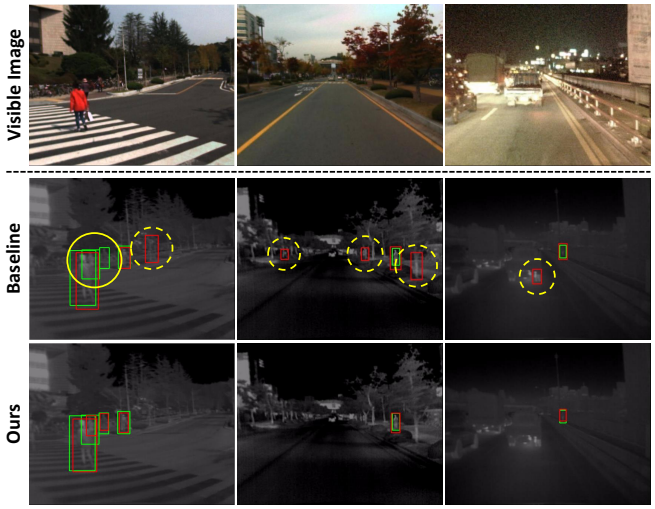

Fig. 4. The visualization of detection results of our framework, comparing with the baseline. The green and red boxes represent the ground-truth and estimated bounding boxes. Note that visible images are not used in the inference time.

Method	All	Day	Night
Halfway Fusion [20]	37.0	38.1	34.4
Park *et al.* [20]	31.4	31.8	30.8
AR-CNN [19]	22.1	24.7	18.1
MBNet [23]	21.1	24.7	13.5
Kim *et al.* [13]	20.0	23.5	12.6
Proposed Framework	**21.7**	**28.6**	**13.8**

Table 4. Detection performances on CVC-14 comparing with the multispectral pedestrian detection frameworks.

5. CONCLUSION

In this paper, we propose a novel thermal image based pedestrian detection framework to perform the robust pedestrian detection by exploiting visible-sensory pedestrian knowledge explicitly. To build the framework, we introduce the Visible-sensory Pedestrian Associating (VPA) Memory which is guided to store and associate the pedestrian knowledge of visible images, so that, the VPA Memory could complement thermal pedestrian features to be more discriminative. Also, we design a memory associating learning to address relevant knowledge of visible-sensory pedestrians properly. Along with that the proposed framework achieves state-of-the-art pedestrian detection performances on thermal images, the qualitative visualization results corroborate the effectiveness of our framework. We hope that our effective framework could be applied usefully into wide range of research areas, such as general object detection.

6. REFERENCES

[1] Wei Ke, Yao Zhang, Pengxu Wei, Qixiang Ye, and Jianbin Jiao, "Pedestrian detection via pca filters based convolutional channel features," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1394–1398.

[2] Zhiheng Yang, Jun Li, and Huiyun Li, "Real-time pedestrian and vehicle detection for autonomous driving," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 179–184.

[3] Chunze Lin, Jiwen Lu, Gang Wang, and Jie Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 732–747.

[4] Tianrui Liu, Jun-Jie Huang, Tianhong Dai, Guangyu Ren, and Tania Stathaki, "Gated multi-layer convolutional feature extraction network for robust pedestrian detection," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3867–3871.

[5] Tiantong Guo, Cong Phuoc Huynh, and Mashhour Solh, "Domain-adaptive pedestrian detection in thermal images," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1660–1664.

[6] My Kieu, Andrew D Bagdanov, Marco Bertini, and Alberto Del Bimbo, "Task-conditioned domain adaptation for pedestrian detection in thermal imagery," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2020, pp. 546–562.

[7] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 5363–5371.

[8] My Kieu, Andrew D Bagdanov, Marco Bertini, and Alberto Del Bimbo, "Domain adaptation for privacy-preserving pedestrian detection in thermal imagery," in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 203–213.

[9] Gemma A Calvert and Thomas Thesen, "Multisensory integration: methodological approaches and emerging principles in the human brain," *Journal of Physiology-Paris*, vol. 98, no. 1-3, pp. 191–205, 2004.

[10] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 1037–1045.

[11] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M López, "Pedestrian detection at day/night time with visible and fir cameras: A comparison," *Sensors*, vol. 16, no. 6, pp. 820, 2016.

[12] Sungjune Park, Jung Uk Kim, Yeon Gyun Kim, Sang-Keun Moon, and Yong Man Ro, "Robust multispectral pedestrian detection via uncertainty-aware cross-modal learning," in *International Conference on Multimedia Modeling (MMM)*. Springer, 2021, pp. 391–402.

[13] Jung Uk Kim, Sungjune Park, and Yong Man Ro, "Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[14] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston, "Key-value memory networks for directly reading documents," *arXiv preprint arXiv:1606.03126*, 2016.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[16] My Kieu, Lorenzo Berlincioni, Leonardo Galteri, Marco Bertini, Andrew D Bagdanov, and Alberto Del Bimbo, "Robust pedestrian detection in thermal imagery using synthesized images," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 8804–8811.

[17] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas, "Multispectral deep neural networks for pedestrian detection," in *27th British Machine Vision Conference (BMVC)*, 2016.

[18] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *29th British Machine Vision Conference (BMVC)*, 2018.

[19] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5127–5137.

[20] Kihong Park, Seungryong Kim, and Kwanghoon Sohn, "Unified multi-spectral pedestrian detection based on probabilistic fusion networks," *Pattern Recognition*, vol. 80, pp. 143–155, 2018.

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

[22] Jung Uk Kim, Sungjune Park, and Yong Man Ro, "Towards human-like interpretable object detection via spatial relation encoding," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3284–3288.

[23] Kailai Zhou, Linsen Chen, and Xun Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2020, pp. 787–803.