Towards Versatile Pedestrian Detector with Multisensory-Matching and Multispectral Recalling Memory

Jung Uk Kim, Sungjune Park, Yong Man Ro*

Image and Video Systems Lab, KAIST, South Korea
{jukim0701, sungjune-p, ymro}@kaist.ac.kr

Abstract

Recently, automated surveillance cameras can change a visible sensor and a thermal sensor for all-day operation. However, existing single-modal pedestrian detectors mainly focus on detecting pedestrians in only one specific modality (*i.e.,* visible or thermal), so they cannot effectively cope with other modal inputs. In addition, recent multispectral pedestrian detectors have shown remarkable performance by adopting multispectral modalities, but they also have limitations in practical applications (*e.g.,* different Field-of-View (FoV) and frame rate). In this paper, we introduce a versatile pedestrian detector that shows robust detection performance in any single modality. We propose a multisensory-matching contrastive loss to reduce the difference between the visual representation of pedestrians in the visible and the thermal modalities. Moreover, to make the proposed method perform robust detection on a single modality, we design a Multispectral Recalling (MSR) Memory. The MSR Memory enhances the visual representation of the single modal features by recalling that of the multispectral modalities. To guide the MSR Memory to store the contexts of the multispectral modalities, we introduce a multispectral recalling loss. It enables the pedestrian detector to encode more discriminative features with a single input modality. We would like to insist that our method is a step forward detector that can be applied to a variety of real-world applications. The comprehensive experimental results verify the effectiveness of the proposed method.

Introduction

Pedestrian detection is a widely studied task in computer vision, and it has been applied to massive real-world applications (Dalal and Triggs 2005; Benenson et al. 2014; Chi et al. 2020). When detecting pedestrians, many methods have utilized a single camera sensor. Traditional methods commonly adopt a visible modality (*e.g.,* RGB), because this modality contains various color information about pedestrians (Zhang et al. 2021). In addition, recent works have adopted a thermal modality, because this modality is robust to various challenging environments, such as low illumination (Li et al. 2019b) and background clutter (Song et al. 2019).

Since the two modalities have different properties, recent surveillance cameras automatically switch the input modality between the visible modality and the thermal modality depending on the situation for around-the-clock operation (Wu et al. 2017, 2020a). In general, the visible modality is adopted in day time where color information is well observed. And, the thermal modality is used in cloudy weather or dark nights where pedestrians cannot be seen but heat information can be detected. Therefore, it is necessary to build a deep network that is robust to changes in the input modality to automatically detect pedestrians. However, previous single-modal methods have a limitation in that they mainly focused on detecting pedestrians in only one specific modality (*i.e.,* visible or thermal), so that they cannot effectively cope with other modal inputs (Wu et al. 2020b; Kieu et al. 2020). For instance, thermal-based pedestrian detectors cannot handle visible modality due to the significant differences between the two modalities (Li et al. 2019a). Some efforts in the person re-identification task have been made to cope with changes in input modality (Wu et al. 2017, 2020a). However, they need the additional modal-specific networks, and the users should know the input modality type in advance to switch modal-specific networks.

Moreover, some recent works have studied on multispectral pedestrian detection which utilizes both visible modality and thermal modality simultaneously (Hwang et al. 2015; Zhang et al. 2019; Kim, Park, and Ro 2021b). Since this task can utilize complementary visual information of the two modalities, multispectral pedestrian detectors can encode more discriminative features (Hwang et al. 2015), showing remarkable performance. However, multispectral pedestrian detection is somewhat difficult to be applied to practical applications for several reasons. For example, the multispectral pedestrian detectors need an image pair of two modalities that have the same Field-of-View (FoV) and same frame rate (Zhang et al. 2019) in the inference phase. Also, the computational cost is much higher than the single modal detectors.

Based on the above discussions of the single modality and the multispectral modalities, it is essential to build a versatile pedestrian detector that overcomes the above limitations. Therefore, we aim to build a pedestrian detector that shows robust performance while freely receiving the visible modality and the thermal modality. To this end, we consider following two issues: (1) how to cope with a single input regardless of the input modality type and (2) how to enhance the visual representation of a single modality for the robust

*Corresponding author

pedestrian detector.

In this paper, we propose a novel pedestrian detector by addressing the raised two issues. First, in order to make the modality-agnostic detector, we propose a multisensory-matching contrastive loss. The multisensory-matching contrastive loss guides the two modal pedestrian features to be similar by referring to the counterpart modal features. It makes the pedestrian features close to each other and separates pedestrian features and background features. Through the multisensory-matching contrastive loss, the proposed detector can encode modality-agnostic pedestrian features. Second, to improve the visual representation of a single modality, we design a Multispectral Recalling (MSR) Memory. The role of the MSR Memory is to enhance the single modal pedestrian features by addressing relevant information to recall the visual appearance of multispectral modalities that contain richer visual information. In order for the MSR Memory to memorize the multispectral contexts, we introduce a multispectral recalling loss. As a result, more discriminative pedestrian features can be encoded from any single modal input through the MSR Memory.

Our work is to design a versatile pedestrian detector that shows robustness for any single modality by recalling the multispectral information. Therefore, we would like to claim that the proposed method is a step forward method that can be applied to various real-world applications, such as recent surveillance cameras that automatically switch between the two modalities. Through the comprehensive experimental results and visualization results, we demonstrate the effectiveness of the proposed pedestrian detector with the multisensory-matching and the MSR Memory.

The major contributions of this work can be summarized as follows:

- We introduce a multisensory-matching contrastive loss in order to guide the network to have similar visual representations about pedestrian regardless of the modality.
- We devise MSR Memory that can recall the visual appearance of the multispectral modalities. To effectively memorize contexts of the multispectral modalities, we propose a multispectral recalling loss.
- Our pedestrian detector shows comparable performance to the multispectral pedestrian detector while receiving a single modal input regardless of modality type.

Related Work

Pedestrian Detection in Single Modality

Pedestrian detection is one of the most widely studied topics in computer vision. It is applied to many human-related works, such as autonomous driving (Liu et al. 2021) and video surveillance (Zhu and Peng 2015). From various previous hand-craft methods (Dalal and Triggs 2005; Benenson et al. 2014) to deep learning methods (Zhang et al. 2020; Wu et al. 2020b), a visible modality (*e.g.,* RGB) is usually adopted, because this modality is the most intuitive modality for humans. Recent visible-based pedestrian detectors have introduced to alleviate the various issues, such as small-scale pedestrian (Song et al. 2018; Han et al. 2019; Wu et al. 2020b) and occlusion (Song et al. 2020; Xie et al. 2020b,a). Recently, some pedestrian detectors have been adopted a thermal modality, because this modality is robust to weather conditions (Kieu et al. 2019), low illumination (Zhou, Chen, and Cao 2020), and background clutter (Song et al. 2019).

Several single modal pedestrian detectors use both visible modality and thermal modality in the training phase to improve the detection results. Xu et al. (Xu et al. 2017) proposed sub-networks that can improve the detection results of the visible modality using the information of the thermal modality. Kieu et al. (Kieu et al. 2019) utilized the visible modality of ImageNet (Krizhevsky, Sutskever, and Hinton 2012) and MS COCO (Lin et al. 2014) datasets prehand when conducting the thermal-based pedestrian detection. The work in (Kieu et al. 2020) proposed an auxiliary day/night classifier for adapting visible modality to the thermal modality. However, they mainly focused on handling one specific modality in the inference phase. As a result, it is difficult to effectively cope with the case when the other modal input comes in. In contrast, our method is flexible to any single input modality. Therefore, our detector can show robust performance day and night regardless of the input modal type, similar to the operation of surveillance camera.

Multispectral Pedestrian Detection

Recently, multispectral pedestrian detection has been received increasing attention by using the visible modality and the thermal modality together, and it shows remarkable detection performance (Park, Kim, and Sohn 2018; Zhang et al. 2019; Zhou, Chen, and Cao 2020; Li et al. 2019b; Kim, Park, and Ro 2021b). Zhang et al. (Zhang et al. 2019) proposed AR-CNN to address the alignment issue of the two modalities. To reduce the effect of the modality discrepancy issue, Zhou et al. (Zhou, Chen, and Cao 2020) introduced MBNet which guides two modal features to be similar and selects the features according to the illumination score.

However, to apply the multispectral pedestrian detector to the real-world applications, image pairs of the two modalities should have the same Field of View (FoV) and same frame rate to avoid the miscalibration problem (Zhang et al. 2019). Kim et al. (Kim, Park, and Ro 2021b) somewhat mitigated this problem by adopting an uncertainty, but it still struggled when the FoV is significantly different (*e.g.,* pedestrians are observed in one modality but not in the other modality). Moreover, the computational cost is higher than that of the single modal detectors. In contrast, our method is free from the problem, since we utilize the single modality in the inference phase. Furthermore, it shows comparable performance with the multispectral pedestrian detectors, because the MSR Memory can recall the information of the multispectral modalities.

Memory Network

Memory augmented neural networks have been introduced, and they are applied to the various research fields (Zhu and Yang 2020; Marchetti et al. 2020; Lee et al. 2021; VS et al. 2021; Kim et al. 2021b). For example, Lee et al. (Lee et al. 2021) proposed LMC-Memory to predict the future frames by recalling long-term motion contexts. MeGA-CDA (VS

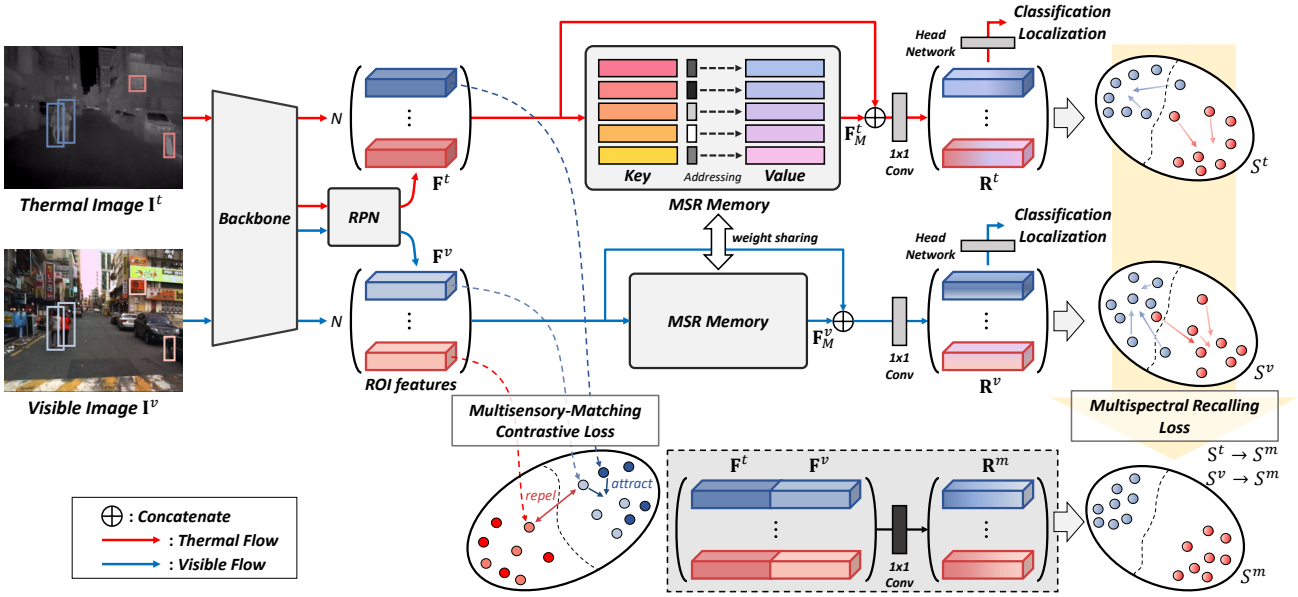


Figure 1: Network configuration of the proposed pedestrian detector in the training phase. Dotted box (*i.e.,* generating $\mathbf{R}^m$) indicates that this procedure is considered only in the training phase. $\bigoplus$ indicates the concatenation operation. Note that, in the inference phase, our detector freely receives any type of the single modality (visible (blue path) or thermal (red path)).

et al. 2021) was introduced to align category-specific features of the source and target domain. Recent works have adopted the key-value memory structure (Marchetti et al. 2020). Given query features, key memory calculates the similarity to address the relevant information of value memory. In this paper, we propose MSR Memory with multispectral recalling loss to recall the multispectral information. Thus, our method can perform robust detection like multispectral pedestrian detector, regardless of the input modality.

Proposed Method

Figure 1 shows the overall architecture of the proposed pedestrian detector in the training phase. A weight-sharing backbone network receives an image pair of visible modality $\mathbf{I}^v$ and thermal modality $\mathbf{I}^t$ to encode image features of the two modalities. The two image features are passed through the RPN (Ren et al. 2015) to estimate N candidate regions, called Region of Interests (ROIs). With the N ROIs, ROI align is conducted (He et al. 2017) to estimate ROI features of the visible modality $\mathbf{F}^v$ and the thermal modality $\mathbf{F}^t$. After that, $\mathbf{F}^v$ and $\mathbf{F}^t$ are passed through the weight sharing Multispectral Recalling (MSR) Memory. Let $\mathbf{F}_M^v$ and $\mathbf{F}_M^t$ denote output features of the MSR Memory. $\mathbf{F}_M^v$ and $\mathbf{F}_M^t$ are concatenated with the original ROI features $\mathbf{F}^v$ and $\mathbf{F}^t$, followed by 1×1 convolution to generate the refined ROI features $\mathbf{R}^v$ and $\mathbf{R}^t$. Finally, classification and localization are conducted through a weight-sharing head network. *Note that, in the inference phase, our pedestrian detector can receive any single input modality (i.e., visible (blue path) or thermal (red path)).*

Multisensory-Matching Contrastive Loss

For the same scene, captured images of the two modalities are from different spectral bands. Therefore, the appearance of the observed pedestrians can be different. For example, in the dark night, the visible modality may lack color information to be judged as a pedestrian, but in the thermal modality, pedestrians may be well represented through the heat information. Therefore, we propose a multisensory-matching contrastive loss to guide the feature representation of the two modal ROI features $\mathbf{F}^v$ and $\mathbf{F}^t$ which are encoded by the backbone network to be similar. At the same time, it guides the ROI features of the pedestrian class close to each other and the pedestrian class and the background to be apart by referring the counterpart modality.

Specifically, $\mathbf{F}^v = \{f_i^v\}_{i=1}^N$ and $\mathbf{F}^t = \{f_i^t\}_{i=1}^N$ ($f_i^v, f_i^t \in \mathbb{R}^{1\times w\times h\times c}$, where w, h, c indicate width, height, and channel, respectively) are passed through the global average pooling layer to generate $\mathbf{F}_G^v = \{f_{G_i}^v\}_{i=1}^N$ and $\mathbf{F}_G^t = \{f_{G_i}^t\}_{i=1}^N$ ($f_{G_i}^v, f_{G_i}^t \in \mathbb{R}^{1\times c}$). Let N_p denote the number of ROI features of the pedestrian class among the N ROI features. When α modality is given ($\alpha = \{v, t\}$), multisensory-matching contrastive loss to the counterpart β modality is defined as:

$$\mathcal{L}_{\alpha\rightarrow\beta} = -\frac{1}{N_p}\sum_{i=1}^{N_p}\log\frac{\sum_{j=1}^{N_p}\exp\left(d(f_{G_i}^{\alpha}, f_{G_j}^{\beta})/\tau\right)}{\sum_{j=1}^{N}\exp\left(d(f_{G_i}^{\alpha}, f_{G_j}^{\beta})/\tau\right)}, \quad (1)$$

$$d(f_{G_i}^{\alpha}, f_{G_j}^{\beta}) = \frac{f_{G_i}^{\alpha}\cdot f_{G_j}^{\beta}}{||f_{G_i}^{\alpha}||\,||f_{G_j}^{\beta}||}, \quad (2)$$

where τ is the temperature parameter to control the softness. The meaning of $\mathcal{L}_{\alpha\rightarrow\beta}$ is to make the N_p pedestrian ROI features of the α modality and N_p pedestrian ROI features of the β modality similar and make the background ROI features dissimilar.

Based on the $\mathcal{L}_{\alpha\rightarrow\beta}$, we introduce the multisensory-matching contrastive loss $\mathcal{L}_{mmc}$ by combining the ways of

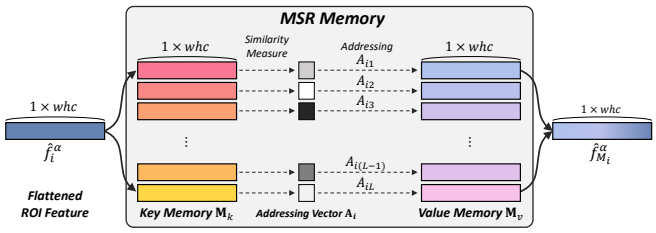


Figure 2: The key-value structure of the MSR Memory. The *i*-th ROI feature of α modality f_i^α ($\alpha = \{v, t\}$) is flattened to generate $\hat{f}_i^\alpha$. Then, $\hat{f}_i^\alpha$ is calculated with key memory $\mathbf{M}_k$ to obtain addressing vector $\mathbf{A}_i$, to control the amount of relevant information of the value memory $\mathbf{M}_v$. Finally, $\hat{f}_{M_i}^\alpha$ is obtained by a weight summation of $\mathbf{A}_i$ and $\mathbf{M}_v$.

the two modalities, which can be represented as follows:

$$\mathcal{L}_{mmc} = \mathcal{L}_{t\to v} + \mathcal{L}_{v\to t}. \tag{3}$$

With the $\mathcal{L}_{mmc}$, the backbone network can encode modality-agnostic ROI features, regardless of the input modality.

MSR Memory

Figure 2 shows the structure of the MSR Memory. The MSR Memory consists of L slot pairs of key-value memory, which can be represented as $\mathbf{M} = \{\mathbf{M}_{k_i}, \mathbf{M}_{v_i}\}_{i=1}^{L} (\mathbf{M}_{k_i}, \mathbf{M}_{v_i} \in \mathbb{R}^{1\times whc})$. The *i*-th ROI feature of the α modality f_i^α ($\alpha = \{v, t\}$) is flattened to generate a vector $\hat{f}_i^\alpha \in \mathbb{R}^{1\times whc}$. Then, the similarity between $\hat{f}_i^\alpha$ and L slots of the key memory $\mathbf{M}_k$ is calculated for obtaining the addressing vector $\mathbf{A}_i = \{A_{i1}, ..., A_{iL}\} \in \mathbb{R}^{1\times L}$ which determines how to read the relevant information of the value memory $\mathbf{M}_v$. The *j*-th element of the $\mathbf{A}_i$ is calculated as follows:

$$A_{ij} = \frac{\exp(z_{ij})}{\sum_{m=1}^{L} \exp(z_{im})}, \; z_{ij} = \frac{\hat{f}_i^\alpha \cdot \mathbf{M}_{k_j}^\top}{\sqrt{c}}. \tag{4}$$

If $\mathbf{M}_{k_j}$ is highly correlated with $\hat{f}_i^\alpha$, A_{ij} will be high. In the opposite case, A_{ij} will be low. After obtaining the addressing vector $\mathbf{A}_i$, flattened i-th output feature maps of the MSR Memory $\hat{f}_{M_i}^\alpha \in \mathbb{R}^{1\times whc}$ is calculated as:

$$\hat{f}_{M_i}^\alpha = \sum_{j=1}^{L} A_{ij} \cdot \mathbf{M}_{v_j}. \tag{5}$$

Finally, $\hat{f}_{M_i}^\alpha$ is reshaped to generate *i*-th output feature maps of the MSR Memory $f_{M_i}^\alpha \in \mathbb{R}^{1\times w\times h\times c}$. Note that, $f_{M_i}^\alpha$ is concatenated with f_i^α, followed by 1×1 convolution to obtain the refined i-th ROI feature $r_i^\alpha \in \mathbb{R}^{1\times w\times h\times c}$. The weight parameters of the MSR Memory (*i.e.,* $\mathbf{M}_k$ and $\mathbf{M}_v$) are initialized with Kaiming initialization (He et al. 2015) and they are updated through the overall training loss of the network.

Multispectral Recalling Loss

Given $\mathbf{F}^v$ and $\mathbf{F}^t$, main purpose of the proposed MSR Memory is to guide $\mathbf{R}^v = \{r_i^v\}_{i=1}^N$ and $\mathbf{R}^t = \{r_i^t\}_{i=1}^N$ to be more discriminative features. It is known that the multispectral modalities contain abundant visual information about the pedestrians (Hwang et al. 2015). Therefore, we propose a multispectral recalling loss to guide the MSR Memory to store the information about the multispectral modalities.

To this end, we first generate the reference feature which can play a role of the teacher. As shown in Figure 1 (dotted box), $\mathbf{F}^v$ is concatenated with $\mathbf{F}^t$ to perform 1×1 convolution to encode the multispectral ROI feature $\mathbf{R}^m = \{r_i^m\}_{i=1}^N$, $r_i^m \in \mathbb{R}^{1\times w\times h\times c}$. Note that, the reference multispectral ROI feature $\mathbf{R}^m$ is generated only in the training phase.

After generating $\mathbf{R}^m$, the multispectral recalling loss guides $\mathbf{R}^v$ and $\mathbf{R}^t$ to be similar with $\mathbf{R}^m$. The multispectral recalling loss consists of feature similarity embedding loss and feature relation guiding loss. First, to guarantee the feature representation of each $\mathbf{R}^v$ and $\mathbf{R}^t$ to be closed with $\mathbf{R}^m$, we introduce feature similarity embedding loss $\mathcal{L}_{fse}$ as:

$$\mathcal{L}_{fse} = \frac{1}{N}\sum_{i=1}^{N} \|r_i^m - r_i^v\|_2^2 + \|r_i^m - r_i^t\|_2^2. \tag{6}$$

Through the $\mathcal{L}_{fse}$, values of each ROI feature of $\mathbf{R}^v$ and $\mathbf{R}^t$ become similar with the corresponding ROI features of $\mathbf{R}^m$.

Second, we also consider the relationship between N ROI features in every single modality. To this end, $\mathbf{R}^v$ and $\mathbf{R}^t$ are passed through the global average pooling layer, followed by ℓ_2 normalization to obtain $l^v = \{l_i^v\}_{i=1}^N$ and $l^t = \{l_i^t\}_{i=1}^N$ ($l_i^v, l_i^t \in \mathbb{R}^{1\times c}$). In the visible modality, similarity relation between N features are measured $C^v = l^v l^{v\top} \in \mathbb{R}^{N\times N}$. Then, we normalize C^v via the softmax along the row direction to obtain similarity relation matrix S^v. The *i*-th row vector S_i^v denotes the probability of how the *i*-th visible ROI feature is similar to the N visible ROI features. Ideally, the similarity between pedestrian ROI features may be high, vice versa between the pedestrian and background ROI features. Similarly, we obtain the similarity relation matrix of the thermal modality S^t and the multispectral modalities S^m.

With the three measurements (*i.e.,* S^v, S^t, and S^m), we use KL divergence $D_{KL}(\cdot)$ to compare the probability distributions. Through the KL divergence, we propose feature relation guiding loss $\mathcal{L}_{frg}$ to guide the relationship between ROI features to be similar with S^m, which is defined as:

$$\mathcal{L}_{frg} = \frac{1}{N}\sum_{i=1}^{N} \underbrace{D_{KL}(S_i^m||S_i^v)}_{\text{visible to multispectral}} + \underbrace{D_{KL}(S_i^m||S_i^t)}_{\text{thermal to multispectral}}. \tag{7}$$

Finally, the multispectral recalling loss $\mathcal{L}_{msr}$ is obtained by adding $\mathcal{L}_{fse}$ and $\mathcal{L}_{frg}$, which is expressed as follows:

$$\mathcal{L}_{msr} = \mathcal{L}_{fse} + \mathcal{L}_{frg}. \tag{8}$$

The $\mathcal{L}_{msr}$ guides each ROI feature as well as the relationship between N ROI features to follow those of the reference features $\mathbf{R}^m$. By doing so, the MSR Memory can memorize information about pedestrian to improve the pedestrian feature representations of the original ROI features (*i.e.,* $\mathbf{F}^v$ and $\mathbf{F}^t$). Consequently, the proposed detector can perform robust pedestrian detection with any single modal inputs by recalling multispectral modal features.

Method	V (Day, Night)			T (Day, Night)			V (Day), T (Night)		
	All	**Day**	**Night**	**All**	**Day**	**Night**	**All**	**Day**	**Night**
MSDS-RCNN (BMVC'18)	82.97	76.04	97.68	36.36	39.53	28.67	N/A	76.04	28.67
AR-CNN (ICCV'19)	77.03	67.54	97.85	17.70	21.95	8.64	N/A	67.54	8.64
MBNet (ECCV'20)	80.20	71.88	100.00	55.56	57.49	46.81	N/A	71.88	46.81
Kim et al. (TCSVT'21)	57.49	37.17	87.84	17.20	22.24	7.21	23.60	37.17	7.21
MLPD (RA-L'21)	23.95	16.88	39.37	16.34	**20.07**	8.22	N/A	16.88	8.22
Proposed Method	**20.29**	**15.28**	**30.45**	**15.87**	20.26	**6.48**	**11.39**	**15.28**	**6.48**

Table 1: Detection results (MR) on KAIST dataset for the single modal inputs (V: visible, or T: thermal). We compare our method with the multispectral pedestrian detectors. It assumes the environments where only one modality is available in the inference (Kim et al. 2019). When the multispectral pedestrian detectors takes one modal images, the counterpart modal images are blackout (fill all zero values), following (Kim et al. 2021a).

Total Loss Function

The total loss function is represented as follows:

$$\begin{aligned} \mathcal{L}_{Total} &= \mathcal{L}_{OD} + \lambda_1 \mathcal{L}_{mmc} + \lambda_2 \mathcal{L}_{msr}, \\ \mathcal{L}_{OD} &= \mathcal{L}_{RPN} + \mathcal{L}_{cls} + \mathcal{L}_{loc}, \end{aligned} \tag{9}$$

where $\mathcal{L}_{OD}$ includes the loss function of the RPN, classification, and the localization of two-stage object detectors (Ren et al. 2015; Lin et al. 2017). λ_1 and λ_2 denote the balancing hyper-parameters.

Experiments

Datasets and Evaluation Metric

To validate the proposed pedestrian detector, we use two public datasets: (1) KAIST Multispectral Pedestrian Detection Dataset (Hwang et al. 2015) (for simplicity, we refer to this as KAIST dataset) and (2) CVC-14 (González et al. 2016). The two datasets contain the pair of visible-thermal images/annotations to detect the pedestrians.

KAIST Dataset The KAIST dataset is a large-scale dataset with well-aligned visible-thermal image pairs (Hwang et al. 2015). It contains 95,328 visible-thermal image pairs with 103,128 bounding-box annotations and 1,182 unique pedestrians. The visible-thermal image pairs of the KAIST dataset are taken in a driving environment with various scenes. It includes the day and night conditions. Image resolution is 512×640. Note that, in the inference phase, we use test set with 2,252 images (visible or thermal).

CVC-14 The visible-thermal image pairs of CVC-14 are also collected in a driving environment with various scenes (González et al. 2016), including day/night conditions. Image resolution is 471×640. Following (Kim, Park, and Ro 2021b), we use 7,085 image pairs in the training phase and 1,433 images (visible or thermal) in the inference phase.

Evaluation Metric Following (Hwang et al. 2015), we adopt miss rate (MR) averaged over the false positive per image (FPPI) with range of $[10^{-2}, 10^{0}]$. The lower MR denotes the better detection performance. We evaluate the performance on '**All**', '**Day**', and '**Night**' environments.

To evaluate the performance, we divide the input modality into three cases: (1) visible (day/night), (2) thermal (day/night), and (3) visible (day) and thermal (night). Note that, case (3) imitates a surveillance camera that can change the input modality depends on the day/night time (Wu et al. 2020a). We call case (3) as '*mixed modality*'.

Method	All	Day	Night
Domain Adapotor (ICIP'19)	46.30	53.37	31.63
Bottom-up (ICIAP'19)	35.20	40.00	20.50
TC Thermal (ECCV'20)	28.53	36.59	11.03
GFD-SSD (ArXiv'19)	28.00	25.80	30.03
TC Det (ECCV'20)	27.11	34.81	10.31
Kieu et al. (ICPR'20)	25.62	31.86	12.92
Kim et al. (ICCV'21)	19.16	24.70	8.26
Proposed Method	**15.87**	**20.26**	**6.48**

Table 2: Detection results (MR) on KAIST dataset for thermal modal inputs. We compare our method with the thermal-based pedestrian detectors. Note that all the methods use the visible modality and thermal modality in the training phase.

Method	All	Day	Night
Baseline (V)	65.12	32.40	82.69
Baseline (T)	68.73	95.41	23.15
Baseline ($V+T$)	28.54	36.73	21.84
Proposed Method	**19.88**	**23.69**	**12.35**

Table 3: Detection results (MR) on CVC-14 for '*mixed modality*'. 'Baseline (k)' indicates k modality is utilized in the training phase to train the single modal detector (V: visible and T: thermal).

Implementation Details

We implement the proposed detector based on Faster R-CNN (Ren et al. 2015) with VGG16 (Simonyan and Zisserman 2014) backbone. In ablation studies, we extend the baseline detector to Feature Pyramid Network (FPN) (Lin et al. 2017) with ResNet (He et al. 2016) backbone to see the generalization ability of the proposed method.

We train the proposed detector with stochastic gradient descent (SGD) which is synchronized over 4 GTX 1080 Ti GPUs with 4 images per mini-batch (1 image per GPU). We train our detector for 4 epochs with 0.008 learning rate. All experiments are conducted based on the Pytorch (Paszke et al. 2017). The number of ROIs is $N = 256$. We use the key-value slots number of the MSR Memory $L = 50$ as the default. Also, we use $\tau = 1$ and $\lambda_1, \lambda_2 = 1$.

Method	$\mathcal{L}_{mmc}$	$\mathcal{L}_{msr}$	V (Day, Night)			T (Day, Night)			V (Day), T (Night)		
			All	**Day**	**Night**	**All**	**Day**	**Night**	**All**	**Day**	**Night**
Baseline (V)	-	-	30.66	22.52	47.70	82.70	85.60	76.23	55.81	22.52	76.23
Baseline (T)	-	-	94.92	90.77	98.08	26.70	29.92	17.98	62.84	90.77	17.98
Baseline ($V+T$)	-	-	26.11	21.78	45.34	23.31	37.78	16.18	20.76	21.78	16.18
Proposed Method	✓	✗	24.10	19.47	34.20	18.58	23.92	8.63	15.89	19.47	8.63
	✗	✓	23.96	17.12	38.31	16.94	21.58	7.47	14.72	17.12	7.47
	✓	✓	**20.29**	**15.28**	**30.45**	**15.87**	**20.26**	**6.48**	**11.39**	**15.28**	**6.48**

Table 4: Effect of the proposed loss (*i.e.,* $\mathcal{L}_{mmc}$ and $\mathcal{L}_{msr}$) on KAIST dataset (V: visible and T: thermal). 'Baseline (k)' indicates k modality is adopted to train the single-modal pedestrian detector in the training phase.

Slot Number (L)	All	Day	Night
-	20.76	21.78	16.18
25	12.98	16.42	7.20
50	**11.39**	**15.28**	**6.48**
100	12.19	15.78	7.47
200	12.58	16.52	7.38

Table 5: Detection results (MR) on KAIST dataset by varying MSR Memory slot number L using '*mixed modality*'.

Results on KAIST Dataset

First, we compare the proposed method with the state-of-the-art multispectral pedestrian detectors (Li et al. 2018; Zhang et al. 2019; Zhou, Chen, and Cao 2020; Kim et al. 2021a; Kim, Park, and Ro 2021b). In the inference phase, a single modality is adopted. It assumes a real-world scenario where both modalities cannot be obtained at the same time (Kim et al. 2019). Therefore, when the multispectral pedestrian detectors take one modal inputs, we follow the settings of (Kim et al. 2021a) where the counterpart modal inputs are blackout (fill all zero values). The results are shown in Table 1. Our method mostly surpasses the multispectral pedestrian detectors. The results show that multispectral pedestrian detectors may not flexibly handle a single modal input, while the proposed method can cope with any input modality.

Next, we also compare our method with the state-of-the-art thermal-based pedestrian detectors (Herrmann, Ruf, and Beyerer 2018; Guo, Huynh, and Solh 2019; Kieu et al. 2019; Zheng, Izzat, and Ziaee 2019; Kieu et al. 2020, 2021; Kim, Park, and Ro 2021a). Note that all the methods adopt the images of the visible modality and the thermal modality in the training phase. As shown in Table 2, our method outperforms the existing thermal-based methods. We analyze that our multisensory-matching and MSR Memory, which can recall information of the multispectral modalities, lead more accurate detection performances.

Results on CVC-14

We also conduct experiments on CVC-14. It is shown in Table 3. The '*Baseline ($V+T$)*' adopting the two modalities in the training phase is better than the '*Baseline (V)*' and '*Baseline (T)*'. In contrast, our method, which considers the multisensory-matching as well as the MSR Memory with multispectral recalling loss, outperforms all the baselines with a large margin.

Detector	Backbone	Proposed Method	All	Day	Night
Faster-RCNN	VGG16	✗	20.76	21.78	16.18
		✓	**11.39**	**15.28**	**6.48**
FPN	ResNet-50	✗	18.50	20.43	15.00
		✓	**10.68**	**13.54**	**6.39**
FPN	ResNet-101	✗	17.98	20.02	14.85
		✓	**10.32**	**13.28**	**6.23**

Table 6: Detection results (MR) on KAIST dataset for various base detector using '*mixed modality*'.

Ablation Studies

We conduct various ablation studies to investigate (1) effect of the proposed losses (*i.e.,* $\mathcal{L}_{mmc}$ and $\mathcal{L}_{msr}$), (2) effect of the slot number L in MSR Memory, and (3) baseline pedestrian detector extension. All the experiments are conducted on the KAIST dataset.

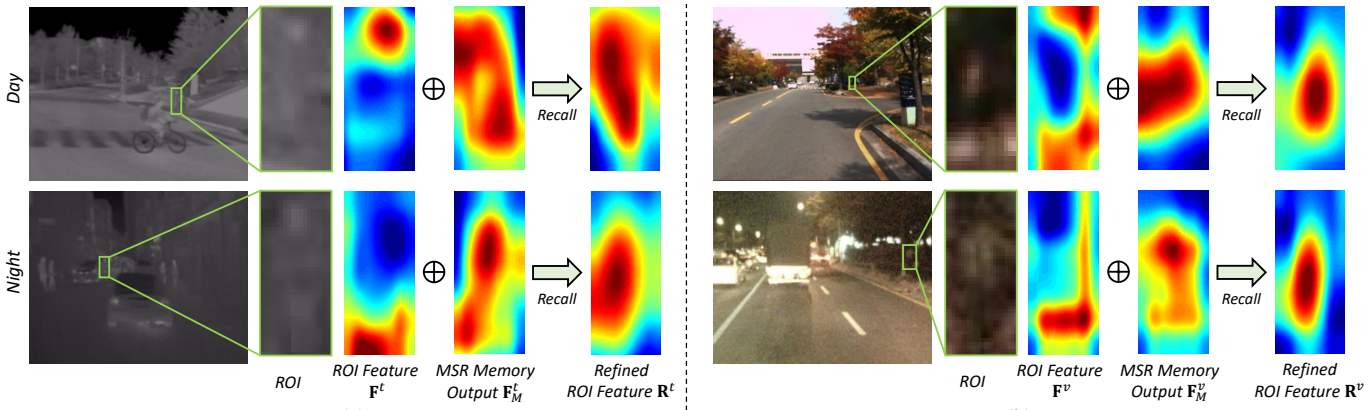
Effect of the Proposed Losses We evaluate the effectiveness of the proposed loss (multisensory-matching contrastive loss and multispectral recalling loss) by changing three types of input modality (*i.e.,* visible, thermal, and '*mixed modality*'). The results are shown in Table 4. When all the proposed losses are considered, the proposed method shows the more improved performance, compared to the baseline single-modal pedestrian detectors.

Slot Number of MSR Memory We also conduct the experiments by modifying the slot number L of the MSR Memory as $L = \{25, 50, 100, 200\}$. Note that, we adopt the '*mixed modality*' in the inference phase. As shown in Table 5, even we change L of the MSR Memory, performances of our method are still high, compared to the single-modal baseline pedestrian detector.

Baseline Detector Extension To see the generalization ability of our method with the MSR Memory, we adopt Feature Pyramid Network (FPN) (Lin et al. 2017) with ResNet-50 and -101 backbone (He et al. 2016). The input modality is the *'mixed-modality'*. The results are shown in Table 6. Even we change the baseline detector, the proposed method with MSR Memory still outperforms the baselines.

Visualization Results

Following (Xu et al. 2020; Kim, Park, and Ro 2020), we conduct feature map visualization on KAIST dataset. Figure



(a) (b)

Figure 3: Feature map visualization ((a): thermal, (b): visible) of the original ROI features $\mathbf{F}^t$, $\mathbf{F}^v$, output features of the MSR Memory $\mathbf{F}_M^t$, $\mathbf{F}_M^v$, and refined ROI features $\mathbf{R}^t$, $\mathbf{R}^v$, following (Xu et al. 2020). Although $\mathbf{F}^t$, $\mathbf{F}^v$ cannot focus the pedestrian regions, $\mathbf{R}^t$, $\mathbf{R}^v$ activate the proper regions (*e.g.,* body) of the pedestrians with the help of the MSR Memory.

3 shows some examples of feature map visualization results. We compare ROI features $\mathbf{F}^t$, $\mathbf{F}^v$, MSR Memory output features $\mathbf{F}_M^t$, $\mathbf{F}_M^v$, and refined ROI features $\mathbf{R}^t$, $\mathbf{R}^v$. Although $\mathbf{F}^t$, $\mathbf{F}^v$ cannot focus the proper regions of the pedestrians, the MSR Memory recalls the information of the multispectral modalities to enhance the visual appearance of the ROI features by focusing on the pedestrian regions (*e.g.,* body). For example, in the thermal modality of the daytime (see the top left of Figure 3), although $\mathbf{F}^t$ mainly looks the pedestrian *head* region, $\mathbf{R}^t$ can focus whole region of the pedestrian through the output feature of the MPR Memory $\mathbf{F}_M^t$.

Discussions

Advantage of the Proposed Method According to Zhang et al. (Zhang et al. 2019), CVC-14 *testset* has a severe misalignment problem, where the observed pedestrian locations of the two modalities are significantly different because of the different FoVs. Therefore, the multispectral pedestrian detectors, which use the two modalities simultaneously, are hard to handle this problem. To mitigate this problem, AR-CNN (Zhang et al. 2019) was introduced, and Kim et al. (Kim, Park, and Ro 2021b) adopted an uncertainty. Thus, as shown in Table 7, they show the improved performance. However, since the proposed method uses a single modality but can act as a multispectral-like detector, so our method is free from the misalignment problem in the inference phase. As a result, our method mostly surpasses the state-of-the-art multispectral pedestrian detectors (Choi et al. 2016; Park, Kim, and Sohn 2018; Zhang et al. 2019; Zhou, Chen, and Cao 2020; Kim, Park, and Ro 2021b; Kim et al. 2021a).

Computational Cost We compare training/inference time and the network parameter number. It is shown in Table 8. Since the proposed method utilizes the visible modality as well as the thermal modality in the training phase, training time is similar to the baseline multispectral pedestrian detector (Faster R-CNN with VGG16). However, in the inference time, our method adopts a single modal input. Therefore, inference time and network parameters are similar to those of the baseline single-modal pedestrian detector.

Method	All	Day	Night
Choi et al.(ICPR'16)	47.30	49.30	43.80
CWF+APF (PR'18)	26.29	28.67	23.48
AR-CNN (ICCV'19)	22.10	24.70	18.10
MLPD (RA-L'21)	21.33	24.18	17.97
MBNet (ECCV'20)	21.10	24.70	13.50
Kim et al. (TCSVT'21)	19.98	**23.52**	12.59
Baseline ($V+T$)	28.54	36.73	21.84
Proposed Method	**19.88**	23.69	**12.35**

Table 7: Detection results (MR) on CVC-14. We compare our method ('*mixed modality*') with the state-of-the-art multispectral pedestrian detectors which adopt the two modal inputs simultaneously.

Method	Training (s) (*per iter*)	Inference (s) (*per image*)	#params
Single Baseline	0.153	0.035	136M
Multispectral Baseline	0.229	0.069	152M
Proposed Method	0.232	0.040	139M

Table 8: The comparisons of training time, inferece time, and number of the parameters.

Conclusion

In this paper, we propose a new perspective for pedestrian detection to effectively handle any single modal input. Here, we design a multisensory-matching contrastive loss to make the pedestrian visual representation of two different modalities similar. Moreover, we devise MSR Memory with multispectral recalling loss to enhance the visual representations of the single modality by recalling the appearance of the multispectral modalities. It enables the proposed pedestrian detector robust in terms of improving feature representation of the pedestrians observed in any single modality. Extensive quantitative and qualitative results verify the effectiveness of the proposed method. We believe that our method can be applied to various real-world applications, such as video surveillance. Also, although this method is developed for pedestrian detection task, it may provide some insight into the studies dealing with various input modalities.

Acknowledgements

This work was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

References

Benenson, R.; Omran, M.; Hosang, J.; and Schiele, B. 2014. Ten years of pedestrian detection, what have we learned? In *ECCV*, 613–627.

Chi, C.; Zhang, S.; Xing, J.; Lei, Z.; Li, S. Z.; and Zou, X. 2020. Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In *AAAI*, 10639–10646.

Choi, H.; Kim, S.; Park, K.; and Sohn, K. 2016. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In *ICPR*, 621–626.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*, 886–893.

González, A.; Fang, Z.; Socarras, Y.; Serrat, J.; Vázquez, D.; Xu, J.; and López, A. M. 2016. Pedestrian detection at day/night time with visible and FIR cameras: A comparison. *Sensors*, 820.

Guo, T.; Huynh, C. P.; and Solh, M. 2019. Domain-adaptive pedestrian detection in thermal images. In *ICIP*, 1660–1664.

Han, B.; Wang, Y.; Yang, Z.; and Gao, X. 2019. Small-scale pedestrian detection based on deep neural network. *TITS*, 21(7): 3046–3055.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 1026–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Herrmann, C.; Ruf, M.; and Beyerer, J. 2018. CNN-based thermal infrared person detection by domain adaptation. In *SPIE*, 1064308.

Hwang, S.; Park, J.; Kim, N.; Choi, Y.; and So Kweon, I. 2015. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, 1037–1045.

Kieu, M.; Bagdanov, A. D.; Bertini, M.; and Del Bimbo, A. 2019. Domain adaptation for privacy-preserving pedestrian detection in thermal imagery. In *ICIAP*, 203–213.

Kieu, M.; Bagdanov, A. D.; Bertini, M.; and Del Bimbo, A. 2020. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In *ECCV*, 546–562.

Kieu, M.; Berlincioni, L.; Galteri, L.; Bertini, M.; Bagdanov, A. D.; and Del Bimbo, A. 2021. Robust pedestrian detection in thermal imagery using synthesized images. In *ICPR*, 8804–8811.

Kim, J.; Kim, H.; Kim, T.; Kim, N.; and Choi, Y. 2021a. MLPD: Multi-Label Pedestrian Detector in Multispectral Domain. *RA-L*, 6(4): 7846–7853.

Kim, J. U.; Park, S.; and Ro, Y. M. 2020. Towards human-like interpretable object detection via spatial relation encoding. In *ICIP*, 3284–3288. IEEE.

Kim, J. U.; Park, S.; and Ro, Y. M. 2021a. Robust Small-Scale Pedestrian Detection With Cued Recall via Memory Learning. In *ICCV*, 3050–3059.

Kim, J. U.; Park, S.; and Ro, Y. M. 2021b. Uncertainty-Guided Cross-Modal Learning for Robust Multispectral Pedestrian Detection. *TCSVT*.

Kim, M.; Hong, J.; Park, S. J.; and Ro, Y. M. 2021b. Multi-modality associative bridging through memory: Speech sound recollected from face video. In *ICCV*, 296–306.

Kim, M.; Joung, S.; Park, K.; Kim, S.; and Sohn, K. 2019. Unpaired Cross-Spectral Pedestrian Detection Via Adversarial Feature Learning. In *ICIP*, 1650–1654.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 1097–1105.

Lee, S.; Kim, H. G.; Choi, D. H.; Kim, H.-I.; and Ro, Y. M. 2021. Video Prediction Recalling Long-term Motion Context via Memory Alignment Learning. In *CVPR*, 3054–3063.

Li, C.; Liang, X.; Lu, Y.; Zhao, N.; and Tang, J. 2019a. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition*, 96: 106977.

Li, C.; Song, D.; Tong, R.; and Tang, M. 2018. Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation. In *BMVC*.

Li, C.; Song, D.; Tong, R.; and Tang, M. 2019b. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognition*, 85: 161–171.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Liu, M.; Zhu, C.; Wang, J.; and Yin, X.-C. 2021. Adaptive Pattern-Parameter Matching for Robust Pedestrian Detection. In *AAAI*, 2154–2162.

Marchetti, F.; Becattini, F.; Seidenari, L.; and Bimbo, A. D. 2020. Mantra: Memory augmented networks for multiple trajectory prediction. In *CVPR*, 7143–7152.

Park, K.; Kim, S.; and Sohn, K. 2018. Unified multi-spectral pedestriaFn detection based on probabilistic fusion networks. *Pattern Recognition*, 143–155.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *NeurIPS Workshop*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 91–99.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, T.; Sun, L.; Xie, D.; Sun, H.; and Pu, S. 2018. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *ECCV*, 536–551.

Song, W.; Li, S.; Chang, T.; Hao, A.; Zhao, Q.; and Qin, H. 2019. Context-interactive CNN for person re-identification. *TIP*, 29: 2860–2874.

Song, X.; Zhao, K.; Chu, W.-S.; Zhang, H.; and Guo, J. 2020. Progressive refinement network for occluded pedestrian detection. In *ECCV*, 32–48.

VS, V.; Gupta, V.; Oza, P.; Sindagi, V. A.; and Patel, V. M. 2021. MeGA-CDA: Memory Guided Attention for Category-Aware Unsupervised Domain Adaptive Object Detection. In *CVPR*, 4516–4526.

Wu, A.; Zheng, W.-S.; Gong, S.; and Lai, J. 2020a. RGB-IR person re-identification by cross-modality similarity preservation. *IJCV*, 128(6): 1765–1785.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-infrared cross-modality person re-identification. In *ICCV*, 5380–5389.

Wu, J.; Zhou, C.; Zhang, Q.; Yang, M.; and Yuan, J. 2020b. Self-mimic learning for small-scale pedestrian detection. In *ACM MM*, 2012–2020.

Xie, J.; Cholakkal, H.; Anwer, R. M.; Khan, F. S.; Pang, Y.; Shao, L.; and Shah, M. 2020a. Count-and similarity-aware R-CNN for pedestrian detection. In *ECCV*, 88–104.

Xie, J.; Pang, Y.; Khan, M. H.; Anwer, R. M.; Khan, F. S.; and Shao, L. 2020b. Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection. *TIP*, 30: 3872–3884.

Xu, C.-D.; Zhao, X.-R.; Jin, X.; and Wei, X.-S. 2020. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, 11724–11733.

Xu, D.; Ouyang, W.; Ricci, E.; Wang, X.; and Sebe, N. 2017. Learning cross-modal deep representations for robust pedestrian detection. In *CVPR*, 5363–5371.

Zhang, L.; Du, G.; Liu, F.; Tu, H.; and Shu, X. 2021. Global-Local Multiple Granularity Learning for Cross-Modality Visible-Infrared Person Reidentification. *TNNLS*.

Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; and Liu, Z. 2019. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *ICCV*, 5127–5137.

Zhang, Z.; Gao, J.; Mao, J.; Liu, Y.; Anguelov, D.; and Li, C. 2020. Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. In *CVPR*, 11346–11355.

Zheng, Y.; Izzat, I. H.; and Ziaee, S. 2019. GFD-SSD: gated fusion double SSD for multispectral pedestrian detection. *arXiv preprint arXiv:1903.06999*.

Zhou, K.; Chen, L.; and Cao, X. 2020. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *ECCV*, 787–803.

Zhu, C.; and Peng, Y. 2015. A boosted multi-task model for pedestrian detection with occlusion handling. *TIP*, 24(12): 5619–5629.

Zhu, L.; and Yang, Y. 2020. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *CVPR*, 4344–4353.