The following publication T. Liu, K. -M. Lam, R. Zhao and G. Qiu, "Deep Cross-Modal Representation Learning and Distillation for Illumination-Invariant Pedestrian Detection," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 1, pp. 315-329, Jan. 2022 is available at https://doi.org/10.1109/TCSVT.2021.3060162.

Deep Cross-modal Representation Learning and Distillation for Illumination-invariant Pedestrian Detection

Tianshan Liu, Kin-Man Lam, *Senior Member, IEEE*, Rui Zhao, and Guoping Qiu

***Abstract*—Integrating multispectral data has been demonstrated to be an effective solution for illumination-invariant pedestrian detection, in particular, RGB and thermal images can provide complementary information to handle light variations. However, most of the current multispectral detectors fuse the multimodal features by simple concatenation, without discovering their latent relationships. In this paper, we propose a cross-modal feature learning (CFL) module, based on a split-and-aggregation strategy, to explicitly explore both the shared and modality-specific representations between paired RGB and thermal images. We insert the proposed CFL module into multiple layers of a two-branch-based pedestrian detection network, to learn the cross-modal representations in diverse semantic levels. By introducing a segmentation-based auxiliary task, the multimodal network is trained end-to-end by jointly optimizing a multi-task loss. On the other hand, to alleviate the reliance of existing multispectral pedestrian detectors on thermal images, we propose a knowledge distillation framework to train a student detector, which only receives RGB images as input and distills the cross-modal representations guided by a well-trained multimodal teacher detector. In order to facilitate the cross-modal knowledge distillation, we design different distillation loss functions for the feature, detection and segmentation levels. Experimental results on the public KAIST multispectral pedestrian benchmark validate that the proposed cross-modal representation learning and distillation method achieves robust performance.**

***Index Terms*—Illumination-invariant pedestrian detection, multispectral fusion, knowledge distillation, cross-modal representation.**

I. Introduction

As one of the crucial research topics in computer vision, pedestrian detection has widespread human-centric applications [1]–[3], such as video surveillance, autonomous driving, human behaviour analysis, etc. Recently, motivated by the success of deep-learning-based methods in object detection [4], [5], remarkable improvements have been made on the performance of pedestrian detectors. However, most of the existing pedestrian detection methods are based on the assumption that the visible images are captured under good illumination conditions. This may restrict them from being deployed in real-world scenarios, with a wide range of illumination variations [6], [7]. Especially, images with weak illumination usually contain low contrast and are of low resolution, which makes the pedestrian detection task difficult even for human beings.

To overcome this limitation, various attempts have been made to explore multispectral data [8], [9], i.e., paired visible (RGB)-thermal images, for illumination-invariant pedestrian detection. The intuition is that thermal data is insensitive to illumination variations and can provide additional information of the targets complementary with RGB data [10], [11]. However, the majority of the current multispectral pedestrian detection methods concentrate on exploring different fusion stages, and simply concatenate the features extracted from RGB and thermal images, without discovering their latent relationships. Considering that paired RGB-thermal images are highly correlated in terms of targets and scenes, more compact yet discriminative representations can be extracted by learning the latent semantic relations between the multispectral data. Therefore, different from the common practice of fusing multispectral features in a holistic manner, we propose a novel cross-modal feature learning (CFL) module, to explicitly explore both the shared features and modality-specific features between RGB and thermal image pairs. The shared features depict the common cues between the two modalities, while the modality-specific features capture unique components of each modality.

Although leveraging multispectral data can alleviate the ill-effects of poor illumination conditions on RGB images, it is impractical to collect thermal images in some real-world applications owing to the limitations of hardware [12]. This poses a new challenge, that of how to train a robust pedestrian detector using multispectral data, while only RGB images are available in the test/inference phase. Inspired by the theory of knowledge distillation explored in numerous multimodal vision tasks [13]–[15], we propose to leverage a deep-learning-based distillation framework to train a student network, which distills cross-modal representations guided by a well-trained multimodal teacher network. In other words, the student network only receives RGB images as input, and is trained to produce the discriminative cross-modal representations, to achieve illumination-invariant pedestrian detection. The teacher model, trained using extra thermal images, can provide privileged information to facilitate the learning of the student model. To distill sufficient knowledge from the

Tianshan Liu, Kin-Man Lam and Rui Zhao are with Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: tianshan.liu@connect.polyu.hk; enkmlam@polyu.edu.hk; rick10.zhao@connect.polyu.hk).

Guoping Qiu is with College of Electronic and Information Engineering, Shenzhen Institute of Artificial Intelligence and Robotics for Scosity, and Guangdong Key Lab for Intelligent Information Processing, Shenzhen University, China, and also with the School of Computer Science, The University of Nottingham, Nottingham, U.K. (guoping.qiu@nottingham.ac.uk).

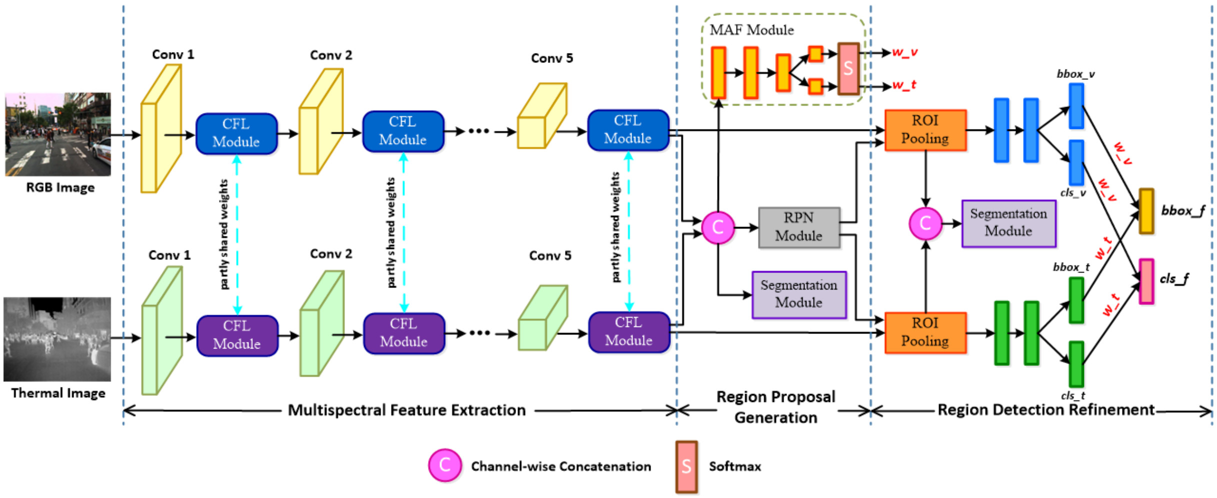


Fig. 1. The overall architecture of the proposed deep cross-modal representation learning-based pedestrian detection network (DCRL-PDN).

multimodal teacher network, we design different levels of distillation loss terms.

In this paper, we first propose a deep cross-modal representation learning-based pedestrian detection network (DCRL-PDN) as the teacher model. The proposed DCRL-PDN is a two-branch-based framework, taking paired RGB-thermal images as input. Each branch is developed, based on a two-stage detector, i.e., Faster R-CNN [5]. We boost the performance of the proposed illumination-invariant pedestrian detector from three aspects. Firstly, to explicitly exploit the shared and modality-specific components between multispectral data, we propose a cross-modal feature learning (CFL) module based on a split-and-aggregation strategy. Our CFL module can be inserted into the backbone CNN in a plug-and-play way, to facilitate information interactions between the two branches. Secondly, motivated by the effectiveness of self-supervision learning, in addition to the detection task, we introduce a segmentation-based auxiliary task by leveraging weakly supervised box-based pedestrian masks. Thirdly, to adaptively assign different importance weights for each modality according to the inputs, we employ a modality attention-based fusion (MAF) strategy to combine the detection results of the two sub-branches. The overall network architecture of our proposed DCRL-PDN is illustrated in Fig. 1.

On the other hand, we leverage a generalized distillation framework, by training a deep cross-modal representation distillation-based pedestrian detection network (DCRD-PDN), as the student model. The proposed DCRD-PDN is fed with RGB images only, and distills cross-modal representations taught by the multimodal teacher network (DCRL-PDN). As illustrated in Fig. 2, the weights of the multimodal teacher detector (DCRL-PDN) are frozen for the learning of the student detector (DCRD-PDN). To distill abundant cross-modal knowledge from the well-trained teacher model, we employ different distillation loss terms at multiple levels, including the feature, detection and segmentation perspectives. Firstly, we employ a Euclidean-distance-based metric loss to explicitly distill cross-modal features, from the backbone CNN of the teacher model. Secondly, considering that pedestrian detection is multi-tasked, we utilize a classification-level distillation loss and a conditional regression distillation loss for two sub-tasks. Thirdly, we propose a pixel-wise segmentation distillation loss, to improve the learning capability of the student network.

The main contributions of this paper can be summarized as follows. First, based on a split-and-aggregation strategy, we propose a cross-modal feature learning (CFL) module, which aims to explicitly discover the shared structure and modality-specific cues between RGB and thermal images. The proposed CFL module is deployed into different layers of the deep multimodal network for illumination-invariant pedestrian detection. Second, we propose a knowledge distillation framework to train a student network, which only receives RGB images as input and distills cross-modal representations guided by the multimodal teacher network. We design multiple losses to conduct cross-modal knowledge distillation at the feature, detection and segmentation levels. Third, extensive experimental results and ablation analysis on the public KAIST multispectral pedestrian data set validate the effectiveness of the proposed method.

II. Related Work

A. Visible Pedestrian Detection

As a canonical sub-task of object detection, pedestrian detection has been intensively studied by the computer vision community, because of its widespread real-world applications. Research works on pedestrian detection have experienced an evolution from handcrafted feature-based methods to deep-learning framework-based methods. The Integral Channel Features (ICF) [16] is a popular pedestrian detector, which combines channel feature pyramids with cascade classifiers. ICF has been a basic descriptor for several variants, such as Aggregated Channel Features (ACF) [17] and Filtered Channel Features (FCF) [18]. With the popularity of CNN in numerous vision tasks, recent pedestrian detectors are based on deep-learning architectures. Since the region-proposal-based detection models [5], [19] have achieved promising performance in

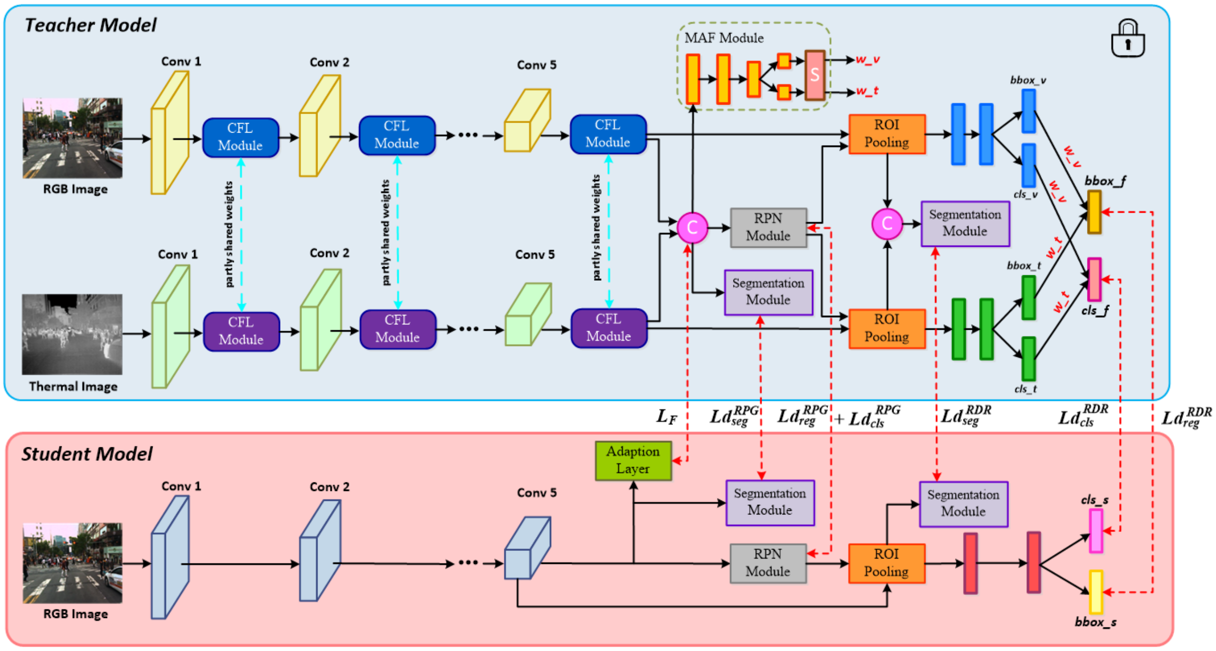


Fig. 2. Illustration of the proposed deep cross-modal representation distillation framework.

object detection, more and more pedestrian detection methods are being developed based on two-stage detectors, e.g., Faster R-CNN [5]. Cai et al. [20] proposed a multi-scale CNN (MS-CNN) by integrating a feature pyramid into the Faster-RCNN. Li et al. [21] developed a Scale-Aware Fast R-CNN (SAF-RCNN), by introducing multiple sub-networks to detect pedestrians at different scales. Instead of using RoIPooling as in typical Faster-RCNN, Zhang et al. [22] extracted convolutional features of the candidates generated by the Region Proposal Network (RPN) [5], and then employed a boosted forest (BF) to mine hard negative examples. Brazil et al. [23] cascaded the RPN and a binary classification network (BCN), with an additional segmentation loss, to guide the learning of the detector. Cai et al. [24] proposed a cascade R-CNN, which consists of multiple detectors, trained with various IoU thresholds, to eliminate false positives in a progressive manner. On the other hand, there are also pedestrian detection methods which explore using a single-stage pipeline to save computational time. Liu et al. [25] proposed an Asymptotic Localization Fitting (ALF) module to evolve the anchor boxes of SSD [26] sequentially to achieve more accurate detection results. Lin et al. [27] integrated fine-grained attention masks into convolutional features, to construct graininess-aware representations for pedestrian detection. Recently, Liu et al. [28] proposed an anchor-free method, which converts the pedestrian detection task to the center and scale prediction task, through convolutions on higher-level semantic features.

Our work is also built based on a two-stage detector, i.e., Faster R-CNN, owing to its excellent detection accuracy. However, different from the above-mentioned visible pedestrian detection methods that mainly tackle the issues of scale [29] and occlusion [30], we focus on leveraging multispectral data to train a robust model to deal with light variations, for illumination-invariant pedestrian detection.

B. Multispectral Pedestrian Detection

Multispectral sensors capture paired RGB-thermal images to provide complementary information about the target pedestrians. Effective fusion of these two modalities can achieve robust detection results. Hwang et al. [8] first collected a large-scale multispectral pedestrian data set, and extended the classical ACF detector to a multispectral version (ACF+T+THOG) by incorporating intensity and HOG features of the thermal channel. Liu et al. [31] integrated Faster R-CNN for multispectral pedestrian detection, and investigated four fusion strategies to combine RGB and thermal branches at different stages. Fusion RPN+BDT [32] was proposed by integrating a multispectral RPN with a boosted decision-tree classifier, to reduce false positives. Li et al. [33] cascaded a multispectral proposal network with a multispectral classification network, and jointly optimized the detection and semantic segmentations losses in an end-to-end manner. Guan et al. [11] developed a built-in day-sub-network and night-sub-network, and applied both of them to process RGB and thermal images. Motivated by the channel attention in squeeze-and-excitation networks (SENets), Zhang et al. [10] proposed a cross-modality interactive attention module to fuse the RGB and thermal features for multispectral pedestrian detection. Li et al. [6] trained an illumination-aware network using coarse day/night labels to obtain two weights, which were further used to combine the detection results from the RGB and thermal sub-networks. Park et al. [34] designed a three-branch-based network, and proposed the channel weighting fusion (CWF) and accumulated probability fusion (APF) layers to dynamically fuse different information flows. To alleviate the problem of position shift, Zhang et al. [7] proposed an aligned region CNN (AR-CNN) to adaptively align the region features between RGB and thermal modalities.

Most of the above-mentioned multispectral pedestrian detection methods simply concatenate features extracted from RGB and thermal modalities, without learning their latent relationships. This restricts the detection performance of these methods. In this paper, we propose a cross-modal feature learning (CFL) module, which specifically explore both shared and modality-specific features between different modalities by a split-and-aggregation strategy. Moreover, the above-mentioned methods are based on the assumption that RGB and thermal images are available in both training and testing stages. However, collection of the thermal data is impractical in some real-world applications. Therefore, we also exploit a generalized distillation framework to train a student network, which distills cross-modal representations from the multi-modal teacher network. The proposed model only receives RGB images as input, but can achieve illumination-invariant pedestrian detection.

C. Cross-modal Representation Learning

Cross-modal representation learning approaches aim to discover discriminative latent connections between different modalities for facilitating the downstream task. Kang et al. [35] proposed to jointly learn the basis matrices of different modalities for handling unpaired data, in cross-modal multimedia retrieval. To fully exploit the rich patterns across the multimodal data, Zhang et al. [36] proposed a cross-modal knowledge transition scheme based on the generative adversarial network, and further built a cross-modal feature fusion network for brain tumor segmentation. To model the urban dynamics from massive geo-tagged social media (GTSM) data, a cross-modal feature learning method, named CrossMap [37], was presented. After detecting the spatio-temporal hotspots, both reconstruction-based and graph-based strategies are applied for projecting all spatial, temporal, and textual components into a shared embedding space. Zhang et al. [38] deployed a series of multi-scale, multi-modal, and multi-level feature fusion modules into a deep network, to achieve robust RGB-T saliency detection.

However, most of these methods only concentrate on discovering shared feature space among different modalities without considering modality-specific structures. In contrast, we propose a cross-modal feature learning (CFL) module, to explicitly exploit both shared and modality-specific features between different modalities, based on channel split and aggregation strategies. The proposed CFL module is flexible and can be deployed into multiple layers of backbone network, to learn the cross-modal representations in diverse semantic levels.

D. Knowledge Distillation

Knowledge distillation [39] is a powerful tool for transferring useful information between different domains, such as high-resolution and low-resolution images, RGB and depth data, etc. Our work is motivated by the knowledge distillation framework explored in various multimodal vision tasks. Hoffman et al. [13] proposed a distillation framework to train a hallucination network, which distills depth features from RGB images for general object detection. Luo et al. [14] presented a graph-based approach to distill knowledge across abundant modalities (RGB, depth, optical flow, skeleton joints) in the training stage, but tested it only using single modality for action detection and recognition. Shi et al. [40] leveraged skeleton data as privileged information, to facilitate the learning of RNN-based models, for action recognition from depth sequences. Garcia et al. [41] trained a hallucination stream, by designing a multi-stage training paradigm and combining a Euclidean-distance metric loss with a generalized distillation loss, for action recognition. The authors further utilized an adversarial strategy [15], instead of the distillation loss, to alleviate the demand of tuning hyper-parameters. Different from these distillation approaches proposed for RGB-D vision tasks, our distillation method is proposed to discover the relationship between RGB and thermal images. In addition, most of the above-mentioned methods train a separate stream to produce the features, mimicking the missing or unavailable modality. In contrast, we train a single network, which produces more compact cross-modal representations for pedestrian detection.

Recently, some research on pedestrian detection has also leveraged knowledge distillation frameworks from different aspects. Shen et al. [42] trained a smaller network guided by a large network, by performing knowledge distillation at both high-dimensional hint layer and prediction layer. The authors also integrated handcrafted features, i.e., ACF, to boost the performance of pedestrian detection. Chen et al. [43] proposed several improved loss functions to distill information from the feature, classification and regression perspectives. Wang et al. [44] proposed a fine-grained feature imitation method for object detection, which forces the student network to concentrate more on distilling knowledge from the teacher network in the near object anchor regions. However, methods [42]–[44] leverage a knowledge distillation framework as a compression tool, which aims to learn a compact yet effective object detection network. In our work, we mainly employ knowledge distillation to transfer cross-modal information from the teacher model to the student model. Xu et al. [45] proposed a region reconstruction network to transfer cross-modal representations between RGB and thermal data, which improves the robustness of pedestrian detectors against bad illumination conditions. However, the method only distills knowledge through an indirect reconstruction task, without considering distillation at high-level detection tasks. Kruthiventi et al. [12] proposed to distill knowledge from a multi-modal teacher model, by only using a weighted L1-hint loss in an intermediate layer. To distill sufficient cross-modal knowledge from the teacher model, our proposed method employs different distillation losses in multiple levels, including the feature, detection and segmentation perspectives.

III. Deep Cross-modal Representation Learning

A. Overall Network Architecture

The overall architecture of the proposed deep cross-modal representation learning-based pedestrian detection network (DCRL-PDN) is illustrated in Fig. 1. The proposed DCRL-PDN framework consists of two sub-branches, which take RGB images and thermal images as input, respectively. Each branch network is based on a Faster R-CNN, involving two stages, i.e., region proposal generation (RPG) and region detection refinement (RDR). We choose VGG-16 [46] as the backbone CNN to extract features for each branch. To improve the representation learning capacity and facilitate information exchange between the two branches, we introduce three modules into the classical Faster R-CNN, including cross-modal feature learning, segmentation auxiliary task, and modality attention-based fusion. The proposed cross-modal feature learning (CFL) module aims to discover the common structure and modality-specific characteristics between the RGB and thermal inputs, in an explicit manner. We insert the proposed CFL module after each convolutional layer from conv1 to conv5, to learn the shared and modality-specific components in different semantic levels. The learned cross-modal representations can serve as a strong basis for the subsequent detection tasks. In addition, to ease the training of the final detection task, we introduce a segmentation-based auxiliary task in both the RPG and RDR stages, by leveraging weakly supervised pedestrian masks. Furthermore, we employ a modality attention-based fusion (MAF) strategy to compute the weights, for combining the detection results of the two sub-branches. The details about these three modules are presented in the following subsections.

The overall objective function, L_T, is formulated as follows:

$$\begin{aligned} L_{RPG} &= L_{cls}^{RPG} + \lambda L_{reg}^{RPG} + \eta L_{seg}^{RPG}, \\ L_{RDR} &= L_{cls}^{RDR} + \lambda L_{reg}^{RDR} + \eta L_{seg}^{RDR}, \\ L_T &= L_{RPG} + L_{RDR}, \end{aligned} \tag{1}$$

where L_{cls} denotes the classification loss function based on the cross-entropy loss, L_{reg} is the bounding-box regression loss function based on the smoothed L1 loss, and L_{seg} is the segmentation loss, which will be described in Section III-C. The parameters λ and η are used to balance the importance between the different loss terms. Both of them are set to be 1, throughout all the experiments.

B. Cross-modal Feature Learning Module

Considering that the RGB and thermal images are highly related in the scenes involving pedestrians, there exist latent semantic relationships between these two modalities. In the deep-learning domain, a crucial insight is that different channels of a feature map represent the diverse patterns learned by the network. For learning discriminative cross-modal representations, some channels are expected to depict the common structures between the RGB and thermal modalities, while the other channels account for capturing the unique cues of each modality. Therefore, we propose a cross-modal feature learning (CFL) module based on a split-and-aggregation strategy, which explicitly explores the shared structures and modality-specific cues between RGB and thermal data. As shown in Fig. 3, for each modality m ($m = 1, 2$, for RGB and thermal images, respectively), given a convolutional feature tensor $\mathbf{x}^m \in \mathbb{R}^{C\times W\times H}$, extracted from the VGG-16 backbone network, where C denotes the number of channels, and W and H are the width and height, respectively. We first split the feature tensor channels into two parts, by a ratio of α, one for learning the shared structures among multispectral data, with the other for learning the modality-specific details. Then, the subsequent convolution operations can be formulated as follows:

$$\begin{aligned} \begin{bmatrix} \mathbf{o}_1^{\hat{m}} \\ \mathbf{o}_2^{\hat{m}} \\ \vdots \\ \mathbf{o}_C^{\hat{m}} \end{bmatrix} &= \begin{bmatrix} \mathbf{W}_{1,1} & \cdots & \mathbf{W}_{1,\alpha C} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{C,1} & \cdots & \mathbf{W}_{C,\alpha C} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^m \\ \vdots \\ \mathbf{x}_{\alpha C}^m \end{bmatrix} \\ &+ \begin{bmatrix} \mathbf{U}_{1,\alpha C+1}^m & \cdots & \mathbf{U}_{1,C}^m \\ \vdots & \ddots & \vdots \\ \mathbf{U}_{C,\alpha C+1}^m & \cdots & \mathbf{U}_{C,C}^m \end{bmatrix} \begin{bmatrix} \mathbf{x}_{\alpha C+1}^m \\ \vdots \\ \mathbf{x}_C^m \end{bmatrix}, \end{aligned} \tag{2}$$

where $\mathbf{o}^{\hat{m}} \in \mathbb{R}^{C\times W\times H}$ represents the output feature tensor, $\mathbf{W}_{i,j}$ denotes the learnable kernels, for learning the shared structure using αC channels, and $\mathbf{U}_{i,j}^m$ represents the learnable parameters of modality m, applied for learning the modality-specific representation with $(1-\alpha)\,C$ channels.

To explore the common structures between the RGB and thermal modalities, and maintain computational efficiency, we employ point-wise convolution, i.e., 1×1 kernels, with the parameters $\mathbf{W}_{i,j}$. As illustrated in Fig. 3, we enforce the parameters $\mathbf{W}_{i,j}$ to be shared between multispectral data, thus, leading to discovering the common patterns of pedestrians across different modalities. In addition, we further divide the modality-specific representation part into several groups. Each group may correspond to an intrinsic characteristic of each modality, such as textures in RGB images, heat signature in thermal data, etc. Then, we utilize group-wise convolution to process these channel groups. Since the group-wise convolution is based on a sparse block-diagonal kernel, each block only accounts for a specific group [47]. To ensure the interactions across all the channels in the modality-specific part, we employ another point-wise convolution to avoid information loss. We apply both group-wise and point-wise convolution on the modality-specific channels, in parallel. Then, these two convolutional features are fused by element-wise summation. The convolution operation on the modality-specific part is formulated as follows:

$$\begin{aligned} \begin{bmatrix} \mathbf{U}_{1,\alpha C+1}^m & \cdots & \mathbf{U}_{1,C}^m \\ \vdots & \ddots & \vdots \\ \mathbf{U}_{C,\alpha C+1}^m & \cdots & \mathbf{U}_{C,C}^m \end{bmatrix} \begin{bmatrix} \mathbf{x}_{\alpha C+1}^m \\ \vdots \\ \mathbf{x}_C^m \end{bmatrix} &= \begin{bmatrix} \mathbf{P}_1^m & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{P}_G^m \end{bmatrix} \begin{bmatrix} \mathbf{z}_1^m \\ \vdots \\ \mathbf{z}_G^m \end{bmatrix} \\ + \begin{bmatrix} \mathbf{V}_{1,\alpha C+1}^m & \cdots & \mathbf{V}_{1,C}^m \\ \vdots & \ddots & \vdots \\ \mathbf{V}_{C,\alpha C+1}^m & \cdots & \mathbf{V}_{C,C}^m \end{bmatrix} \begin{bmatrix} \mathbf{x}_{\alpha C+1}^m \\ \vdots \\ \mathbf{x}_C^m \end{bmatrix}&, \end{aligned} \tag{3}$$

where the modality-specific part, learned with $(1-\alpha)\,C$ channels, is partitioned into G groups, and each group $\mathbf{z}_g^m$, $g = 1, 2, ..., G$, consists of $(1-\alpha)\,C/G$ channels. $\mathbf{P}_g^m$ represents the group-wise convolutional kernels for the g-th group.

After generating both the shared representation $\mathbf{S}^m \in \mathbb{R}^{C\times W\times H}$ and modality-specific representation $\mathbf{R}^m \in \mathbb{R}^{C\times W\times H}$, we employ an attention-based fusion strategy to combine these two types of features, without introducing extra parameters. Specifically, as shown in Fig. 3, global average pooling (GAP) is applied to $\mathbf{S}^m$ and $\mathbf{R}^m$ along the spatial

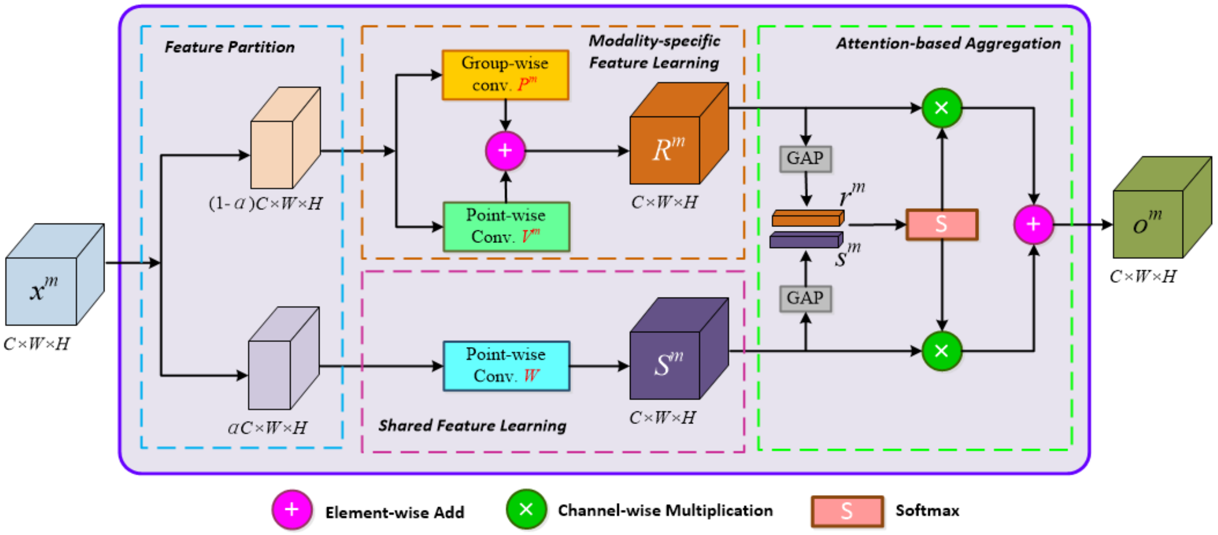


Fig. 3. Schematic diagram of the proposed cross-modal feature learning (CFL) module.

dimension, to obtain the channel-wise statistics $\mathbf{s}^m \in \mathbb{R}^C$ and $\mathbf{r}^m \in \mathbb{R}^C$, respectively. Then, we utilize a self-attention strategy based on softmax operation, to generate the fusion-weight vectors $\beta^m \in \mathbb{R}^C$ and $\gamma^m \in \mathbb{R}^C$, as follows:

$$\beta_c^m = \frac{\exp(s_c^m)}{\exp(s_c^m) + \exp(r_c^m)}, \tag{4}$$

$$\gamma_c^m = 1 - \beta_c^m, \tag{5}$$

where β_c^m and γ_c^m, $c = 1, 2, ..., C$, denote the c-th element of the weight vectors β^m and γ^m, respectively. The final output of the CFL module is computed by fusing the shared feature $\mathbf{S}^m$ and the modality-specific feature $\mathbf{R}^m$, using the weight vectors β^m and γ^m, as follows:

$$\mathbf{o}_c^m = \beta_c^m \mathbf{S}_c^m + \gamma_c^m \mathbf{R}_c^m, \tag{6}$$

By Eq. (6), Eq. (2) becomes as follows:

$$\begin{aligned}
\begin{bmatrix} \mathbf{o}_1^m \\ \mathbf{o}_2^m \\ \vdots \\ \mathbf{o}_C^m \end{bmatrix} &= \beta^m \odot \begin{bmatrix} \mathbf{W}_{1,1} & \cdots & \mathbf{W}_{1,\alpha C} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{C,1} & \cdots & \mathbf{W}_{C,\alpha C} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^m \\ \vdots \\ \mathbf{x}_{\alpha C}^m \end{bmatrix} \\
&+ \gamma^m \odot \begin{bmatrix} \mathbf{P}_1^m & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{P}_G^m \end{bmatrix} \begin{bmatrix} \mathbf{z}_1^m \\ \vdots \\ \mathbf{z}_G^m \end{bmatrix} \\
&+ \gamma^m \odot \begin{bmatrix} \mathbf{V}_{1,\alpha C+1}^m & \cdots & \mathbf{V}_{1,C}^m \\ \vdots & \ddots & \vdots \\ \mathbf{V}_{C,\alpha C+1}^m & \cdots & \mathbf{V}_{C,C}^m \end{bmatrix} \begin{bmatrix} \mathbf{x}_{\alpha C+1}^m \\ \vdots \\ \mathbf{x}_C^m \end{bmatrix},
\end{aligned} \tag{7}$$

where $\odot$ denotes the channel-wise multiplication. These are also the detailed operations of the proposed CFL module.

C. Segmentation-based Auxiliary Task

Auxiliary task has had demonstrated success in a number of vision-based topics [48], [49], because it can ease the training of the downstream task and improve the learning capacity of the trunk network. Motivated by the works in [23] and [6], we introduce a segmentation-based auxiliary task into both the RPG and RDR stages, by using box-based pedestrian masks. Since the goal of the auxiliary task is to facilitate the feature-learning efficiency of the trunk network, rather than to achieve high segmentation performance, we implement the segmentation module as a shallow layer, with 1×1 convolutional kernels. The segmentation module outputs the masks indicating the likelihood of pixels belonging to pedestrian or background segments. As shown in Fig. 1, by concatenating the cross-modal representations of each modality along the channel dimension, the segmentation module is attached to predict the pedestrian masks. Each ground-truth pedestrian mask $\mathbf{q} \in \mathbb{R}^{W\times H}$ is generated by labelling all the regions in a bounding box as foreground, i.e., $\mathbf{q}_{x,y} = 1$, with the other regions as background, i.e., $\mathbf{q}_{x,y} = 0$, where W and H are the width and height, respectively, of the mask. The segmentation loss in the RPG stage is defined as follows:

$$L_{seg}^{RPG} = \frac{1}{H \times W} \sum_{x,y} l_s\left(\mathbf{q}_{x,y}, \mathbf{q}_{x,y}^*\right), \tag{8}$$

where $l_s(\cdot)$ is the cross-entropy loss. $\mathbf{q}$ and $\mathbf{q}^*$ represent the predicted and ground-truth pedestrian masks, respectively. Similarly, in the RDR stage, the segmentation loss is formulated as follows:

$$L_{seg}^{RDR} = \frac{1}{H \times W \times J} \sum_{x,y,j} l_s\left(\mathbf{q}_{x,y,j}, \mathbf{q}_{x,y,j}^*\right), \tag{9}$$

where the subscript j denotes the j-th region of interest (ROI), and J is the total number of ROIs.

D. Modality Attention-based Fusion

Since each sub-branch outputs the modality-specific detection results, involving confidence score and bounding-box regression offsets, it is vital to adopt an effective fusion strategy to obtain the final detection results. We observe that the importance of each modality may vary in different scenarios. In good illumination conditions, e.g., during daytime, the colour and texture patterns in RGB images may provide more useful cues for pedestrian detection. In poor illumination conditions, e.g., during nighttime, the pedestrian silhouettes in thermal images are more reliable. Since the proposed CFL module explores both the shared and modality-specific components to provide discriminative representations, it is reasonable to adaptively generate weights for each modality based on the features learned by the CFL module. Therefore, we employ a modality attention-based fusion module, to compute the relative importance weight for each modality, according to the input image pairs. Given the features $\mathbf{o}^m$ output from the CFL module after conv5 layer, in each branch, the attention weights are computed as follows:

$$\mathbf{w} = D_{\theta^{at}} \left(\oplus_{m=1}^{M} fl(\mathbf{o}^m) \right), \tag{10}$$

$$\mu_m = \frac{\exp(w_m)}{\sum_{k=1}^{M} \exp(w_k)}, \tag{11}$$

where $\oplus$ is the concatenation operator, $fl(\cdot)$ denotes the flatten operation, D represents a multi-layer perceptron (MLP) with the learnable parameter θ^{at}, and w_m denotes the m-th element of vector $\mathbf{w}$. The final detection results are produced by a linear combination, based on the modality-attention weights, as follows:

$$d_{final} = \sum_{m=1}^{M} \mu_m \cdot d_m, \tag{12}$$

$$t_{final} = \sum_{m=1}^{M} \mu_m \cdot t_m, \tag{13}$$

where d_m and t_m are the classification score and bounding-box regression offsets from each sub-branch, respectively. Since the modality-attention-weighted results, defined in Eq. (12) and Eq. (13), are differentiable with respect to the weights and sub-branch results, the overall network is trainable end-to-end.

IV. Cross-modal Representation Distillation

A. Overview

Since capturing multispectral data may be difficult in some practical applications due to limitations of hardware, we propose to solve the problem by leveraging a knowledge distillation framework to train a student network, which receives only RGB images as input and distills cross-modal representations guided by the multimodal teacher network. Considering that visual patterns in RGB images are less discriminative in regions of low visibility for pedestrian detection, a well-trained teacher model using extra thermal data can provide privileged information for student-model learning. As illustrated in Fig. 2, the DCRL-PDN model is chosen as the teacher network. The student network is also built based on a Faster R-CNN, involving two stages, i.e., region proposal generation (RPG) and region detection refinement (RDR). To ensure the basic learning capacity, we also introduce a segmentation-based auxiliary task into these two stages. The training paradigm for knowledge distillation consists of two steps. In the first step, the teacher network is trained leveraging RGB and thermal image pairs, as depicted in Section III. In the second step, the teacher network is frozen to provide a stable target for the student network to learn. We only feed RGB images into the student model, and train it to distill cross-modal representation for illumination-invariant pedestrian detection. The cross-modal knowledge distillation is conducted at multiple levels, including the feature, detection and segmentation perspectives. We introduce these three aspects in detail, in the following subsections.

The overall objective function, L_{TS}, for the proposed teacher-student learning model is defined as follows:

$$\begin{aligned} Ld_{RPG} &= Ld_{cls}^{RPG} + \lambda Ld_{reg}^{RPG} + \eta Ld_{seg}^{RPG}, \\ Ld_{RDR} &= Ld_{cls}^{RDR} + \lambda Ld_{reg}^{RDR} + \eta Ld_{seg}^{RDR}, \\ L_{TS} &= Ld_{RPG} + Ld_{RDR} + \rho L_F, \end{aligned} \tag{14}$$

where L_F denotes the feature-level distillation loss function that forces the student network to produce cross-modal representation as the teacher network. Ld_{cls} and Ld_{reg} represent the classification loss and bounding-box regression loss, respectively. Both of them belong to the detection-level distillation. Ld_{seg} is the segmentation-level distillation loss function. All these loss functions are described in the subsections that follow. The parameters λ, η and ρ balance the weights between the different distillation loss terms. We set them to be 1, 1, and 0.5, respectively, throughout all the experiments.

B. Feature-level Distillation

To explicitly distill cross-modal representation from the multimodal teacher network, we employ a feature-level distillation loss function, based on Euclidean distance, as follows:

$$L_F = \left\| A(\mathbf{E}^s) - \mathbf{E}^t \right\|_2^2, \tag{15}$$

where $\mathbf{E}^t \in \mathbb{R}^{2C \times W \times H}$ denotes the multispectral feature maps in the teacher network, generated by concatenating the cross-modal representations of each modality after the conv5 layer, as $\mathbf{E}^t = \oplus_{m=1}^{2} \mathbf{o}^m$. $\mathbf{E}^s \in \mathbb{R}^{C \times W \times H}$ represents the feature map output from the conv5 layer of the student network. To match the number of channels between the feature maps in the teacher and the student networks, an adaption layer $A(\cdot)$, based on 1×1 convolution, is adopted to map the student features to the teacher feature space. Thus, the output size of the adaption layer is $2C$. Minimizing L_F implies distilling cross-modal representations at the intermediate layers, guided by the well-trained multimodal teacher network.

C. Detection-level Distillation

Since the pedestrian detection task consists of two sub-tasks, i.e., classification and bounding-box regression, we design the detection-level distillation loss from these two aspects. Motivated by the classical knowledge distillation framework, we employ a generalized distillation loss [50] for the classification task by simultaneously considering hard and soft labels, as follows:

$$Ld_{cls} = \zeta g \left(y^h, \varphi(d^s) \right) + (1 - \zeta) \, g \left(y^s, \varphi(d^s) \right), \tag{16}$$

where φ is a softmax function and $g(\cdot)$ denotes a cross-entropy loss. d^s is the classification score of the student network. The hyper-parameter $\zeta \in [0,1]$ balances the contribution of the hard labels (ground truth) y^h and the soft labels y^s in the classification distillation loss. Given the classification score of the teacher network as d^t, the soft labels y^s are computed by $\varphi\left(d^t/T\right)$, where T is the temperature parameter to soften the prediction vector. The soft labels involve discriminative information learned from multispectral data by the teacher network. Integrating the soft labels into the distillation loss can force the student network to distill such privileged information.

The bounding-box regression task aims to calibrate the position and size of the proposals. Considering that the regression results are real numbers and unbounded, the inaccurate regression of the teacher model may provide wrong guidance to the student model. Motivated by the work in [43], we employ a conditional regression distillation loss, which utilizes the regression results of the teacher model as a bound, instead of using a soft target to guide the regression learning directly. The specific bounding-box regression distillation loss is formulated as follows:

$$L_b\left(t_s, t_t, y\right) = \begin{cases} \left\|t_s - y\right\|_2^2, & \text{if } \left\|t_s - y\right\|_2^2 + \delta > \left\|t_t - y\right\|_2^2 \\ 0, & \text{otherwise} \end{cases},$$

$$Ld_{reg} = L_{sL1}\left(t_s, y_{reg}\right) + \varepsilon L_b\left(t_s, t_t, y_{reg}\right), \qquad (17)$$

where y_{reg} is the regression ground truth, and δ denotes a margin. t_s and t_t represent the regression output of the student network and teacher network, respectively. ε is a weight parameter. $L_{sL1}\left(\cdot\right)$ denotes a smoothed L1 loss function. The regression distillation loss, defined in Eq. (17), implies that when the performance of the student model is lower than the teacher model by a certain margin, additional penalties will be introduced to regularize the learning of the student network.

D. Segmentation-level Distillation

To ensure the basic learning capability of the student network, we also introduce a segmentation-based auxiliary task into both the RPG and RDR stages. The segmentation module is implemented based on a 1×1 convolutional layer, similar to the teacher network depicted in Section III-C. To distill segmentation-level knowledge from the multimodal teacher network, we exploit a pixel-wise distillation loss as follows:

$$\begin{aligned} Ld_{seg} = {} & \xi \frac{1}{H \times W} \sum_{x,y} l_s\left(\mathbf{q}^s_{x,y}, \mathbf{q}^{seg}_{x,y}\right) \\ & + \left(1 - \xi\right) \frac{1}{H \times W} \sum_{x,y} l_s\left(\mathbf{q}^s_{x,y}, \mathbf{q}^t_{x,y}\right), \end{aligned} \qquad (18)$$

where $\mathbf{q}^{seg}$ is the ground-truth pedestrian mask, and $l_s\left(\cdot\right)$ is the cross-entropy loss function. $\mathbf{q}^s$ and $\mathbf{q}^t$ represent the predicted pedestrian masks by the student network and teacher network, respectively. ξ is a weight parameter, balancing the supervision provided between the ground-truth mask and the teacher's predicted mask.

V. Experiments

A. Data Sets

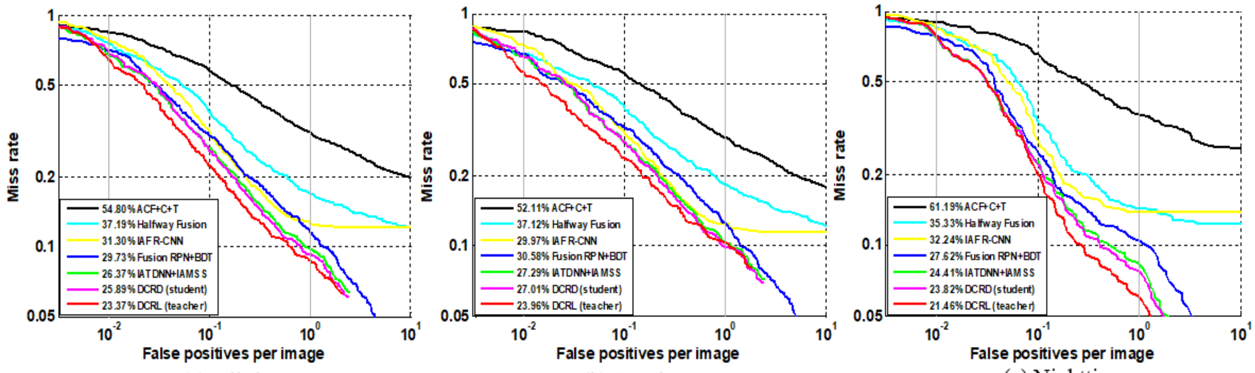
We evaluate the proposed method on the public KAIST multispectral pedestrian benchmark [8]. The KAIST data set consists of 50,172 aligned colour-thermal image pairs, captured by visible and thermal cameras, under different light conditions. Following the works in [32], we utilize 25,086 multispectral image pairs for training. The testing set involves 2,252 pairs of colour-thermal images, in which 1,455 and 797 pairs are captured during daytime and nighttime, respectively. We follow the reasonable evaluation settings presented in [8]. Since the choice of annotations is not consistent in the literature, we conduct a comprehensive evaluation by considering three widely used annotations, including the original annotations (OA) [8], improved annotations (IA) [31], sanitized annotations (SA) [33], and paired annotations (PA) [7]. The log-average miss rate (MR) is adopted to evaluate the performance of different pedestrian detection methods. We average the miss rate over the false positive per image (FPPI) in the range of $[10^{-2}, 10^{0}]$.

B. Implementation Details

The convolutional layers, conv1-5 in the backbone-CNN VGG-16, are initialized with the weights pretrained on ImageNet, for both the teacher (DCRL-PDN) and student (DCRD-PDN) networks. All the other layers or modules are initialized with a random Gaussian distribution. An anchor is considered to be a pedestrian (positive), when its Intersection over Union (IoU) satisfies IoU > 0.5. For the proposed CFL module, we adopt half of the channels for the shared components, i.e., $\alpha = 0.5$, and the modality-specific part is partitioned into 2 groups, i.e., $G = 2$. The batch size is set to be 64. For the training of the proposed DCRL-PDN model (teacher network), we adopt the ADAM optimization algorithm. The learning rate is initialized at 0.001, and decayed by 0.1 after 50 and 75 epochs, with a total of 100 epochs. For the training of the student network, the weights of the teacher network are frozen to provide a stable soft target. Moreover, we adopt the SGD algorithm with an initial learning rate of 0.001 and a momentum of 0.9 for 150 epochs. The weight parameters ζ, ε and ξ in Eq. (16), Eq. (17) and Eq. (18), respectively, are all set as 0.5.

C. Comparison with State-of-the-Art Methods

The proposed DCRL-PDN (teacher model) and DCRD-PDN (student model) networks are compared with several state-of-the-art multispectral pedestrian detection methods. The curves of MR against FPPI using original annotations are shown in Fig. 4. The MR results using improved annotations and paired annotations are presented in Table I. The comparison methods include handcrafted representations, such as ACF+C+T [8], and deep-learning-based methods, such as Halfway Fusion [31], IAF R-CNN [6], Fusion RPN+BDT [32], CMT-CNN [45], IATDNN+IAMSS [11], MSDS-RCNN [33], AR-CNN [7], and AS-MPD [51]. From Fig. 4, it can be observed that the proposed DCRL-PDN network achieves the best detection results, in all of the all-day, daytime and nighttime evaluation settings. Although the state-of-the-art methods, listed in Fig. 4, explore different fusion strategies or stages for processing colour-thermal image pairs, simple concatenation of the features from different branches leads to the ignorance



(a) All-day (b) Daytime (c) Nighttime

Fig. 4. Miss rate curves of different methods on the KAIST pedestrian data set using the reasonable evaluation settings.

TABLE I. Miss rate (%) of different methods using improved annotations ($\mathrm{MR}^{\mathrm{IA}}$), sanitized annotations ($\mathrm{MR}^{\mathrm{SA}}$) and paired annotations (MR^{R} and MR^{T}) on the KAIST pedestrian data set.

Method	$\mathrm{MR}^{\mathrm{IA}}$			$\mathrm{MR}^{\mathrm{SA}}$	MR^{R}			MR^{T}		
	All	Day	Night	All	All	Day	Night	All	Day	Night
ACF+C+T [8]	41.65	39.18	48.29	-	41.74	39.30	49.52	41.36	38.74	48.03
CMT-CNN [45]	36.83	34.56	41.82	-	36.25	34.12	41.21	-	-	-
Halfway Fusion [31]	25.75	24.88	26.59	-	25.10	24.29	26.12	25.51	25.20	24.90
Fusion RPN+BDT [32]	15.91	16.49	15.15	-	15.98	16.60	15.28	16.52	17.56	14.48
IAF R-CNN [6]	15.73	14.55	18.26	-	15.65	14.95	18.11	16.00	15.22	17.56
IATDNN+IAMSS [11]	14.95	14.67	15.72	-	15.14	14.82	15.87	15.08	15.02	15.20
MSDS-RCNN [33]	11.63	10.60	13.73	7.49	11.28	9.91	14.21	12.51	12.02	13.01
AS-MPD [51]	9.68	-	-	5.68	-	-	-	-	-	-
AR-CNN [7]	9.34	9.94	8.38	-	8.86	8.45	9.16	8.26	9.08	7.04
DCRD (student)	12.58	13.12	11.65	8.12	13.64	13.15	13.98	-	-	-
DCRL (teacher)	9.16	9.86	8.18	5.36	9.56	9.08	9.94	8.42	9.22	7.25

of the latent relations in the multispectral data. In contrast, the proposed cross-modal feature learning (CFL) module explicitly explores the common structure and modality-specific cues between RGB and thermal images. In addition, the CFL module is deployed into multiple layers in the DCRL-PDN network, which can learn the cross-modal representations in diversified semantic levels. Compared with the state-of-the-art IATDNN+IAMSS [11], the performance gap (3%) in the all-day setting demonstrates the effectiveness of the shared-specific representation learning. Furthermore, we find that the proposed DCRD-PDN network (student model) achieves the second-best performance, and outperforms other state-of-the-art multispectral pedestrian detectors. This suggests that the proposed multi-stage-based distillation strategy can force the student model to distill the discriminative cross-modal representations guided by the well-trained multispectral teacher model, by only receiving RGB images as input. Although the coarse original annotations contain some inaccuracies, the robust MR results indicate that the proposed DCRL-PDN and DCRD-PDN networks perform better than other approaches, when handling erroneous annotation labels.

From Table I, we can find that the proposed DCRL-PDN network achieves 9.16% MR, which performs better than the recent methods AR-CNN [7], MSDS-RCNN [33] and AS-MPD [51], when using the improved annotations. Moreover, by using the sanitized annotations, the proposed DCRL-PDN network achieves 5.36% MR, which outperforms the MSDS-RCNN [33] and AS-MPD [51] by 2.13% and 0.32%, respectively. The AS-MPD approach mainly investigates multispectral data augmentation strategies based on an anchor-free framework. These augmentation techniques are orthogonal to our proposed method. Moreover, the proposed DCRD-PDN model outperforms the cross-modality transfer learning method CMT-CNN [45] by a large margin, i.e., 24.25%, which validates the effectiveness of the proposed cross-modal knowledge distillation framework and multi-level distillation loss functions. The CMT-CNN model only designs a region-reconstruction task for transferring knowledge between the RGB and thermal modalities, which cannot provide sufficient discriminative representations for pedestrian detection. The paired annotations are presented to explore the effect of the position-shift problem. We also utilize the paired annotations to evaluate the performance of the proposed method. MRR and MRT in Table I indicate the log-average miss rate, in terms of the RGB modality and thermal modality, respectively. The performance of the proposed DCRL-PDN model is comparable with AR-CNN, without using any specific feature alignment operations. Furthermore, we can find that performance gaps

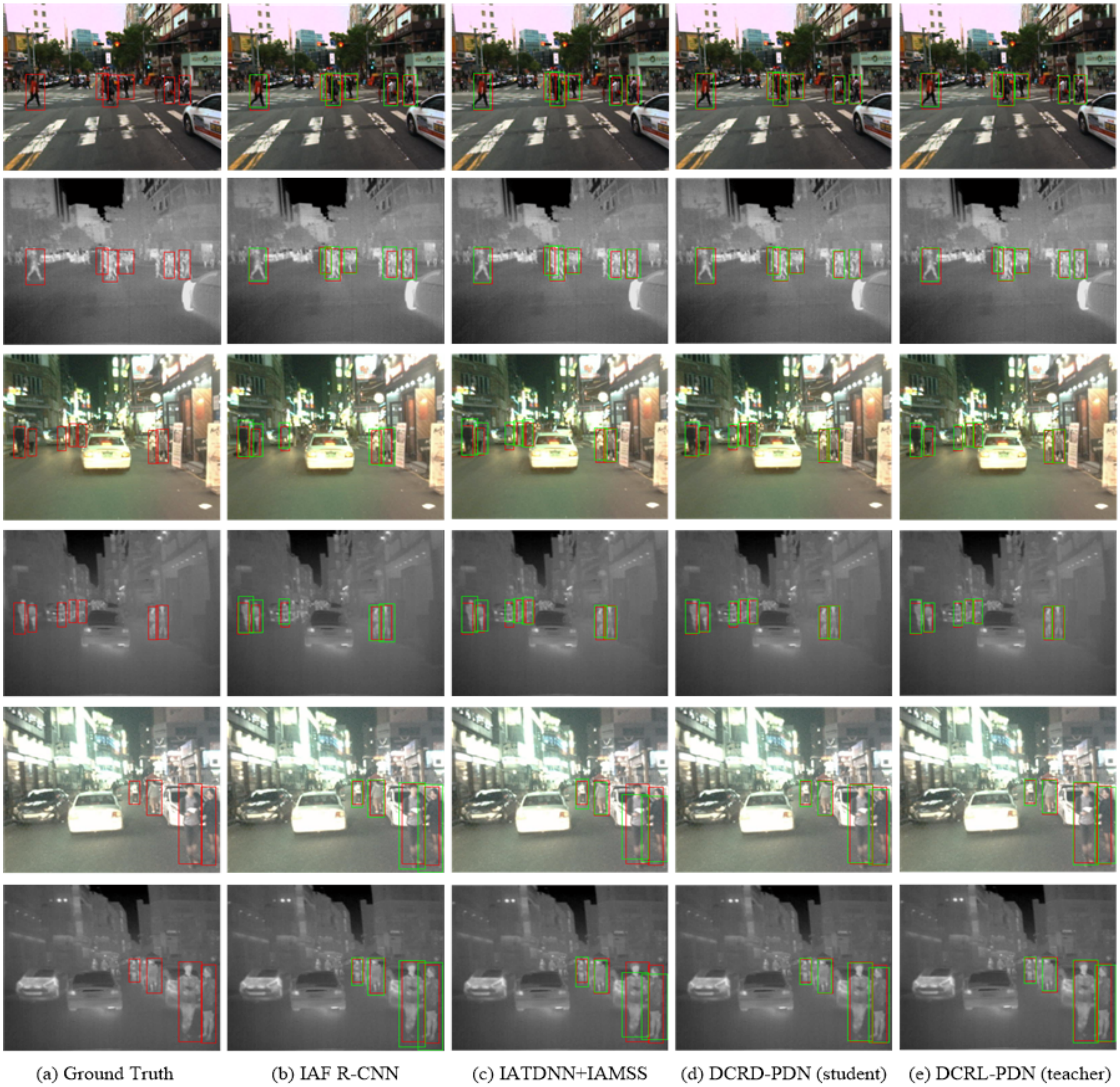


Fig. 5. Sample detection results of different methods evaluated on the KAIST pedestrian data set. The red bounding boxes and green bounding boxes indicate the ground-truth annotations and predicted results, respectively.

exist, when using these three different annotations for the evaluated methods, which reveals the impact of annotations.

To qualitatively demonstrate the effectiveness of the proposed DCRL-PDN and DCRD-PDN networks, we further visualize some detection results of the different methods, as shown in Fig. 5. The first two rows are colour-thermal image pairs captured during daytime, and the other four rows are nighttime image pairs. The first column is the input image pairs with ground-truth annotations, and the other four columns illustrate the detection results generated by IAF R-CNN [6], IATDNN+IAMSS [11], DCRD-PDN (student model) and DCRL-PDN (teacher model), respectively. The red bounding boxes and green bounding boxes indicate the ground-truth annotations and predicted results, respectively. Note that the student model (DCRD-PDN) does not require thermal images as its input, in both the training and testing images. Therefore, we obtain the detection results of the student network from RGB images only, and then, directly visualize the results on thermal images for comparison with other multispectral pedestrian detectors. It can be observed that the proposed DCRL-PDN and DCRD-PDN networks achieve robust detection results under various illumination conditions. For example, in the fifth row of Fig. 5, both the DCRL-PDN and DCRD-PDN models can generate more accurate bounding boxes, by tackling the challenging issues, such as low contrast and low resolution caused by the poor illumination situations.

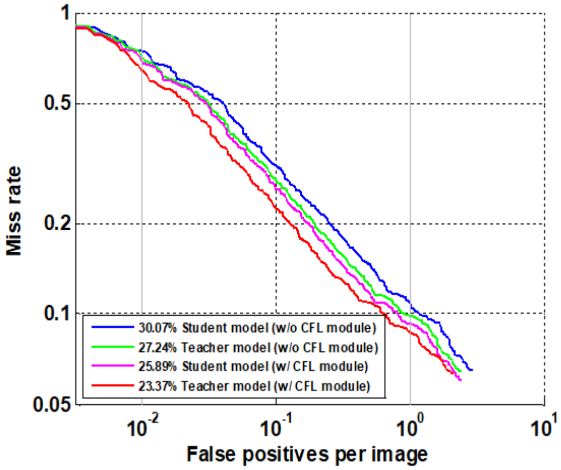


Fig. 6. Comparison of detection performances with (w/) and without (w/o) the proposed CFL module.

D. Ablation Study

1) Effect of the Cross-modal Feature Learning Module: We evaluate the performance of DCRL-PDN (teacher model), with and without the proposed cross-modal feature learning (CFL) module. In addition, we further investigate its corresponding effect on the knowledge distillation of DCRD-PDN (student model). The comparison results are illustrated in Fig. 6. The performance gap is 3.87%, which validates the benefits of explicitly learning shared and modality-specific components between paired RGB and thermal images. Without using the CFL module, the teacher model simply concatenates the features from different modalities and ignores the underlying semantic relations between them. In contrast, the CFL module incorporated into the teacher model can facilitate the information-flow interactions between different modalities in multiple semantic levels, which improves the performance of the final pedestrian detection. Moreover, we can find that the teacher model, trained without the CFL module, has a side effect on the subsequent knowledge distillation. The MR of the student model increases from 25.89% to 30.07%, when learned by the guidance of the teacher model without any CFL module. The student model cannot distill the discriminative cross-modal representations from the teacher model, thus, resulting in a worse detection performance. This suggests that training an effective teacher model is crucial to the cross-modal representation distillation of the student model.

Since the CFL module mainly consists of three components, i.e., shared feature learning (SFL), modality-specific feature learning (MFL), and attention-based aggregation s-trategy (AAS), we further conducted ablation experiments to investigate the contribution of each of the components in the CFL module. The results are presented in Table II. Without shared feature learning, i.e., $\alpha = 0$, the CFL module is identical to employing the group-wise convolution and point-wise convolution in parallel, to refine the intermediate feature maps of each modality. The MR increases from 23.37% to 24.06%, as the CFL module cannot explicitly discover the common structure between the RGB and thermal modalities. When considering another extreme case, i.e., $\alpha = 1$, this implies that the CFL module is forced to only concentrate on learning the shared representations between different modalities by using fully shared parameters. The performance gap (2.05%) confirms the importance of modality-specific features, which provide unique, yet discriminative, cues for pedestrian detection. Without utilizing the attention-based aggregation strategy, both the learned shared and modality-specific components are combined by a basic average pooling operation, which increases the detection error by 1.48%. This validates the effectiveness of AAS, as it adaptively generates the channel-wise aggregation weights by measuring the confidence of the shared and modality specific features.

TABLE II. Ablation results of the proposed CFL module using the reasonable all-day evaluation setting.

Ablation Setting	Miss Rate (%)
CFL w/o MFL	25.42
CFL w/o AAS	24.85
CFL w/o SFL	24.06
CFL	**23.37**

We leverage the visualization toolbox provided by [52], to illustrate the learned shared and modality-specific components generated by the proposed CFL module. The results are shown in Fig. 7. The first sample, i.e., the first row of Fig. 7, is captured under good illumination conditions, and the regions involving pedestrians are highlighted by yellow bounding boxes. It can be observed that the modality-specific features exhibit unique patterns, which are not shareable between the different modalities. The patterns in the RGB-specific features are relatively richer than the ones in the thermal-specific feature maps. The scene in the second row of Fig. 7 is captured under poor light conditions. The RGB-specific component cannot capture discriminative patterns, especially in the pink and green rectangular regions, as the RGB modality is sensitive to illumination variations. The shared features exhibit a compromise in terms of performance, because they aim to explore the common structures between these two modalities. We can find that the thermal-specific features capture rich patterns in all of these three highlighted regions, which provide more reliable information for pedestrian detection.

2) Effect of Modality Attention-based Fusion: To validate the effectiveness of the modality attention-based fusion (MAF) module, we compare it with two other strategies, including average pooling and linearly weighted fusion. The experimental results on the KAIST data set are presented in Table III. For average pooling, we average the classification scores and bounding-box regression offsets from both sub-branch networks. For the linearly weighted fusion strategy, the detection results from each sub-branch network are combined by a set of learnable weights. It can be observed that the modality attention-based fusion module outperforms the other two fusion strategies, in all of the all-day, daytime and nighttime evaluation settings. The linearly weighted fusion is input independent, since the importance weights for each modality (sub-branch) are fixed after training. The MAF module outperforms

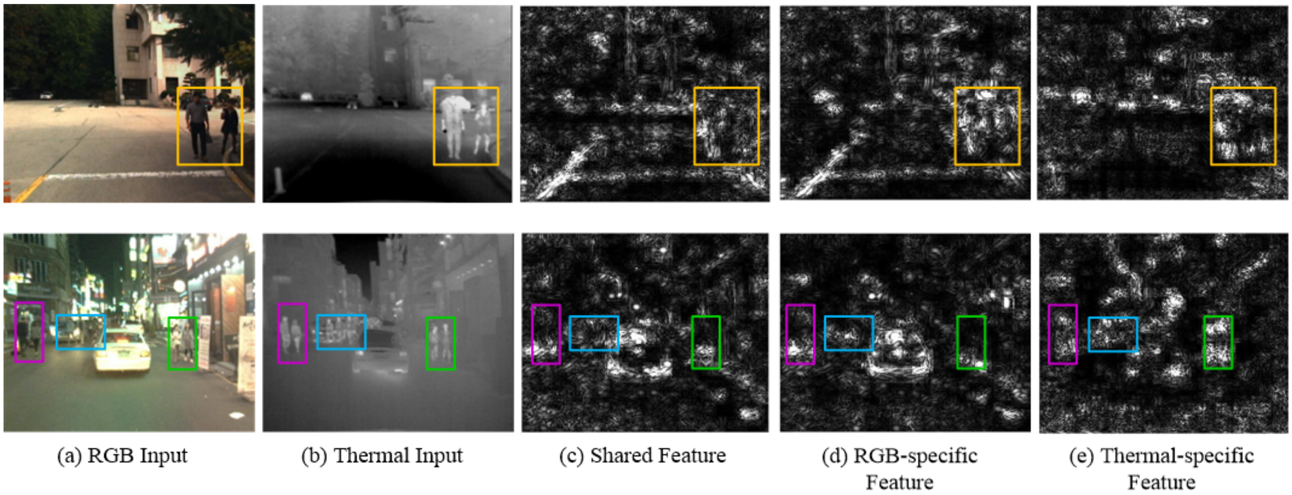


Fig. 7. Visualization of the learned shared and modality-specific components generated by the proposed CFL module on two samples.

TABLE III. Miss rate (%) of the proposed DCRL-PDN network using different fusion strategies.

Fusion Strategy	All-day	Daytime	Nighttime
Average pooling	24.63	25.13	22.82
Linearly weighted fusion	25.35	26.08	23.74
MAF	**23.37**	**23.96**	**21.46**

it by a margin of 1.98%. The underlying reason for this is that the MAF module can dynamically assign input-specific weight for each modality. Since the illumination conditions may undergo drastic changes from daytime to nighttime, the MAF module can adaptively generate importance weights to help the proposed DCRL-PDN network to achieve illumination-invariant pedestrian detection.

3) Effect of the Distillation Loss Function: To evaluate the contribution of each part of the proposed knowledge distillation method, we conduct experiments using different loss terms. The comparison results are summarized in Table IV, where FEA, CLS, BBR and SEG represent the feature-level distillation loss (Eq. (15)), classification-level distillation loss (Eq. (16)), bounding-box regression-level distillation loss (Eq. (17)), and segmentation-level distillation loss (Eq. (18)), respectively. For comparison, we train a baseline model based on the proposed DCRD-PDN network by only receiving RGB images as input, without guidance from the teacher model. We can find that the knowledge distillation, with only feature-level loss, improves the performance over the baseline by a margin of 7.88%, which demonstrates the effectiveness of distilling the cross-modal representations. By combining the detection-level loss functions (CLS+BBR) with the feature-level (FEA) loss function, the performance of the proposed DCRD-PDN network can further improve, because the classification-level distillation loss and bounding-box regression-level distillation loss are directly related to the final detection task. Moreover, by introducing the segmentation-level distillation loss, the DCRD-PDN network achieves the best performance, as it can distill additional segmentation-level knowledge from the teacher network. Since the FEA, CLS, BBR and SEG loss functions are employed for distilling knowledge from different perspectives, they have functions complementary to the learning of the DCRD-PDN network.

TABLE IV. Miss rate (%) of the proposed DCRD-PDN network using different distillation loss terms in reasonable all-day setting.

Ablation Configurations	Miss Rate (%)
Baseline	46.23
FEA (Eq. (15))	38.35
CLS (Eq. (16))	40.46
BBR (Eq. (17))	37.97
SEG (Eq. (18))	41.52
CLS+BBR	34.18
CLS+BBR+FEA	29.86
CLS+BBR+FEA+SEG	**25.89**

4) Effectiveness of Cross-modal Knowledge Distillation: To demonstrate the effectiveness of the proposed cross-modal knowledge distillation framework, we further evaluate the proposed DCRD-PDN network (student model) on two single-modality data sets, i.e., the Caltech [53] and NightOwls [54] pedestrian data sets. The Caltech data set consists of 250,000 RGB images collected from urban traffic scenes, with 350,000 bounding-box annotations for about 2,300 pedestrians. We follow the Caltech-reasonable evaluation settings as in [45]. The NightOwls data set involves 279,000 images captured at night, with bounding-box annotations for 42,273 pedestrians. In addition to the low illumination conditions, this data set also contains several challenging factors, such as occlusion, blur, reflections, etc. We follow the training and testing of split settings, as in [54]. We implement a baseline model, which is equipped with the same network architecture as the DCRD-PDN network. But the baseline model only takes RGB images as input in both training and testing stages. The

TABLE V. Miss rate (%) of different methods on the KAIST, Caltech, and NightOwls pedestrian data sets.

Method	KAIST	Caltech	NightOwls
Faster R-CNN [5]	-	20.98	20.00
CMT-CNN [45]	49.55	10.69	-
A-Faster-R-CNN [55]	-	10.27	18.81
RPN+BF [22]	-	9.58	23.26
SDS-RCNN [23]	-	7.36	17.80
S^3D [56]	-	9.28	-
MS-J [57]	-	8.81	-
Baseline	52.08	22.26	20.28
DCRD (student)	25.89	7.52	16.73

proposed DCRD-PDN network is first trained to distill cross-modal representations on the KAIST data set using RGB-thermal data, then we fine-tune the DCRD-PDN model on the Caltech and NightOwls data sets only using RGB images. For fair comparison, the baseline model is also pretrained on the KAIST data set before fine-tuning on the Caltech and NightOwls data sets. The comparison results on the KAIST, Caltech and NightOwls pedestrian data sets are presented in Table V. On the KAIST data set, DCRD-PDN performs better than the baseline model by a large margin (26.19%), as the DCRD-PDN is trained to produce cross-modal features to deal with illumination changes. The baseline model only explores RGB modality, which results in poor performance in bad lighting conditions. On the Caltech data set, DCRD-PDN also outperforms the baseline model by 14.74%, which shows the benefits of cross-modal knowledge distillation. We can find that the performance gap on the Caltech data set is less than that obtained on the KAIST data set. The underlying reason is that the images in the Caltech data set exhibit relatively good illumination conditions. The thermal information is generally more useful in handling poor lighting cases, as in the KAIST data set. Moreover, the proposed DCRD-PDN network performs better than the cross-modality transfer learning method CMT-CNN on both the KAIST and Caltech data sets. This is mainly due to the fact that the CMT-CNN transfers knowledge between RGB and thermal modalities via region reconstruction, which is not task-oriented. In contrast, the proposed method conducts knowledge distillation at multiple levels, which can provide abundant discriminative cues for pedestrian detection. On the NightOwls data set, DCRD-PDN achieves a satisfactory MR result, i.e., 16.73%. This suggests that the distilled cross-modal representations are helpful for night scenarios. In addition, the performance gaps between different pedestrian detectors are smaller, which reveals that the NightOwls data set is more challenging due to various interference factors.

5) Effect of Segmentation-based Auxiliary Task: To investigate the effect of the segmentation-based auxiliary task for the proposed DCRL-PDN network, we conducted experiments on the KAIST data set using different ablation settings. The results are summarized in Table VI. Without employing the segmentation task, the baseline model achieves a detection error of 25.23%. The error decreases to 23.37%, when introducing the segmentation task into both the RPG and RDR stages. This suggests that the box-based segmentation supervision is beneficial for pedestrian detection. We also explore the effect of introducing the segmentation task into RPG or RDR alone. We can find that the detection performance is improved in both of these cases. Integrating box-based segmentation supervision in RPG brings greater improvement than that in RDR (1.31% vs. 0.55%). Since the bounding-box annotations in the KAIST data set are generally coarse, the inaccuracy will be accumulated when generating segmentation masks for local regions in the RDR stage.

TABLE VI. Ablation results of introducing the segmentation-based auxiliary task into different stages for the DCRL-PDN network on the KAIST data set.

Ablation Configurations		MR (%)
RPG stage	RDR stage	All-day
		25.23
√		23.92
	√	24.68
√	√	**23.37**

TABLE VII. Runtime comparison results of different pedestrian detection methods in seconds per frame (s/f).

Method	Testing Time (s/f)
Fusion RPN+BDT [32]	0.593
CMT-CNN [45]	0.326
Halfway Fusion [31]	0.296
IATDNN+IAMSS [11]	0.185
MSDS-RCNN [33]	0.169
DCRD (student)	0.128
DCRL (teacher)	0.175

6) Runtime Analysis: The computational efficiency of the different pedestrian detection methods is evaluated on a s-ingle RTX 2080 Ti GPU. We summarize the testing time comparison results in Table VII. The Fusion RPN+BDT [32] model integrates the RPN network with boosting trees, which significantly increases the runtime compared with other deep-learning-based approaches. The proposed DCRL-PDN network (teacher model) improves the performance of the state-of-the-art method, MSDS-RCNN [33], with a reduction of the detection error by 2.47%, while only sacrifices a small computational overhead, i.e., 0.006s/f. In the inference phase, our DCRD-PDN network (student model) takes 0.128 seconds only to process one image, which is faster than the domain transfer-learning-based approach CMT-CNN.

VI. Conclusion

In this paper, we first propose a deep cross-modal representation learning-based network, for illumination-invariant pedestrian detection. Different from the common practice of fusing multispectral features by simple concatenation, we propose a cross-modal feature learning (CFL) module, based on a split-and-aggregation strategy, to explicitly explore both the shared structures and modality-specific cues between paired

RGB and thermal images. The proposed CFL module is further deployed into multiple layers of Faster R-CNN, to learn the cross-modal representations in diverse semantic levels. On the other hand, to alleviate the reliance on the thermal data, we propose to leverage a knowledge distillation framework to train a student network, which only receives RGB images as its input and distills cross-modal representations with the guidance from a multimodal teacher network. Specifically, we employ different loss terms to conduct cross-modal knowledge distillation in multiple levels, including the feature, detection, and segmentation perspectives. Extensive evaluations and ablation studies on the public KAIST multispectral pedestrian data set demonstrate that our method achieves superior performance, compared with state-of-the-art multispectral detection approaches.

Acknowledgment

The work described in this paper was supported by a research grant from the Key-Area Research and Development Program of Guangdong Province 2020 under Project 76, and was partially supported by the Education Department of Guangdong Province, PR China, under project No 2019KZDZX1028.

References

[1] C. Lin, J. Lu, G. Wang, and J. Zhou, “Graininess-aware deep feature learning for robust pedestrian detection,” *IEEE Transactions on Image Processing*, vol. 29, pp. 3820–3834, 2020.

[2] Y. Jiao, H. Yao, and C. Xu, “Pen: Pose-embedding network for pedestrian detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1-1, 2020.

[3] Y. Zhao, Z. Yuan, and B. Chen, “Accurate pedestrian detection by human pose regression,” *IEEE Transactions on Image Processing*, vol. 29, pp. 1591-1605, 2020.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 779-788.

[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[6] C. Li, D. Song, R. Tong, and M. Tang, “Illumination-aware faster rcnn for robust multispectral pedestrian detection,” *Pattern Recognition*, vol. 85, pp. 161–171, 2019.

[7] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, “Weakly aligned cross-modal learning for multispectral pedestrian detection,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019, pp. 5126-5136.

[8] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, “Multispectral pedestrian detection: Benchmark dataset and baseline,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1037-1045.

[9] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, and A. M. López, ”Pedestrian detection at day/night time with visible and fir cameras: A comparison,” *Sensors*, vol. 16, no. 6, p. 820, 2016.

[10] L. Zhang, Z. Liu, S. Zhang, X. Yang, H. Qiao, K. Huang, and A. Hussain, ”Cross-modality interactive attention network for multispectral pedestrian detection,” *Information Fusion*, vol. 50, pp. 20–29, 2019.

[11] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, “Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection,” *Information Fusion*, vol. 50, pp. 148–157, 2019.

[12] S. S. S. Kruthiventi, P. Sahay, and R. Biswal, “Low-light pedestrian detection from rgb images using multi-modal knowledge distillation,” in *2017 IEEE International Conference on Image Processing (ICIP)*, Sep. 2017, pp. 4207-4211.

[13] J. Hoffman, S. Gupta, and T. Darrell, “Learning with side information through modality hallucination,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 826-834.

[14] Z. Luo, J.-T. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, “Graph distillation for action detection with privileged modalities,” in *European Conference on Computer Vision (ECCV)*, 2018, pp. 174-192.

[15] N. C. Garcia, P. Morerio, and V. Murino, “Learning with privileged information via adversarial discriminative modality distillation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2581-2593, 2020.

[16] P. Dollár, Z. Tu, P. Perona, and S. Belongie, “Integral channel features,” in *British Machine Vision Conference (BMVC)*, 2009, pp. 1-11.

[17] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast feature pyramids for object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532-1545, 2014.

[18] S. Zhang, R. Benenson, and B. Schiele, “Filtered channel features for pedestrian detection,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1751-1760.

[19] R. Girshick, “Fast r-cnn,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1440-1448.

[20] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in *European Conference on Computer Vision (ECCV)*, 2016, pp. 354-370.

[21] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, “Scale-aware fast r-cnn for pedestrian detection,” *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985-996, 2018.

[22] L. Zhang, L. Lin, X. Liang, and K. He, “Is faster r-cnn doing well for pedestrian detection?” in *European Conference on Computer Vision (ECCV)*, 2016, pp. 443-457.

[23] G. Brazil, X. Yin, and X. Liu, “Illuminating pedestrians via simultaneous detection and segmentation,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 4960-4969.

[24] Z. Cai and N. Vasconcelos, “Cascade r-cnn: Delving into high quality object detection,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 6154-6162.

[25] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, “Learning efficient single-stage pedestrian detectors by asymptotic localization fitting,” in *European Conference on Computer Vision (ECCV)*, 2018, pp. 643-659.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European Conference on Computer Vision (ECCV)*, 2016, pp. 21-37.

[27] C. Lin, J. Lu, G. Wang, and J. Zhou, “Graininess-aware deep feature learning for pedestrian detection,” in *European Conference on Computer Vision (ECCV)*, 2018, pp. 745-761.

[28] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, “High-level semantic feature detection: A new perspective for pedestrian detection,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 5182-5191.

[29] C. Lin, J. Lu, and J. Zhou, “Multi-grained deep feature learning for robust pedestrian detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3608-3621, 2019.

[30] W. Ouyang, X. Zeng, and X. Wang, “Partial occlusion handling in pedestrian detection with a deep model,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 2123-2137, 2016.

[31] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, “Multispectral deep neural networks for pedestrian detection,” in *British Machine Vision Conference (BMVC)*, 2016, pp. 73.1-73.13.

[32] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, “Fully convolutional region proposal networks for multispectral person detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 243-250.

[33] C. Li, D. Song, R. Tong, and M. Tang, “Multispectral pedestrian detection via simultaneous detection and segmentation,” in *British Machine Vision Conference (BMVC)*, 2018, pp. 1-12.

[34] K. Park, S. Kim, and K. Sohn, “Unified multi-spectral pedestrian detection based on probabilistic fusion networks,” *Pattern Recognition*, vol. 80, pp. 143–155, 2018.

[35] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, “Learning consistent feature representation for cross-modal multimedia retrieval,” *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 370-381, 2015.

[36] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, and Y. Yu, “Cross-modality deep feature learning for brain tumor segmentation,” *Pattern Recognition*, vol. 110, p. 107562, 2021.

[37] C. Zhang, K. Zhang, Q. Yuan, H. Peng, Y. Zheng, T. Hanratty, S. Wang, and J. Han, “Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning,” in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 361-370.

[38] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1-1, 2020.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv e-prints arXiv:1503.02531*, Mar. 2015.

[40] Z. Shi and T. Kim, "Learning and refining of privileged information-based rnns for action recognition from depth sequences," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4684-4693.

[41] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 106-121.

[42] J. Shen, N. Vesdapunt, V. N. Boddeti, and K. M. Kitani, "In Teacher We Trust: Learning Compressed Models for Pedestrian Detection," *arXiv e-prints arXiv:1612.00478*, Dec. 2016.

[43] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 742-751.

[44] T. Wang, L. Yuan, X. Zhang, and J. Feng, "Distilling object detectors with fine-grained feature imitation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 4928-4937.

[45] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning crossmodal deep representations for robust pedestrian detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4236-4244.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[47] T. Zhang, G. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 4383-4392.

[48] A. Diba, V. Sharma, L. Van Gool, and R. Stiefelhagen, "Dynamonet: Dynamic action and motion network," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019, pp. 6191-6200.

[49] W. Lee, J. Na, and G. Kim, "Multi-task self-supervised object detection via recycling of bounding box annotations," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 4979-4988.

[50] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," in *International Conference on Learning Representations (ICLR)*, 2016, pp. 1-10.

[51] A. Wolpert, M. Teutsch, M. Saquib Sarfraz, and R. Stiefelhagen, "Anchor-free small-scale multispectral pedestrian detection," in *British Machine Vision Conference (BMVC)*, 2020, pp. 1-14.

[52] U. Ozbulak, "Pytorch cnn visualizations," https://github.com/utkuozbulak/pytorch-cnn-visualizations, 2019.

[53] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 304-311.

[54] L. Neumann, M. Karg, S. Zhang, C. Scharfenberger, E. Piegert, S. Mistr, O. Prokofyeva, R. Thiel, A. Vedaldi, A. Zisserman, and B. Schiele, "Nightowls: A pedestrians at night dataset," in *Asian Conference on Computer Vision*, 2018, pp. 691-705.

[55] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4457-4465.

[56] X. Wang, C. Liang, C. Chen, J. Chen, Z. Wang, Z. Han, and C. Xiao, "S3d: Scalable pedestrian detection via score scale surface discrimination," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3332-3344, 2020.

[57] Y. Pang, J. Cao, J. Wang, and J. Han, "Jcs-net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 12, pp. 3322-3331, 2019.

Tianshan Liu received the B.Sc. and M.Sc. degrees from the School of Internet of Things Engineering, Jiangnan University, China. He is currently pursuing the Ph.D. degree with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. His research interests include computer vision, pedestrian detection, human activity analysis, and multimodal fusion.

Kin-Man Lam (SM'14) received his Associateship in Electronic Engineering with distinction from the Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic) in 1986, the M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College of Science, Technology and Medicine, London, U.K., in 1987, and the Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Australia, in 1996. From 1990 to 1993, Prof. Lam was a lecturer at the Department of Electronic Engineering of The Hong Kong Polytechnic University. He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University again as an Assistant Professor in 1996. He became an Associate Professor in 1999, and has been a Professor since 2010. Currently, he is also an Associate Dean of the Faculty of Engineering. He was actively involved in professional activities. He has been a member of the organizing committee or program committee of many international conferences. Prof. Lam was the Chairman of the IEEE Hong Kong Chapter of Signal Processing between 2006 and 2008. He was a General Co-Chair of the 2012 IEEE International Conference on Signal Processing, Communications & Computing (ICSPCC 2012), APSIPA Annual and Summit 2015, and 2017 IEEE International Conference on Multimedia and Expo (ICME 2017), which were held in Hong Kong, and the Technical Chair of the 2020 IEEE International Conference on Visual Communications and Image Processing. Prof. Lam was the Director-Student Services and the Director-Membership Services of the IEEE Signal Processing Society between 2012 and 2014, and between 2015 and 2017, respectively. He was an Associate Editor of IEEE Trans. on Image Processing between 2009 and 2014, and Digital Signal Processing between 2014 and 2018. He was also an Editor of HKIE Transactions between 2013 and 2018, and an Area Editor of the IEEE Signal Processing Magazine between 2015 and 2017. Currently, he is the VP-Publications of the Asia-Pacific Signal and Information Processing Association (APSIPA). Prof. Lam serves as an Associate Editor of APSIPA Trans. on Signal and Information Processing, and EURASIP International Journal on Image and Video Processing. His current research interests include human face recognition, image and video processing, and computer vision.

Rui Zhao received his B.Sc. degree in electrical engineering from Xi'an Jiaotong University, China, in 2015, and his M.Sc. degree in electrical and electronic engineering from Imperial College London, U.K., in 2017. He is currently a Ph.D. candidate in the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. His research interests include image processing, computer vision, and facial expression recognition.

Guoping Qiu is a Distinguished Professor of Information Engineering, Director of Shenzhen U-niversity Intelligent Robotics Centre at Shenzhen University, China, and a Chair Professor of Visual Information Processing at the University of Nottingham, Nottingham, UK. He has taught in universities in the UK and Hong Kong and also consulted for multinational companies in Europe, Hong Kong and China. His research interests include image processing, pattern recognition, and machine learning. He is particularly known for his pioneering research in high dynamic range imaging and machine learning based image processing technologies. He has published widely and holds several European and US patents. Technologies developed in his lab have laid the cornerstone for successful spinout companies that are developing advanced digital photography software enjoyed by tens of millions of global users.