Illumination Distribution-Aware Thermal Pedestrian Detection

Songtao Li, Mao Ye, *Senior Member, IEEE*, Luping Ji, *Member, IEEE*, Song Tang, *Member, IEEE*, Yan Gan, *Member, IEEE*, and Xiatian Zhu

***Abstract*— Pedestrian detection is an important task in computer vision, which is also an important part of intelligent transportation systems. For privacy protection, thermal images are widely used in pedestrian detection problems. However, thermal pedestrian detection is challenging due to the significant effect of temperature variation on the illumination of images and that fine-grained illumination annotations are difficult to be acquired. The existing methods have attempted to exploit coarse-grained day/night labels, which however even hampers the model performance. In this work, we introduce a novel idea of regressing conditional thermal-visible feature distribution, dubbed as *Illumination Distribution-Aware adaptation* (IDA). The key idea is to predict the conditional visible feature distribution given a thermal image, subject to their pre-computed joint distribution. Specifically, we first estimate the thermal-visible feature joint distribution by constructing feature co-occurrence matrices, offering a conditional probability distribution for any given thermal image. With this pairing information, we then form a conditional probability distribution regression task for model optimization. Critically, as a model agnostic strategy, this allows the visible feature knowledge to be transferred to the thermal counterpart implicitly for learning more discriminating feature representation. Experiment results show that our method outperforms the prior art methods, which use extra illumination annotations. Besides, as a plug-in, our method can averagely reduce about 2% MR on KAIST dataset, and improve about 1% mAP on FLIR-aligned and Autonomous Vehicles datasets without extra calculation for test. Code is available at https://github.com/HaMeow-lst1/IDA.**

***Index Terms*— Domain adaptation, feature co-occurrence, illumination variation, thermal object detection.**

I. Introduction

PEDESTRIAN detection is an important problem in computer vision [1], which plays an important role in autonomous driving, surveillance, and intelligent transportation systems. Obviously, a good intelligent transportation system requires a fast and high-precision pedestrian detection algorithm. There are pedestrian detection methods based on visible images [2]. But these detectors do not work in the insufficient light (nighttime) or bad weather (rain) cases. So multi-spectral pedestrian detection methods emerge based on multiple sensors [3], [4], [5], [6], which work very well. However, there exist some cases in which only thermal imaging sensor exists, for example, privacy concerns. So pedestrian detection based on only thermal images has attracted many attentions.

There are three routes to realize pedestrian detection in thermal image: *direct reconstruction*, *visual knowledge adaptation*, and *illumination-aware*. *Direct reconstruction* tries to design a new network structure for better performance [7], [8], or utilizes the imaging characteristics of thermal images [9], [10]. For the first approach, the key to success is how to extract features unique to thermal images, which itself is sufficiently challenging because of insufficient details existing in the thermal image; while for the latter approach, the existing datasets always lack descriptions of image devices, which makes it difficult to mine the special properties of thermal images. Except the above mentioned approaches, for data scarcity, data augmentation is also employed to improve detection performance in thermal domain [11].

The second line of methods try to train the detector by *visual knowledge adaptation* [12], [13], [14]. In this route, the current methods can be roughly divided into two categories. The first category is *image adaptation* [7], [14], [15]. Since visible images generally have more information than thermal images, a generator model is trained to transform thermal images into visible images. However, this approach requires high quality image generator which itself is also a very difficult problem. Another category is *feature adaptation* [12], [13], [16], [17], which aligns the distributions between visible and thermal features such that detector network can extract visible features. These methods try to extract the common features between visible and thermal domains, but unique thermal modal features are ignored.

The third line of methods use extra *illumination annotations* to endow the detector with the ability of extracting discriminate features [18]. Different from other methods, this approach uses extra illumination annotations, which denotes this image is obtained at day or night. A hard-label classification auxiliary task is employed such that the detector backbone can extract more discriminate features. But illumination annotations are

Manuscript received 25 January 2024; revised 29 May 2024 and 27 July 2024; accepted 4 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62276048, in part by Chengdu Science and Technology Projects under Grant 2023-YF06-00009-HZ, and in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20233323. The Associate Editor for this article was A. Y. S. Lam. *(Corresponding author: Mao Ye.)*

Songtao Li, Mao Ye, and Luping Ji are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: lisongtao@std.uestc.edu.cn; cvlab.uestc@gmail.com; jiluping@uestc.edu.cn).

Song Tang is with the Institute of Machine Intelligence (IMI), University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: steventangsong@gmail.com).

Yan Gan is with the College of Computer Science, Chongqing University, Chongqing 400044, China (e-mail: shiyangancq@cqu.edu.cn).

Xiatian Zhu is with the Surrey Institute for People-Centred Artificial Intelligence, Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH Guildford, U.K. (e-mail: xiatian.zhu@surrey.ac.uk).

Digital Object Identifier 10.1109/TITS.2024.3441628

not always available. Furthermore, simply dividing the images into day and night is not accurate, since thermal imaging is affected by environment temperatures. As shown in Fig.1(a), besides day and night samples, there are also some ambiguous samples, which we define as mixed and indistinguishable day and night samples. So the third route is reasonable but how to design an appropriate auxiliary task is the key to success.

In this work, we propose a novel method, dubbed as Illumination Distribution-Aware adaptation (IDA). Instead of explicit illumination annotations, given a thermal image the conditional visible feature distribution is used as the auxiliary task supervision soft-label signal, because day, night and ambiguous thermal samples have different conditional distributions respectively as shown in Fig.1(b). Specifically, our method consists two parts. The first part is the *Feature Co-occurrence Matrices construction* (FCM) module, the thermal-visible feature joint distribution is constructed which is approximated by feature co-occurrence matrices by calculating and counting the thermal-visible feature for each thermal-visible image pair. Another part is the *Visible Knowledge Adaptation* (VKA) module. For each thermal image, the corresponding visible feature distribution is obtained by querying the feature co-occurrence matrices. Then this distribution is used to as a supervision signal to train an auxiliary task which endows the detector network with the ability of learning more discriminate features and implicitly extracting visible knowledge.

Our contributions can be summarized as follows: (1) We analyze the shortcoming of illumination-aware methods, which need illumination annotations and ignore ambiguous samples. (2) An illumination distribution-aware adaptation strategy is proposed for thermal pedestrian detection. Instead of using explicit day and night annotations, a conditional visible feature distribution regression scheme is proposed. (3) Our methods can be used in many detectors as a plug-in to improve performance.

II. Related Works

A. Pedestrian Detection in Visible Images

Recent advances promoted the development of visible pedestrian detection because visible images have rich information such as color and clear outline. Methods are roughly classified into five strategies. The first one is based on handcrafted features, such as AKBING [19], Zhao et al. [20], DMP [21] and Shen et al. [22]. The second one uses CNN for detection, such as PAMS-FCN [23], CompACT [24], and MCF [25]. They modify the classic object detection model for pedestrian detection. The third one uses attention for detection, such as GDFL [26] and MDFL [27]. These methods add extra attention modules for better feature extraction. The fourth one is occlusion processing. These methods aim to detect occluded persons for better performance, such as MGAN [2], SA-DPM [28] and Zhang et al. [29]. The final one is domain adaptation, which uses thermal images to auxiliary train a detector for visible pedestrian detection, such as CMT-CNN [30].

B. Pedestrian Detection in Multispectral Images

Since visible and thermal images have complementary information, many multispectral pedestrian detection methods

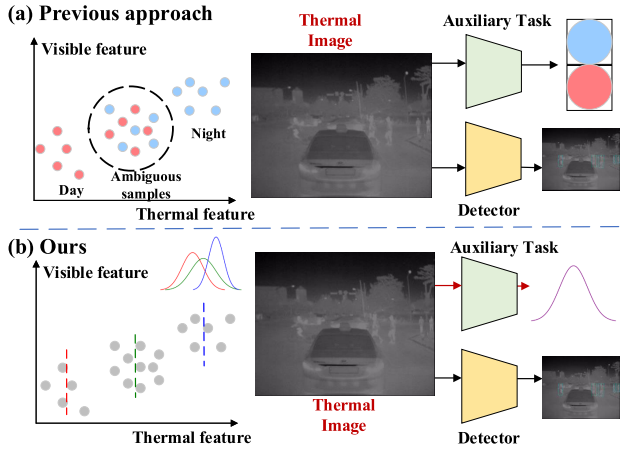

Fig. 1. Comparison between previous illumination-aware approach and our method. (a) An auxiliary task is designed to help the detector network extract discriminate features which simply classifies the image as day or night at training stage. (b) Our method endows detection network with the ability of extracting more discriminate features and implicit visible knowledge by designing an auxiliary network regressing a conditional distribution.

emerge. The first approach is based on *feature fusion*, which uses convolution neural networks to combine both visible and thermal features [31], [32] or attention mechanism for feature fusion [3], [33], [34]. The second route is based on *illumination-aware* module which addresses modality imbalance problems because different illumination conditions can affect the feature distribution of visible and thermal images [4], [35], [36], [36], [37]. They use an extra network to predict whether the image was taken during the day or at night, then add different weights for thermal and visible features according to the predictions.

C. Pedestrian Detection in Thermal Images

Due to insufficient light (nighttime) or bad weather (rain), many methods focus on pedestrian detection in thermal images. As mentioned earlier, thermal pedestrian detection methods have three strategies. The first one is *direct reconstruction*. For example, GPCAnet [9] is designed to predict the pedestrian position via an additional task. Kim et al. [10] proposed a memory network to transform low-scale features into large-scale features. Bongini et al. [11] proposed a novel data augmentation approach compositing 3d fake thermal objects in a real thermal scene. The second strategy is *visible knowledge adaptation*. For image adaptation, Guo et al. [15] used CycleGAN to generate visible images for detection from thermal images. While for feature adaptation, Kim et al. [38] proposed a memory network to transform thermal features into visible features. The final strategy is *illumination-aware* such as TC Thermal [18] and TC Det [18]. They use an extra network to predict whether the image is taken at the day or night, then use the auxiliary network features to guide the detector to extract more discriminate features.

D. Co-Occurrence Learning

Co-occurrence learning aims to count the number of co-occurrence between different elements as co-occurrence

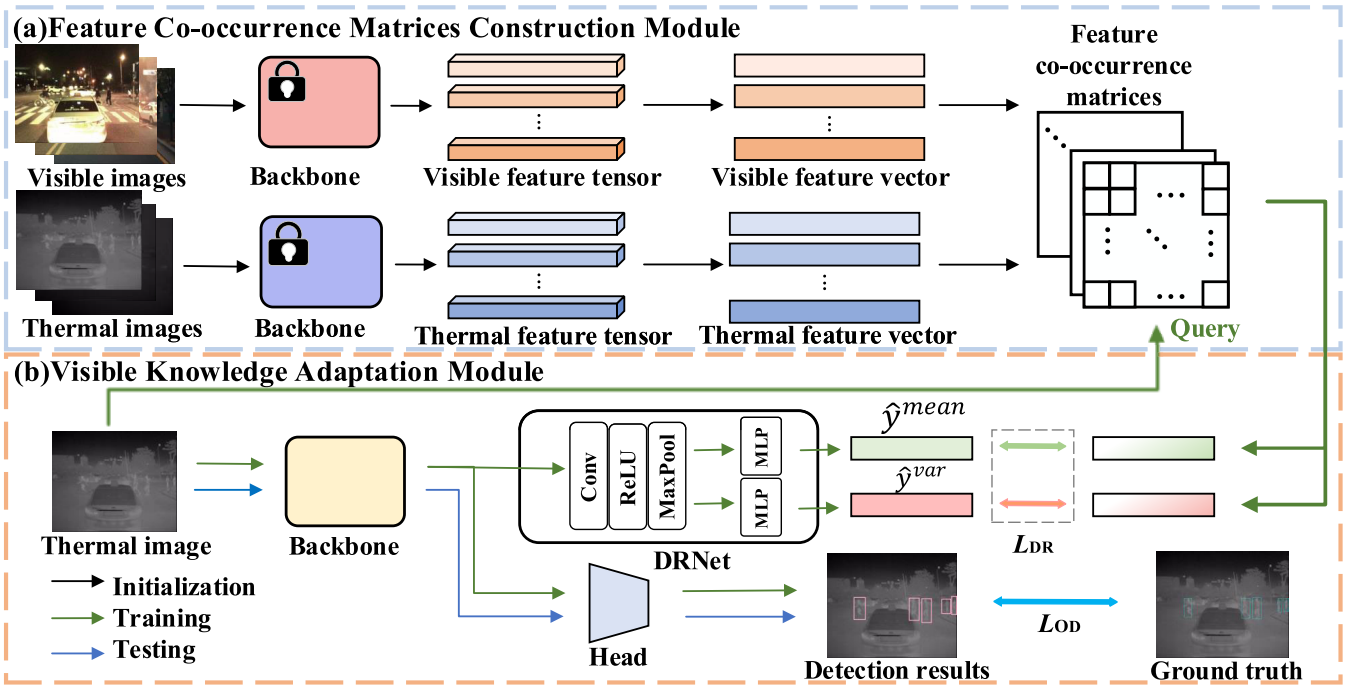

Fig. 2. Overview of the proposed **IDA** method. The black, green, and blue lines indicate the initialization, training, and testing flow respectively. (a) The feature co-occurrence matrices construction module is used to approximate the thermal-visible joint feature distribution. (b) Visible knowledge adaptation module endows detector with the ability of learning more discriminate features and implicitly extracting visible knowledge.

frequency for analysis. The existing methods for computer vision can be roughly classified into two categories. The first category is *handcrafted* based method, which counts co-occurrence frequency between different position pixel values as features in an image or a region for downstream tasks, for example, CoF [39] for background blurring and image recoloring, CoTM [40] for template matching, and COS [41] for mining consistent feature correspondences. Another category is based on *deep learning* framework which counts feature co-occurrence frequency of different channels in an image as an attention map for feature augment, such as CoL [42] for image classification and semantic pixel labeling, CoCNN [43] for recommendation and COOC [44] for object recognition. The existing approaches count co-occurrence frequency between different parts in one sample or in one domain. Instead of previous approaches, our method counts feature co-occurrence frequency between the thermal and visible domains, which is used as a bridge to connect the thermal and visible knowledge.

III. The Proposed Method

A. Problem Statement

Let $D_s = \{(x_{s,i}^V, x_{s,i}^T, y_{s,i})\}_{i=1}^{N_s}$ denotes the training dataset. $x_{s,i}^V \in \mathbf{R}^{w\times h\times c}$ is the i-th visible image, and $x_{s,i}^T \in \mathbf{R}^{w\times h\times c}$ is the corresponding i-th paired thermal image, where w, h and c are the width, height and channel number respectively. $y_{s,i}$ denotes the object annotations of the i-th paired visible-thermal images. N_s denotes the total number of thermal-visible image pairs in training set. Suppose the test dataset $D_t = \{x_{t,j}^T\}_{j=1}^{N_t}$, where $x_{t,j}^T \in \mathbf{R}^{w\times h\times c}$ and N_t denotes the total number of test thermal images. Our goal is to construct an auxiliary task to endow the detector network with the ability of extracting discriminate features and visible knowledge by using the paired thermal-visible image pairs at train stage such that it can obtain better performance at test stage than that only trained based on the thermal images.

1) Overview: The proposed thermal pedestrian detection method is based on illumination distribution-aware adaptation (IDA) which consists of two parts. As shown in Fig.2, one is the Feature Co-occurrence Matrices construction (FCM) module; another one is Visible Knowledge Adaptation (VKA) module. The detector baseline in Fig.2 can be one-stage detector such as YOLOv3 [45] or two-stage detector such as Faster RCNN [46]. FCM module (Fig.2(a)) first uses the pre-trained and freezen backbone to extract the features of each pair of thermal-visible images. Then they are reduced to the paired thermal-visible feature vectors. In the end, these thermal-visible feature vectors are used to create feature co-occurrence matrices, which are used to approximate a joint thermal-visible feature distribution.

After the initialization of feature co-occurrence matrices, VKA module uses them as a supervision signal for an auxiliary task. For each thermal image, by querying the feature co-occurrence matrices, a Gaussian distribution is used to approximate the conditional visible feature distribution. Specifically, the queried samples are used to calculate a mean vector and a variance vector. The auxiliary task is designed to predict the mean and variance vectors. In this way, the detector backbone can extract the discriminate features to distinguish the day, night, and ambiguous image samples. Meanwhile, through the alignment with visible feature distribution, the detector also implicitly obtains the ability of extracting the corresponding visible knowledge.

B. Feature Co-Occurrence Matrices Construction

In this section, the aim is to obtain the conditional visible feature distribution given a thermal image. Before this, we need to acquire the joint distribution between thermal and visible features. Here we use feature co-occurrence matrices to represent this joint distribution. After normalization and quantification, the continuous feature values are discretized into multiple ranges. The problem of distribution computation is approximated by counting the co-occurrences of thermal-visible features contained in each range.

Based on the above idea, first all of the visible and thermal images in the training set are used to train a visible detector and a thermal detector respectively. Then the corresponding thermal and visible feature extraction backbones $\widetilde{\phi}_T$ and $\widetilde{\phi}_V$ are frozen. For each pair of thermal and visible images, the corresponding features $\widetilde{F}_i^T, \widetilde{F}_i^V \in \mathbf{R}^{W\times H\times C}$ can be obtained as follows,

$$\widetilde{F}_i^T = \widetilde{\phi}_T(x_{s,i}^T), \tag{1}$$
$$\widetilde{F}_i^V = \widetilde{\phi}_V(x_{s,i}^V), \tag{2}$$

where $x_{s,i}^T$ and $x_{s,i}^V$ denote the i-th thermal and visible images respectively. $\widetilde{\phi}_T$ and $\widetilde{\phi}_V$ are the corresponding trained thermal and visible backbones. W, H and C denote the width, height and channel number of feature tensor respectively.

For simplicity, the feature tensor is further reduced to a feature vector. Specifically, the feature tensor is divided into K parts according to the channel dimension. The feature vector is calculated as follows,

$$\widetilde{z}_i^{V/T}[j] = \frac{K}{W\times H\times C}\sum_{m,n}\sum_{p=(C\times j)/K}^{(C\times(j+1))/K-1}\widetilde{F}_i^{V/T}[m,n,p], \tag{3}$$

where $\widetilde{F}_i^{V/T}[m,n,p]$ denotes the (m,n,p) coordinate value of $F_i^{V/T}$ with the size $W\times H\times C$. $\widetilde{z}_i^T, \widetilde{z}_i^V \in \mathbf{R}^K$. $z[j]$ denotes the j-th element of a vector z for $j = 1,\cdots,K$. Then the vectors $\widetilde{z}_i^{V/T}[j]$ are normalized to ensure that the element value is between 0 to 1 for $j = 1,\cdots,K$. They are written as follows,

$$\widetilde{f}_i^{V/T}[j] = \frac{\widetilde{z}_i^{V/T}[j] - \min\{\widetilde{z}_i^{V/T}[j]\}_{i=1}^{N_s}}{\max\{\widetilde{z}_i^{V/T}[j]\}_{i=1}^{N_s} - \min\{\widetilde{z}_i^{V/T}[j]\}_{i=1}^{N_s}}, \tag{4}$$

where $\widetilde{f}_i^T$ and $\widetilde{f}_i^V$ denote the i-th normalized thermal and visible feature vectors respectively. $min\{\cdot\}$ and $max\{\cdot\}$ denotes the corresponding minimum and maximum values. Based on this computation, the joint thermal-visible feature distribution is approximated by K two-dimensional random variables $[\widetilde{f}^T[j],\ \widetilde{f}^V[j]]$ for $j = 1,\cdots,K$.

Then the feature value is discretized into N ranges. The discrete random variable has N possible values $(0, 1,\cdots, N-1)/N$. In this way, the joint features $\widetilde{f}_i^T[j]$ and $\widetilde{f}_i^V[j]$ are quantified to $N\times N$ parts. Correspondingly, the quantified two-dimensional discrete random variables $[\bar{f}^T[j],\ \bar{f}^V[j]]$ has $N\times N$ possible values. As shown in Fig.3(a), this joint distribution is approximated by a feature co-occurrence matrix constructed by counting the co-occurrence frequencies. Totally, K matrices are computed to approximate the whole joint distribution. Specifically, the feature co-occurrence matrix $A_j \in R^{N\times N}$ for the j-th part is

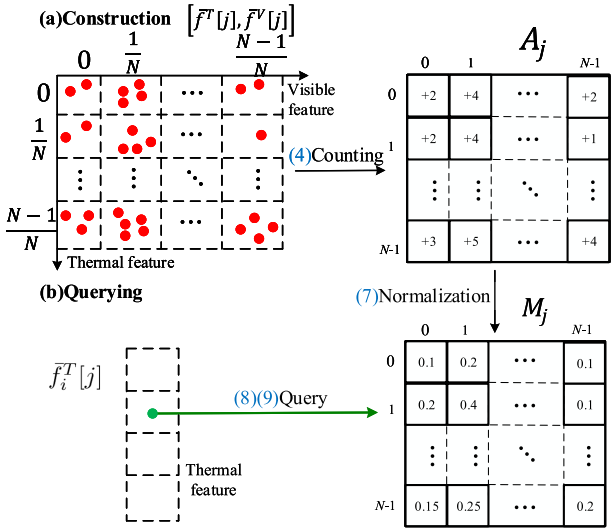

Fig. 3. Illustration of constructing and using feature co-occurrence matrices. (n) denotes this part is related to formula (n).

$$A_j^{m,n} = \varepsilon + \sum_{i=1}^{N_s}\mathbb{1}(I_i^T[j]=m)\cdot\mathbb{1}(I_i^V[j]=n), \tag{5}$$

where $\mathbb{1}(\cdot)$ equals 1 if the value inside the bracket is true and 0 otherwise. $I_i^T[j]$ and $I_i^V[j]$ denote the j-th quantized indexes respect to the i-th thermal and visible images, which are calculated as

$$I_i^{T/V}[j] = \begin{cases} \lfloor \widetilde{f}_i^{T/V}[j]\times N\rfloor, & \widetilde{f}_i^{T/V}[j] < 1, \\ N-1, & \widetilde{f}_i^{T/V}[j] = 1 \end{cases} \tag{6}$$

where $\lfloor\cdot\rfloor$ denotes the round down operation. Here, for avoiding the case of zero values of a certain row such that the conditional distribution can not be calculated, a small positive parameter ε is set as 1.0×10^{-6}.

The feature co-occurrence matrices A record the frequency relationship between different thermal-visible feature values. Since we aim to predict the conditional visible feature distribution, the feature co-occurrence matrices are row-wise normalized as shown in Fig.3(b), which are written as

$$M_j^{m,n} = \frac{A_j^{m,n}}{\sum_{i=1}^{N}A_j^{m,i}}, \tag{7}$$

for $j = 1,\cdots,K$. We call them conditional feature co-occurrence matrices. For every row, the sum of all values is 1, which can be regarded as a discrete random variable whose values are from $(0,\cdots,N-1)/N$. Each discrete random variable is used to approximate a Gaussian distribution to supervise an auxiliary task.

Remark: The features are always high dimensional. So K matrices are constructed to approximate the joint thermal-visible feature distribution. Besides, since the channel of the feature is usually an integer power of 2 such as 1024 or

2048, the value of K is also set as an integer power of 2 such as 16, 32, or 64. If the value of K is high, the distribution is more completely approximated but more storage space is needed and the conditional distribution vectors are difficult to be predicted. On the other hand, if the value of K is low, although less space is required and the conditional distribution vectors are easy to be predicted, the joint distribution can not be fine approximated. Similarly, the discretization slot number N obeys the same phenomena.

C. Visible Knowledge Adaptation

After creating conditional feature co-occurrence matrices, we can obtain the corresponding visible feature distribution given a thermal feature. Without the loss of generality, given a thermal image, suppose that the conditional visible feature obeys Gaussian distribution, an auxiliary task is constructed to predict the mean and variance vectors instead of specific distribution. In this way, the detector backbone can extract more discriminate features and implicitly align the thermal domain with the visible domain.

For the i-th thermal image, the corresponding quantified features are $\bar{f}_i^T[j]$ for $j = 1, \cdots, K$. The feature $\bar{f}_i^T[j]$ corresponds to the visible feature row $I_i^T[j]$ of the j-th conditional feature co-occurrence distribution matrix as shown in Fig.3(b). Suppose the elements of this row are sampled from a Gaussian distribution, then the corresponding mean and variance can be calculated as follows,

$$y_i^{mean}[j] = \sum_{n=0}^{N-1} M_j^{I_i^T[j],n} \cdot \frac{n}{N}, \tag{8}$$

$$y_i^{var}[j] = -(y_i^{mean}[j])^2 + \sum_{n=0}^{N-1} M_j^{I_i^T[j],n} \cdot (\frac{n}{N})^2, \tag{9}$$

for $j = 1, \cdots, K$, where y_i^{mean} and y_i^{var} denote the mean and variance vectors of the visible features corresponding to the i-th thermal image respectively.

Then an auxiliary network is proposed named Distribution Regression Net (DRNet) to predict the above computed mean and variance vectors. The input of DRNet is the feature map of the ith thermal image $x_{s,i}^T$, which can be presented as follows,

$$F_i^T = \phi_T(x_{s,i}^T), \tag{10}$$

where ϕ_T denotes the detector backbone to be learned. The outputs of DRNet are two K-length vectors, which are the predictions of mean and variance vectors respectively. DRNet contains two parts. The first part consists of a convolution layer shaping channel number to 256, a ReLU activation function and a MaxPool layer reshaping feature maps to quarters, which can be denoted as

$$\varphi_i = MaxPool \circ ReLU \circ Conv_{1\times1}(F_i^T). \tag{11}$$

The effect of the first part is to reduce the feature dimensionality to simplify calculations. Next, we resize φ_i to a vector φ_i^* and use two MLPs networks to regress the mean and variance vectors as the following,

$$\hat{y}_i^{mean} = MLP_1(\varphi_i^*), \quad \hat{y}_i^{var} = MLP_2(\varphi_i^*), \tag{12}$$

where $\hat{y}_i^{mean}$ and $\hat{y}_i^{var}$ denote the predictions of mean and variance vectors respectively. Here, MLP network consists of four parts: A linear layer shaping the vector length to 256, a ReLU activation function, a dropout layer with probability of $p = 0.5$, and an another linear layer.

The auxiliary task is to minimize the difference between the predictions and ground truth of the mean and variance vectors to train the backbone ϕ_T such that it can extract discriminate features and also implicitly align the thermal domain to the visible domain. The distribution regression loss is defined as

$$L_{DR} = \lambda_m \|\hat{y}_i^{mean} - y_i^{mean}\|_2 + \lambda_v \|\hat{y}_i^{var} - y_i^{var}\|_2. \tag{13}$$

The whole network is trained end-to-end. The total loss is defined as

$$L_{total} = L_{OD} + L_{DR}, \tag{14}$$

where L_{OD} is the standard detection loss of any detector such as YOLO and Faster-CNN. The hyper-parameters λ_m and λ_v are two balance factors. To better illustrate our method, the pseudocode is shown as Algorithm 1.

Remark: Based on our conditional feature co-occurrence matrices, different thermal images correspond to different rows of matrices, so the mean and variance vectors of day, night, and ambiguous samples are different. Therefore, day, night, and ambiguous samples can be distinguished by predicting different mean and variance vectors. Thus, the discriminate features can be extracted. Furthermore, by predicting the mean and variance vectors of the corresponding visible features, our auxiliary task can also help the detector backbone extract visible knowledge implicitly. At test phase, the auxiliary task is not needed any more. So our model can be used as a plug-in to any detector backbones. Experimental results also confirm our strategy works. Moreover, our method generalizes over different illumination and temperature scenarios, as our model design does not assume any specific conditions of these factors. The feature co-occurrence matrices can directly capture the changes in illumination and temperatures as encoded in the training data.

IV. Experiments

Datasets: Our experiments are conducted on the KAIST dataset, FLIR-aligned dataset, and Autonomous Vehicles dataset. (1) The KAIST dataset consists of 59328 thermal-visible image pairs for training and 45156 image pairs for test. As is common practice [47], [48], [49], [50], we sample every two frames from training videos and exclude heavily occluded and small person instances ($<$ 50 pixels). Meanwhile, we use the training annotations from [49] and test annotations from [50]. The final dataset consists 7601 thermal-visible image pairs for training and 2252 image pairs for test. (2) To verify the generalization of our method, we also use the FLIR-aligned dataset [51] for evaluation. This dataset contains 4129 visible-thermal image pairs for training and 1013 pairs for test on three categories (person, bicycle, and car). Only thermal images are used for test. To confirm the extention to complex scenarios, Autonomous Vehicles dataset [52] is used which contains five categories: bike, car, car_stop, color_cone,

Algorithm 1 Algorithm With IDA for Training

Input: Pre-trained visible backbone $\widetilde{\phi}_V$, pre-trained thermal backbone $\widetilde{\phi}_T$, backbone ϕ_T, detector head D, DRNet F_{DR}, paired visible-thermal images and annotations $D_s = \{(x_{s,i}^V, x_{s,i}^T, y_{s,i})\}_{i=1}^{N_s}$.
Output: backbone ϕ_T, detector head D.

```
// collect feature tensors
```
Calculate each visible-thermal image pair's feature tensor by formulas (1) and (2);
```
// reduce dimension
```
Calculate each visible-thermal image pair's feature vector by formula (3);
```
// normalization
```
Normalize each visible-thermal image pair's feature vector by formula (4);
for $j = 1 : K$ **do**
  // init a $N \times N$ matrix with ϵ
  $A_j \leftarrow Init\_Matrix(N \times N, \epsilon)$;
  // calculate the feature co-occurrence matrix
  for $i = 1 : N_s$ **do**
    $I_i^T \leftarrow \max(int(\widetilde{f}_i^T[j] \times N), N-1)$;
    $I_i^V \leftarrow \max(int(\widetilde{f}_i^V[j] \times N), N-1)$;
    $A_j^{m,n} \leftarrow A_j^{m,n} + 1.0$;
  end
  Calculate the M by formula (7);
  // calculate the ground truth of mean-var vectors
  $y_i^{mean}[j] \leftarrow 0.0$;
  $y_i^{var}[j] \leftarrow 0.0$;
  for $i = 1 : N_s$ **do**
    for $n = 0 : N-1$ **do**
      $y_i^{mean}[j] \leftarrow y_i^{mean}[j] + M_j^{I_i^T[j],n} \cdot \frac{n}{N}$;
      $y_i^{var}[j] \leftarrow y_i^{var}[j] + M_j^{I_i^T[j],n} \cdot (\frac{n}{N})^2$;
    end
    $y_i^{var}[j] \leftarrow y_i^{var}[j] - (y_i^{mean}[j])^2$;
  end
end
for $k = 1 : epochs$ **do**
  for $i = 1 : N_s$ **do**
    $F_i^T \leftarrow \phi_T(x_{s,i}^T)$;
    // predict the mean-var vectors
    $\hat{y}_i^{mean}, \hat{y}_i^{var} \leftarrow F_{DR}(F_i^T)$
    Calculate the loss by formula (13);
    $\hat{y} \leftarrow D(F_i^T)$;
    Calculate the total loss by formula (14);
    Update the network;
  end
end

and person. We use the RGB images as visible images, and FIR images as thermal images. This dataset contains 6009 paired images for training and 1503 for test.

Evaluation Metric: (1) For the KAIST dataset, we follow the setting in the works [47], [48]. The evaluation metric is the log-average miss rate (MR) for thresholds in the range of $[10^{-2}, 100]$. We set 0.5 as the Intersection over Union (IoU) threshold to calculate True Positives (TP), False Positives (FP), and False Negatives (FN). We also include floating-point operations per second (FLOPS) and the number of parameters (Param) in the evaluation metrics. (2) For the FLIR-aligned dataset, models are evaluated with mean Average Precision (mAP), mAP50, and mAP75. Note that mAP50/mAP75 means the IOU value is 0.5/0.75. Besides, mAP means we calculate the average of 10 results with IOU as [0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95].

Implementation Details: The networks are pre-trained on the COCO dataset. We set our method as a plug-in combined with YOLOv3 [45], Faster RCNN [46], and Cascade RCNN [53]. This is because YOLOv3 is the most common single-stage detector for thermal pedestrian detection. Besides, FasterRCNN and CascadeRCNN are great two-stage detectors. The input image size is 640×512. For these three models, MMDetection is used [54]. (1) For YOLOv3, the batch size is set as 4. We choose Darknet53 for the backbone. The hyper-parameters K and N are set as 32 and 40 respectively. λ_v is 200.0. Besides, λ_m is 75.0 on the KAIST dataset and 100.0 on the FLIR-aligned dataset. (2) For Faster RCNN and Cascade RCNN, the backbone is Resnet50 [55]. The epoch number is 4. SGD is used to train the detector with an initial learning rate as 0.005 for the first three epochs and 0.005 for the last epoch. The baseline is Feature Pyramid Network (FPN) [56]. The batch size is 4. K and N are set as 32 and 40. For FasterRCNN, λ_m is 0.1 and λ_v is 0.125. For CascadeRCNN, λ_m is 0.1 and λ_v is 0.2. On the FLIR-aligned dataset, the parameter setting is almost the same. K and N are set as 16 and 40 for CascadeRCNN. For YOLOX on FLIR-aligned dataset, K and N are 32 and 40, respectively. λ_m is 10.0 and λ_v is 20.0. We use SGD with learning rate as 5.0×10^{-5}.

A. Comparisons to State-of-the-Arts

1) Compared Methods: On KAIST dataset, we choose three categories of methods for comparison whose codes are public available. The first category is the *direct reconstruction* which is based on single thermal domain such as TPIHOG [57], FasterRCNN-T [50], and Ghose et al. [8]. The second category is *visual knowledge adaptation* such as SSD300 [7], VGG16-two-stage [15], ResNet101-two-stage [15], Bottom-up [13], Top-down [13], Mixed 40_60 [14], Mixed 80_20 [14], Mixed 90_10 [14], Xie et al. [17], and DIP [58]. The third category is *illumination-aware* approach such as TC Thermal [18] and TC Det [18]. We implement our method IDA based on the detection frameworks YOLOv3, FasterRCNN, and CascadeRCNN. The detection results of the compared methods are from the corresponding papers. For FLIR-aligned dataset, we compare our methods with YOLOv3(Xie et al.) [17], DIP [58], and TIRDet [59]. All these methods use paired visible-thermal images for training. Besides, to explore the generation capability of our method, we also choose Autonomous Vehicles dataset for the experiment. Compared with the KAIST and FLIR-aligned datasets, the scenario of this dataset is more complex. On Autonomous Vehicles dataset, we compare our

TABLE I
COMPARISONS ON THE KAIST DATASET AT DAY AND NIGHT IN TERMS OF MR. FLOPS AND PARAM ARE ALSO USED FOR COMPARISON. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Method	All↓	Day↓	Night↓	FLOPS↓	Param↓
TPIHOG [57]	-	-	57.38	-	-
FasterRCNN-T [50]	47.59	50.13	40.93	$280.514^{\dagger}$	$49.007^{\dagger}$
Ghose et al. [8]	-	30.40	21.00	$328.755^{\dagger}$	$192.845^{\dagger}$
SSD300 [7]	69.81	-	-	$\mathbf{30.428}^{\dagger}$	$\mathbf{23.745}^{\dagger}$
VGG16-two-stage [15]	46.30	53.37	31.63	$157.304^{\dagger}$	$191.103^{\dagger}$
ResNet101-two-stage [15]	42.65	49.59	26.70	$289.586^{\dagger}$	$101.681^{\dagger}$
Bottom-up [13]	35.20	40.00	20.50	62.103^{*}	61.524^{*}
Top-down [13]	36.30	42.30	20.40	62.103^{*}	61.524^{*}
Mixed 40_60 [14]	34.78	43.45	14.53	$74.520^{\dagger}$	$110.047^{\dagger}$
Mixed 80_20 [14]	25.88	33.01	11.12	$74.520^{\dagger}$	$110.047^{\dagger}$
Mixed 90_10 [14]	25.62	31.86	12.92	$74.520^{\dagger}$	$110.047^{\dagger}$
YOLOv3(Xie et al.) [17]	23.49	30.11	9.64	$272.575^{\dagger}$	$114.500^{\dagger}$
DIP [58]	19.12	25.16	7.49	156.040^{*}	59.090^{*}
TC Thermal [18]	28.53	36.59	11.03	62.105^{*}	63.659^{*}
TC Det [18]	27.11	34.81	10.31	62.105^{*}	63.659^{*}
YOLOv3 + IDA (Ours)	22.53	28.46	9.27	62.030	61.520
FasterRCNN + IDA (Ours)	**18.88**	**24.97**	6.61	75.580	41.120
CascadeRCNN + IDA (Ours)	18.93	25.63	**5.86**	103.390	68.930

method with YOLOv3, FasterRCNN, and CascadeRCNN as shown in Table III. Note that in the table the symbol $*$ indicates that the method is open-source. We use the official code to calculate the FLOPS and Param. The symbol † denotes the method is not open-source; for these methods, we reproduced the codes closest to the results reported in their papers. Their FLOPS and Param are calculated based on the reproduced codes.

2) Quantitative Comparisons: From Table I, by adding our method as a plug-in to YOLOv3, FasterRCNN and CascadeRCNN, all three detectors get good performance. Interestingly, comparing TC Det [18] to YOLOv3 + IDA (Ours), our method gets better performance in the day, -2.40% from 34.81% to 32.41% vs -0.04% from 10.31% to 10.27% at night. This is because our method implicitly aligns the extracted features with the conditional visible feature distribution. Visible features in the day are richer than those at night and thermal features in the day are also not informative. So the performance is boosted much in the day. This proves that our method has the ability to obtain the visible features. Moreover, as a plug-in, our method is lightweight without extra networks and calculations for test. However, other one-stage-based methods need extra networks to transform thermal features into visible features. This leads to much computations because of large size of feature maps. In some two-stage-based methods, a network is proposed to transform thermal ROI features into visible ROI features. This also leads to computation burden because of many ROIs.

The quantitative comparisons on the FLIR-aligned dataset are shown in Table II. With our method as a plug-in, we can improve mAP $+2.1\%$ for YOLOv3, $+1.4\%$ for FasterRCNN, and $+0.6\%$ for CascadeRCNN. This demonstrates the generalization ability of our method across different datasets. In addition, as shown in Table III, our method can also be applied to more complex scenario, i.e. Autonomous Vehicles dataset. Detection performance is improved.

TABLE II
COMPARISONS ON FLIR-ALIGNED DATASET WITH PERSON, BICYCLE, AND CAR. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Method	mAP	mAP50	mAP75
Xie et al. [17]	-	76.38	-
DIP [58]	-	77.28	-
TIRDet [59]	44.3	**81.4**	41.1
YOLOv3 + IDA (Ours)	37.5	75.0	30.3
FasterRCNN + IDA (Ours)	40.3	75.4	35.8
CascadeRCNN + IDA (Ours)	41.6	76.2	36.7
YOLOX + IDA (Ours)	**46.3**	81.1	**43.4**

TABLE III
COMPARISONS ON AUTONOMOUS VEHICLES DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Method	mAP	mAP50	mAP75
YOLOv3	34.6	71.9	29.9
FasterRCNN	39.9	72.5	39.8
CascadeRCNN	42.2	72.9	42.7
YOLOv3 + IDA (Ours)	35.7	73.0	30.8
FasterRCNN + IDA (Ours)	41.0	73.0	40.6
CascadeRCNN + IDA (Ours)	**43.1**	**73.1**	**43.8**

In conclusion, (1) compared with the *direct reconstruction* based methods, our methods get better performance because our methods use the visible knowledge to train the detector. For each thermal image, the backbone can also extract some visible features for better performance because of our additional auxiliary task. (2) For *visual knowledge adaptation* methods which predicts the visible features, our methods also obtain better performance by predicting conditional visible distribution. This is because these visual knowledge adaptation methods only learn the knowledge of a specific visible image corresponding to the paired thermal image, however, our method implicitly learns more similar visible samples by regressing visible feature distribution. Meanwhile, we choose to predict the conditional visible distribution instead of searching the visible features for fusion. This is because

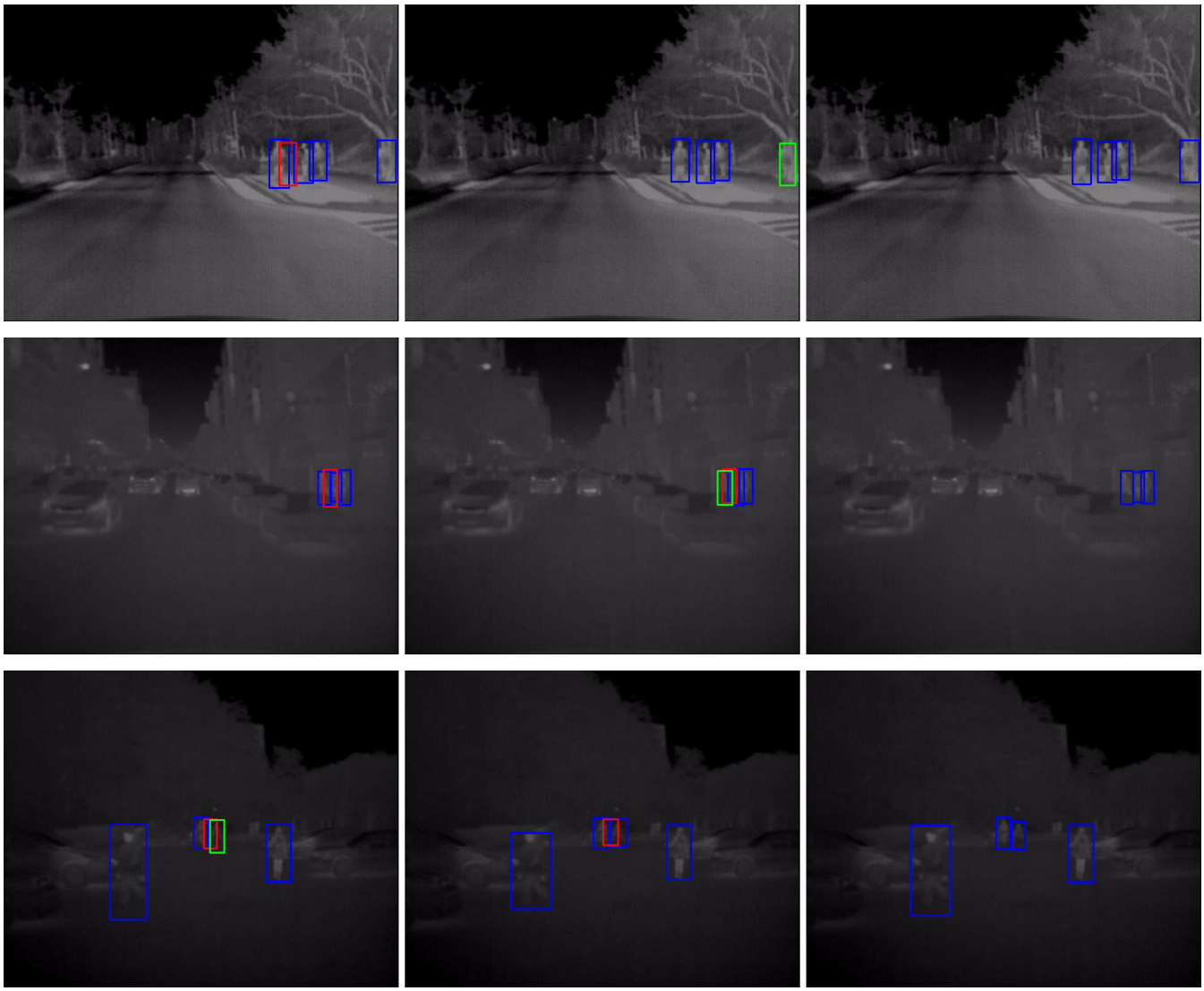

Fig. 4. Pedestrian detection results on the KAIST dataset. The first, second and third columns are the results of YOLOv3, TC Det and YOLOv3+IDA (Ours) respectively. Blue, green and red boxes denote the true positive detections, false negatives, and false positives, respectively. The first, second and third rows are results of day, night and ambiguous samples respectively. Images are cropped to save space when displayed here.

prediciting full dimensional features is hard and needs much computation. (3) Furthermore, our method gets better performance than the state-of-the-art *illumination-aware* methods. Because our method can distinguish day, night, and ambiguous samples, more discriminate features and visible knowledge are extracted. Other methods try to distinguish ambiguous samples as day or night, which leads to bad performance.

3) Visualization Comparison: We print the prediction results and annotations on some images and save them for visual comparison with those open-source methods. The method YOLOv3 + IDA (Ours) is compared with YOLOv3 and TC Det on the KAIST dataset. As shown in Fig.4, our method can reduce false positive predictions and improve true positive predictions for day, night, and ambiguous samples. All of the visual results confirm that the detector with our proposed plug-in can get better performance.

B. Further Studies on YOLOv3+IDA

In this section, we give further studies of our method based on YOLOv3 backbone.

TABLE IV

ABLATION STUDY OF YOLOV3 ON KAIST (LEFT) AND FLIR-ALIGNED (RIGHT) DATASETS. M AND V DENOTE WHETHER THE METHOD PREDICTS MEAN AND VARIANCE VECTORS RESPECTIVELY

M	V	All	Day	Night	mAP	mAP50	mAP75
×	×	25.04	31.89	10.45	35.4	71.4	29.9
✓	×	24.22	30.54	10.36	35.5	**75.7**	27.2
×	✓	24.48	32.06	9.48	36.4	74.2	**31.5**
✓	✓	**22.53**	**28.46**	**9.27**	**37.5**	75.0	30.3

1) Ablation Studies: To explore the effectiveness of the components of our method, we conduct an ablation study on the KAIST dataset and FLIR-aligned dataset. On the KAIST dataset, as shown in Table IV, compared with the baseline YOLOv3, our method reduces the MR by -2.51% from 25.04% to 22.53%. Only predicting mean vector or variance vector will reduce -0.82% from 25.04% to 24.22% or -0.56% from 25.04% to 24.48% respectively. Predicting both mean and variance vectors improves performance better because using both mean and variance can accurately describe

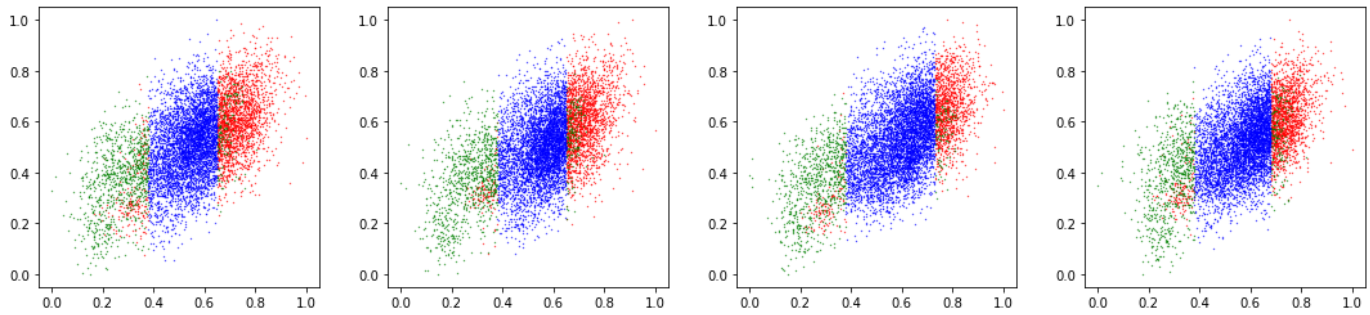

Fig. 5. Feature distribution of thermal and visible images. Red, green and blue points denote the day, night and ambiguous samples respectively. Note that blue points are red and green points mixed together and can be regarded as ambiguous samples. We draw them blue for distinguishing.

TABLE V
DETECTION RESULTS (MR) OF YOLOV3+IDA WITH VARYING K AND N ON KAIST DATASET

K	All	Day	Night	N	All	Day	Night
8	24.48	31.00	10.56	10	24.95	31.86	11.01
16	23.90	30.78	9.68	25	23.92	30.20	11.61
32	**22.53**	**28.46**	**9.27**	40	**22.53**	**28.46**	**9.27**
64	23.92	30.87	9.34	55	23.63	30.32	10.48
128	24.65	31.98	9.68	70	23.25	30.28	9.95

TABLE VI
DETECTION RESULTS (MAP) OF YOLOV3+IDA WITH VARYING K AND N ON FLIR-ALIGNED DATASET

K	mAP	mAP50	mAP75	N	mAP	mAP50	mAP75
8	34.6	73.8	27.0	10	33.2	70.4	25.4
16	32.5	70.2	26.4	25	34.6	73.9	27.4
32	**37.5**	75.0	**30.3**	40	**37.5**	**75.0**	**30.3**
64	35.4	75.2	27.7	55	33.8	73.1	26.3
128	35.6	**75.8**	27.1	70	36.7	**75.0**	29.0

a Gaussian distribution. Similarly results can be obtained based on mAP on the FLIR-aligned dataset.

2) Length of the Mean and Variance Vector: On the KAIST dataset, to explore the effectiveness of the length of mean and variance vectors, we let the value of K be 8, 16, 32, 64 and 128 respectively. As shown in Table V, the detector performs best when $K = 32$. The low value of K means that the information losses more in the process of feature dimensionality reduction; while if the value of K is high, predicting high dimensional mean and variance vectors will be difficult, which also leads to bad performance. So we choose $K = 32$. On the FLIR-aligned dataset, as shown in Table VI, the same result happens. It shows the setting of K is stable across different datasets.

3) Size of the Conditional Feature Co-Occurrence Matrix: On the KAIST dataset, to explore the effectiveness of the size of the conditional feature co-occurrence matrix, we let the value of N be 10, 25, 40, 55 and 70 respectively. As shown in Table V, the detector performs best when $N = 40$. The low value of N leads to the matrices cannot accurately represent the joint Gaussian distribution, because two values with large difference may be quantized into a same interval. The high value of N leads to the matrices being sparse. Since the total number of samples is a constant number N_s, too few samples will be allocated to each matrix row. So frequency cannot approximate a probability anymore. Similarly as the parameter K, the same thing applies to FLIR-aligned dataset as shown in Table VI.

4) Parameter Sensitivity Analysis: The parameter sensitive analysis of λ_m and λ_v in Eq.(13) is performed on the KAIST dataset based on the method YOLOv3+IDA. The value of λ_m is selected as 25.0, 50.0, 75.0, 100.0 and 125.0 when $\lambda_v = 200.0$. Similarly, λ_v is selected as 100.0, 150.0, 200.0, 250.0 and 300.0 when $\lambda_m = 75.0$.

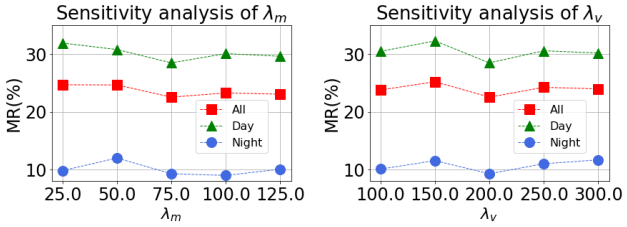

Fig. 6. Hyperparameter sensitive analysis of λ_m and λ_v on the KAIST dataset.

As shown in Fig.6, the performance is the best when $\lambda_m = 75.0$ and $\lambda_v = 200.0$. Meanwhile, our model can keep a relatively stable result in a wide value range respect to λ_m and λ_v. The low value of λ_m and λ_v means that the detector cannot accurately distinguish day, night, and ambiguous samples without using the conditional probability distribution regression task; while the high value of λ_m and λ_v leads to bad performance because of insufficient constrain on detection loss L_{OD}. It can also be observed that our method can achieve better improvements in the day. This further proves that our method can effectively obtain the visible features to compensate the information loss in thermal features. Similarly on the FLIR-aligned dataset, we let the value of λ_m be 50.0, 75.0, 100.0, 125.0 and 150.0, and λ_v be 100.0, 150.0, 200.0, 250.0 and 300.0, respectively. As shown in Table 7, the detector performs best when $\lambda_m = 100.0$ and $\lambda_v = 200.0$.

5) Visualization of Feature Distribution: In this part, based on the KAIST dataset, more visualization results based on YOLOv3+IDA are shown here. For the i-th thermal-visible image pair in the training dataset, we calculate the normalized feature vectors $\widetilde{f}_i^{T/V}$, which are K-length vectors. For $j = 1, \cdots, K$, we denote by a point at coordinates $(\widetilde{f}_i^T[j], \widetilde{f}_i^V[j])$ for each sample in the training dataset. Here, we get K

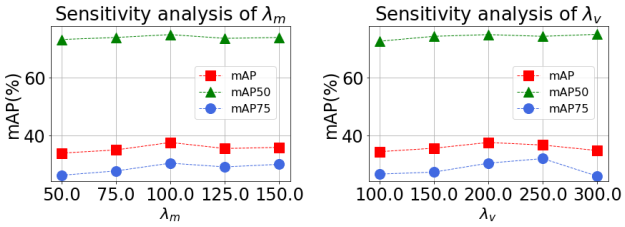

Fig. 7. Hyperparameter sensitive analysis of λ_m and λ_v on the FLIR-aligned dataset.

TABLE VII

ABLATION STUDY OF FASTERRCNN ON KAIST (LEFT) AND FLIR-ALIGNED (RIGHT) DATASETS. M AND V DENOTE WHETHER THE METHOD PREDICTS MEAN AND VARIANCE VECTORS RESPECTIVELY

M	V	All	Day	Night	mAP	mAP50	mAP75
×	×	21.07	26.83	8.56	38.9	74.3	34.3
✓	×	20.06	26.44	6.91	39.9	**75.5**	35.8
×	✓	20.86	26.88	7.67	40.0	75.3	**35.9**
✓	✓	**18.88**	**24.97**	**6.61**	**40.3**	75.4	35.8

visualization results and randomly select four to show here. As shown in Fig. 5, samples are distributed as day, night, and ambiguous cases instead of simple day and night two categories. This shows that the conditional distributions of day, night, and ambiguous samples are different. Therefore, predicting the distribution can make the backbone learn specific visible features, improving performance.

C. Further Studies on FasterRCNN+IDA

To prove the versatility of IDA as a plug-in, in this subsection, we give more analysis on FasterRCNN, which is a two-stage model different from YOLOv3 as a single-stage model.

1) Ablation Studies: To explore the effectiveness of our method, we conduct an ablation study on the KAIST dataset. As shown in Table VII (left), our method reduces the MR by -2.19% from 21.07% to 18.88%. Only predicting mean vector or variance vector reduces -1.06% from 21.07% to 20.06% or -0.21% from 21.07% to 20.86%, respectively. Predicting both mean and variance vectors improves performance better because only using mean and variance can accurately describe the Gaussian distribution. The same applies to FLIR-aligned dataset. As shown in Table VII (right), with IDA, our method improves by $+1.4\%$ from 38.9% to 40.3%.

2) Length of the Mean and Variance Vectors: On the KAIST dataset, to explore the effectiveness of the length of the mean and variance vector, We modify the value of K as $\{8, 16, 32, 64, 128\}$ respectively. As shown in Table VIII, the detector performs best when $K = 32$. The low value of K means that the feature losses too much in the process of feature dimensionality reduction. If the value of K is high, predicting the mean and variance vectors will be difficult, which leads to bad performance. The same applies to FLIR-aligned dataset as shown in Table VIII.

3) Size of the Conditional Feature Co-Occurrence Matrices: On the KAIST dataset, to explore the effectiveness of the size of the conditional feature co-occurrence matrices, we modify

TABLE VIII

DETECTION RESULTS (MR) OF FASTERRCNN WITH VARYING K AND N ON KAIST DATASET

K	All	Day	Night	N	All	Day	Night
8	20.58	26.33	9.50	10	19.83	25.06	8.55
16	20.84	27.13	8.13	25	21.08	26.71	9.45
32	**18.88**	**24.97**	**6.61**	40	**18.88**	**24.97**	**6.61**
64	21.03	26.86	8.48	55	19.79	25.82	8.17
128	21.05	27.03	9.27	70	21.27	27.55	8.60

TABLE IX

DETECTION RESULTS (MAP) OF FASTERRCNN WITH VARYING K AND N ON FLIR-ALIGNED DATASET

K	mAP	mAP50	mAP75	N	mAP	mAP50	mAP75
8	39.4	74.9	34.7	10	39.5	75.1	35.1
16	39.1	**75.7**	34.3	25	39.4	75.2	34.9
32	**40.3**	75.4	**35.8**	40	**40.3**	**75.4**	**35.8**
64	39.3	74.2	35.0	55	39.7	75.2	35.7
128	39.8	74.9	34.9	70	39.0	74.0	34.9

the value of N by changing $\{10, 25, 40, 55, 70\}$. As shown in Table VIII, the detector performs best when $N = 40$. The low value of N leads to the matrices cannot accurately represent the joint Gaussian distribution, because two values with large difference maybe be quantized into the same interval. The high value of N leads to the matrices being sparse. Since the total samples are a constant number N_s, the number assigned to a specific row of the matrix will be small. Therefore, frequency cannot be used to approximate a probability. The same applies to FLIR-aligned dataset as shown in Table IX.

4) Sensitivity Analysis: We perform analysis on λ_m and λ_v on the KAIST dataset. (1) We modify the value of λ_m by changing $\{0.05, 0.075, 0.1, 0.125, 0.15\}$ when $\lambda_v = 0.125$. As shown in the left of Fig.8, the performance is the best when $\lambda_m = 0.1$. Our model can keep a relatively stable result in a wide range of λ_m. Besides, the low-value λ_m means that the detector cannot accurately distinguish day, night, and ambiguous samples by predicting the mean of the conditional probability distribution. The high-value λ_m leads to bad performance because of insufficient training for detection L_{OD}. (2) We modify the value of λ_v by changing $\{0.075, 0.1, 0.125, 0.15, 0.175\}$ when $\lambda_m = 0.1$. As shown in the right of Fig.8, the performance is the best when $\lambda_v = 0.125$. Our model can keep a relatively stable result in a wide range of λ_v. Besides, the low-value λ_v means that the detector cannot accurately distinguish day, night, and ambiguous samples by predicting the variance of the conditional probability distribution. The high-value λ_v leads to bad performance because of insufficient training for detection L_{OD}. The same applies to FLIR-aligned dataset as shown in Fig. 9.

D. Noise Impact

This part aims to explore the impact of noise on the feature co-occurrence matrices. In the FCM module, we add Gaussian noise on the feature vectors based on Autonomous Vehicles dataset. The mean of Gaussian noise is zero. The standard deviation of Gaussian noise is sampled from the set $\{0.001,$

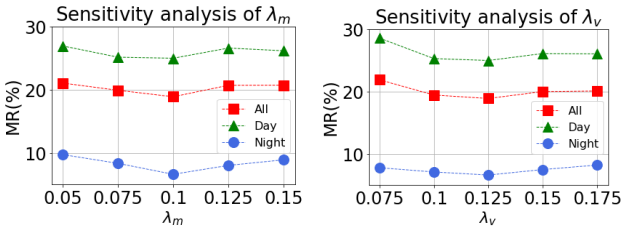

Fig. 8. Hyperparameter analysis for FasterRCNN on the KAIST dataset: (left) λ_m and (right) λ_v.

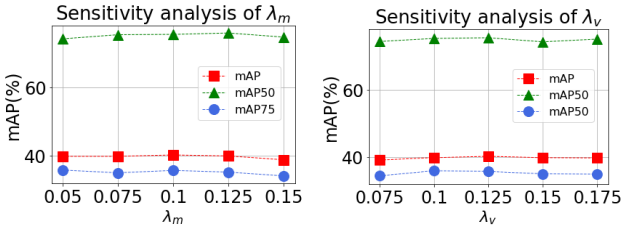

Fig. 9. Hyperparameter analysis for FasterRCNN on the FLIR-aligned dataset: (left) λ_m and (right) λ_v.

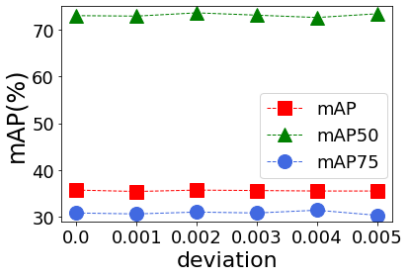

Fig. 10. Performance impacts on YOLOv3+IDA with different standard deviation Gaussian noises in Autonomous Vehicles dataset.

0.002, 0.003, 0.004, 0.005}. As shown in Fig.10, the results are stable with few impacts of noise. This is because feature values are discretized. Small value changes hardly affect the discretization results. While for a few changed values, since there are many samples in the row of the matrices, the distribution is also stable. Therefore, our method is robust for a small amount of noise.

E. Generalization Capability

This part aims to explore whether the feature co-occurrence matrices constructed in a dataset can be used to another dataset. We use YOLOv3 as the detector. The feature co-occurrence matrices are calculated in KAIST dataset. The test dataset is FLIR-aligned dataset. As shown in Table X, the detector also improves mAP in FLIR-aligned dataset using co-occurrence matrices from another dataset. This is because for day, night, and ambiguous samples, the conditional visible features can be obtained although they are not very correct, which also enhance discriminality for better performance. This experiment also verifies that the constructed co-occurrence matrices have a certain degree of generalization ability. Of course, Table XI also shows using the co-occurrence matrices from the same dataset is the best. How to transfer the co-occurrence matrices to a new target dataset is an intereing problem.

TABLE X
DETECTION RESULTS (MAP) OF YOLOV3 WITH DIFFERENT FEATURE CO-OCCURRENCE MATRICES ON FLIR-ALIGNED DATASET

method	mAP	mAP50	mAP75
Baseline	35.4	71.4	29.9
KAIST $\rightarrow$ FLIR-aligned	36.9	**75.2**	30.1
FLIR-aligned $\rightarrow$ FLIR-aligned	**37.5**	75.0	**30.3**

F. Limitations

As described in the algorithm, as many SOTA methods, our method also needs paired visible-thermal images at training phase. However, the paired visible-thermal images are hard to be acquired and aligned. Therefore, developing a method that uses unpaired visible and thermal images or large models is an important future research direction. Besides, previous methods need two backbones for visible and thermal domains respectively, which lead to more parameters for training. For simplicity, using a pre-trained large model such as the Segment Anything Model (SAM) [60] for obtaining visible information is a promising way.

V. CONCLUSION

We proposed a novel method, dubbed as Illumination Distribution-Aware adaptation (IDA), which can distinguish day, night and ambiguous samples without extra illumination annotations and implicitly extract the compensated visible features. Based on this strategy, conditional feature co-occurrence matrices are proposed, which record the visible feature distribution given a thermal feature. The detector can distinguish day, night and ambiguous samples by querying and predicting the conditional distribution. In this way, the extracted features are also implicitly aligned to the visible features. Experiments confirm the effectiveness of our method. Theoretically, as feature co-occurrence matrices have some generalization capability, our method can also be extended to more complex scenarios, such as the combination of more categories and different shooting time.

REFERENCES

[1] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4913–4934, Sep. 2022.

[2] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network for occluded pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4967–4975.

[3] L. Zhang et al., "Cross-modality interactive attention network for multispectral pedestrian detection," *Inf. Fusion*, vol. 50, pp. 20–29, Oct. 2019.

[4] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, pp. 161–171, Jan. 2019.

[5] 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," *Inf. Fusion*, vol. 50, pp. 148–157, Oct. 2019.

[6] Y. Gong, L. Wang, and L. Xu, "A feature aggregation network for multispectral pedestrian detection," *Int. J. Speech Technol.*, vol. 53, no. 19, pp. 22117–22131, Oct. 2023.

[7] C. Herrmann, M. Ruf, and J. Beyerer, "CNN-based thermal infrared person detection by domain adaptation," *Proc. SPIE*, vol. 10643, pp. 38–43, May 2018.

[8] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 988–997.

[9] Z. Xu, C.-M. Vong, C.-C. Wong, and Q. Liu, "Ground plane context aggregation network for day-and-night on vehicular pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6395–6406, Oct. 2021.

[10] J. U. Kim, S. Park, and Y. M. Ro, "Robust small-scale pedestrian detection with cued recall via memory learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3050–3059.

[11] F. Bongini, L. Berlincioni, M. Bertini, and A. D. Bimbo, "Partially fake it till you make it: Mixing real and fake thermal images for improved object detection," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5482–5490.

[12] M. Kieu, A. D. Bagdanov, M. Bertini, and A. D. Bimbo, "Domain adaptation for privacy-preserving pedestrian detection in thermal imagery," in *Proc. 20th Int. Conf. Image Anal. Process. (ICIAP)*, Trento, Italy. Springer, Sep. 2019, pp. 203–213.

[13] M. Kieu, A. D. Bagdanov, and M. Bertini, "Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1, pp. 1–19, Feb. 2021.

[14] M. Kieu, L. Berlincioni, L. Galteri, M. Bertini, A. D. Bagdanov, and A. del Bimbo, "Robust pedestrian detection in thermal imagery using synthesized images," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8804–8811.

[15] T. Guo, C. P. Huynh, and M. Solh, "Domain-adaptive pedestrian detection in thermal images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1660–1664.

[16] S. Park, D. Hwi Choi, J. Uk Kim, and Y. M. Ro, "Robust thermal infrared pedestrian detection by associating visible pedestrian knowledge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4468–4472.

[17] Q. Xie, T.-Y. Cheng, Z. Dai, V. Tran, N. Trigoni, and A. Markham, "Illumination-aware hallucination-based domain adaptation for thermal pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 315–326, Jan. 2024.

[18] M. Kieu, A. D. Bagdanov, M. Bertini, and A. D. Bimbo, "Task-conditioned domain adaptation for pedestrian detection in thermal imagery," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 546–562.

[19] J. Baek, J. Hyun, and E. Kim, "A pedestrian detection system accelerated by kernelized proposals," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 1216–1228, Mar. 2020.

[20] C. Zhou, M. Wu, and S.-K. Lam, "Group cost-sensitive BoostLR with vector form decorrelated filters for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 12, pp. 5022–5035, Dec. 2020.

[21] X. Liu, K.-A. Toh, and J. P. Allebach, "Pedestrian detection using pixel difference matrix projection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1441–1454, Apr. 2020.

[22] J. Shen, X. Zuo, L. Zhu, J. Li, W. Yang, and H. Ling, "Pedestrian proposal and refining based on the shared pixel differential feature," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2085–2095, Jun. 2019.

[23] P. Yang, G. Zhang, L. Wang, L. Xu, Q. Deng, and M.-H. Yang, "A part-aware multi-scale fully convolutional network for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1125–1137, Feb. 2021.

[24] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2015, pp. 3361–3369.

[25] J. Cao, Y. Pang, and X. Li, "Learning multilayer channel features for pedestrian detection," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3210–3220, Jul. 2017.

[26] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 732–747.

[27] C. Lin, J. Lu, and J. Zhou, "Multi-grained deep feature learning for robust pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3608–3621, Dec. 2019.

[28] Y. Luo, C. Zhang, W. Lin, X. Yang, and J. Sun, "Sequential attention-based distinct part modeling for balanced pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15644–15654, Sep. 2022.

[29] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.

[30] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4236–4244.

[31] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5127–5137.

[32] J. Xie et al., "Learning a dynamic cross-modal network for multispectral pedestrian detection," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4043–4052.

[33] K. Dasgupta, A. Das, S. Das, U. Bhattacharya, and S. Yogamani, "Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15940–15950, Sep. 2022.

[34] R. Li, J. Xiang, F. Sun, Y. Yuan, L. Yuan, and S. Gou, "Multiscale cross-modal homogeneity enhancement and confidence-aware fusion for multispectral pedestrian detection," *IEEE Trans. Multimedia*, vol. 26, pp. 852–863, 2023.

[35] X. Yang, Y. Qian, H. Zhu, C. Wang, and M. Yang, "BAANet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2920–2926.

[36] K. Chen, J. Liu, and H. Zhang, "IGT: Illumination-guided RGB-T object detection with transformers," *Knowl.-Based Syst.*, vol. 268, May 2023, Art. no. 110423.

[37] Y. Zhang, H. Yu, Y. He, X. Wang, and W. Yang, "Illumination-guided RGBT object detection with inter- and intra-modality fusion," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[38] J. U. Kim, S. Park, and Y. M. Ro, "Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, pp. 1157–1165.

[39] R. J. Jevnisek and S. Avidan, "Co-occurrence filter," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3816–3824.

[40] R. Kat, R. Jevnisek, and S. Avidan, "Matching pixels using co-occurrence statistics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1751–1759.

[41] G. Xiao, S. Wang, H. Wang, and J. Ma, "Mining consistent correspondences using co-occurrence statistics," *Pattern Recognit.*, vol. 119, Nov. 2021, Art. no. 108062.

[42] I. Shevlev and S. Avidan, "Co-occurrence neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4792–4799.

[43] M. Chen, T. Ma, and X. Zhou, "CoCNN: Co-occurrence CNN for recommendation," *Exp. Syst. Appl.*, vol. 195, Jun. 2022, Art. no. 116595.

[44] Y.-F. Shih, Y.-M. Yeh, Y.-Y. Lin, M.-F. Weng, Y.-C. Lu, and Y.-Y. Chuang, "Deep co-occurrence feature learning for visual object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7302–7311.

[45] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[47] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[48] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.

[49] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," 2018, *arXiv:1808.04818*.

[50] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," 2016, *arXiv:1611.02644*.

[51] F. Group. (May 2019). *Flir Thermal Dataset for Algorithm Training*. [Online]. Available: https://www.flir.co.uk/oem/adas/adas-dataset-form/

[52] K. Takumi, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada, "Multispectral object detection for autonomous vehicles," in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 35–43.

[53] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[54] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[56] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
[57] J. Baek, S. Hong, J. Kim, and E. Kim, "Efficient pedestrian detection at nighttime using a thermal camera," *Sensors*, vol. 17, no. 8, p. 1850, Aug. 2017.
[58] Y. Hu, N. Zhang, and L. Weng, "Retrieve the visible feature to improve thermal pedestrian detection using discrepancy preserving memory network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 1125–1129.
[59] Z. Wang, F. Colonnier, J. Zheng, J. Acharya, W. Jiang, and K. Huang, "TIRDet: Mono-modality thermal InfraRed object detection based on prior thermal-to-visible translation," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 2663–2672.
[60] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

Songtao Li received the B.S. degree in computer science and technology from the University of Electronic Science and Technology of China, Chengdu, in 2021, where he is currently pursuing the M.S. degree. His research interests include computer vision, deep learning, and object detection.

Mao Ye (Senior Member, IEEE) received the B.S. degree in mathematics from Sichuan Normal University, Chengdu, China, in 1995, the M.S. degree in mathematics from the University of Electronic Science and Technology of China, Chengdu, in 1998, and the Ph.D. degree in mathematics from The Chinese University of Hong Kong, China, in 2002. He has been a short-time Visiting Scholar with The University of Queensland, and the University of Pennsylvania. He is currently a Professor and the Director of the CVLab, University of Electronic Science and Technology of China. His research interests include machine learning and computer vision. In these areas, he has published over 90 papers in leading international journals or conference proceedings. He has served on the editorial board of *Engineering Applications of Artificial Intelligence*. He was a co-recipient of the Best Student Paper Award at the IEEE ICME 2017.

Luping Ji (Member, IEEE) received the B.S. degree in mechanical and electronic engineering from Beijing Institute of Technology, Beijing, China, in 1999, and the M.S. and Ph.D. degrees in computer application and technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2005 and 2008, respectively. He is currently a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include neural networks and pattern recognition.

Song Tang (Member, IEEE) received the B.S. degree in electronic engineering from Hainan University, Haikou, China, in 2005, the M.S. degree in control engineering from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2013, and the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2017. From 2005 to 2009, he was a Senior Engineer with Shanghai Mitsubishi Elevator Company Ltd. (SMEC). From 2017 to 2019, he was a Research Associate with the Department of Informatics, Universitét Hamburg, Hamburg, Germany. He is currently an Associate Professor with the IMI, University of Shanghai for Science and Technology, Shanghai, China. He has authored or co-authored more than 50 papers in journals and conferences. His current research interests include machine learning, computer vision, and robot learning.

Yan Gan (Member, IEEE) received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, China, in 2020. He is currently a Research Assistant with the School of Computer Science, Chongqing University, Chongqing, China. His current research interests include generative models and computer vision.

Xiatian Zhu received the Ph.D. degree from the Queen Mary University of London. He was a Research Scientist with the Samsung AI Centre, Cambridge, U.K. He is currently a Senior Lecturer with the Surrey Institute for People-Centred Artificial Intelligence and the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, U.K. His research interests include computer vision and machine learning. He won the Sullivan Doctoral Thesis Prize 2016.