# Thermal Pedestrian Detection Based on Different Resolution Visual Image

**Songtao Li**
University of Electronic Science and Technology of China
**Jinzhong Cui**
University of Electronic Science and Technology of China
**Mao Ye**
University of Electronic Science and Technology of China
**Ting Li** ( tingli@std.uestc.edu.cn )
University of Electronic Science and Technology of China
**Liang Tian**
University of Electronic Science and Technology of China

**Research Article**

**Keywords:** Domain Adaptation, Pedestrian Detection, Low-Resolution Thermal Images, Disentanglement

**Posted Date:** March 15th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2676851/v1



**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at Signal, Image and Video Processing on July 26th, 2023. See the published version at https://doi.org/10.1007/s11760-023-02667-z.

# Thermal Pedestrian Detection Based on Different Resolution Visual Image

Songtao Li[1], Jinzhong Cui[2], Mao Ye[3], Ting Li[4*] and Liang Tian[5]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.

[2]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.

[3]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.

[4*]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.

[5]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China.

*Corresponding author(s). E-mail(s): tingli@std.uestc.edu.cn;
Contributing authors: 2268172676lst@gmail.com;
jzcui@uestc.edu.cn; cvlab.uestc@gmail.com; tbeatl@163.com;

**Abstract**

Thermal pedestrian detection is a core problem in computer vision. Usually, the corresponding visual image knowledge are used to improve the performance in thermal domain. However, existing methods always assume the same resolution between visible and thermal images. But in reality, there is a problem with this setting. Since thermal imaging acquisition equipment is expensive, the resolution of thermal images is always lower than visible images. To address this issue, we propose a new method, named as Disentanglement Then Restoration (DTR).

The key idea is to disentangle the features into content features and modal features, and restore the complete content features of thermal images by learning the changes of content features caused by different resolutions. Specifically, we first train an object detector such as YOLO to initialize our model. Then, a feature disentanglement network is trained, which can disentangle the features from the backbone as content features and modal features. In the end, the feature disentanglement network is frozen. By forcing the content feature consistency between visual image and upsampled thermal image, the complete content features of low-resolution thermal images are restored. Experiment results on public datasets show that our method performs very well.

**Keywords:** Domain Adaptation, Pedestrian Detection, Low-Resolution Thermal Images, Disentanglement

# 1 Introduction

Pedestrian detection is an important problem in computer vision [1], which plays an important role in autonomous driving, surveillance and other security fields. There are pedestrian detection methods based on visible images [2–6]. But these detectors will fail in the insufficient light (nighttime) or bad weather (rain) cases. There also exist some methods focusing on multispectral pedestrian detection [7–13]. However, in actual use, only thermal imaging sensors exist, for example, privacy concerns. So pedestrian detection in only thermal images has attracted many attentions. The existing works can be roughly divided into two categories. The first approach tries to design new network structure for better performance [14, 15]. The key to success is how to extract features unique to thermal images, which itself is sufficiently challenging and not well solved. Another approach utilizes the particularity of thermal images [16–18]. But existing datasets lack descriptions of image devices, which makes it difficult to mine the special properties of thermal images.

Since thermal images lack details compared with the visual images, therefore, some works try to use visible domain knowledge to train the detector; while at the inference phase, the detector is used without any visual images [19–21]. In this route, the current methods can be roughly divided into two categories. The first category is image adaptation [14, 21, 22]. Since visible images generally have more information than thermal images, a generator model can be trained to transform thermal images into visible images. This method requires high quality generator. Another category is feature adaptation [19, 20]. Visual images are used to teach the detector to extract object-specific features at the training phase which is lost in the thermal domain.

All existing methods assume the same resolution of thermal and visible images as shown in Fig.1(a). Since the thermal camera equipment is expensive, the resolution of thermal images is always lower than the resolution of visible images. So the true situation is shown in Fig.1(b), which makes the existing

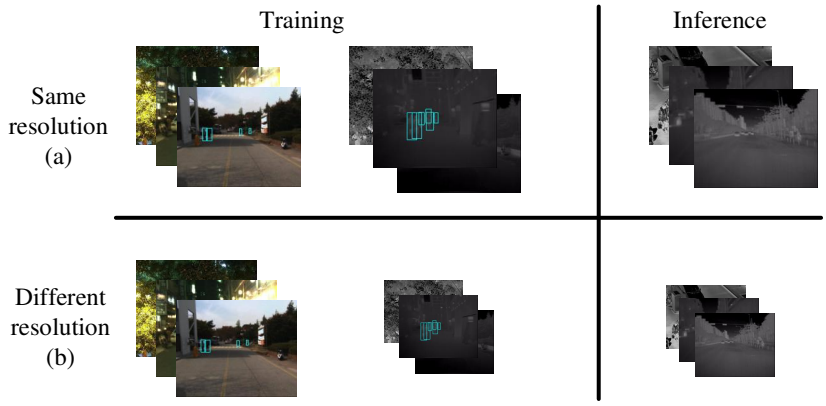


**Fig. 1** Comparisons between previous assumption (a) and true situation (b). For training, both high-resolution visible images and low-resolution thermal images can be used; while for inference, pedestrian detection is performed only in low-resolution thermal images.

methods hard to work for unaligned images and features. Based on the observation that the features can be disentangled to two parts: content features and modal features. Content features mean the domain invariant features, which are the same between the paired visible and thermal images. Without loss of generality, we can consider that high-resolution visible and low-resolution thermal images have complete and incomplete content features, respectively. So it is natural to require the content features of thermal image to be consistent with the visual image to learn the information lost in the low resolution thermal image.

Based on the above motivations, we propose a novel method, dubbed as Disentanglement Then Restoration (DTR). The key idea is to learn a disentanglement network and then use it to extract content features, and restore lost information by forcing content feature consistency. Specifically, our method contains three steps: initialization, disentanglement, and restoration. At the initialization step, the up-sampled thermal images, the down-sampled and then up-sampled visible images, and visible images are used to train a YOLOv3 detector. Here, up-sample means changing the low-resolution to high-resolution respect to thermal image and visual image. Down-sample is just the opposite. At the second step, similar as the traditional disentanglement learning [23–26], the feature disentanglement network is learned by minimizing the content feature difference between the up-sampled thermal images and the down-sampled and then up-sampled visible images abbreviated as up-down sampled image. In the end, feature disentanglement network is frozen. We train the YOLOv3 backbone by minimizing the content feature difference between the up-sampled thermal images and visible images to learn to restore lost information in thermal image.

Our contributions are three-folds: (1) We proposed a new problem setting which exists in practical applications. Traditional methods cannot be well

applied because the visual and thermal images can not be well aligned due to different resolution problem. (2) A disentanglement then restoration method is proposed. By requiring content feature consistency, detector network can learn to restore lost information. (3) Experiments on public datasets confirm the effectiveness of our method.

## 2 Related Works

**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 [27], Cao et al. [28], DMP [29] and Shen et al. [30]. The second one uses CNN for detection, such as PAMS-FCN [31], CompACT [32], and MCF [33]. They modify the classic object detection model for pedestrian detection. The third one uses attention for detection, such as GDFL [34] and MDFL [35]. 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 [36], SA-DPM [37] and Zhang et al. [38]. The final one is domain adaptation, which uses thermal images to auxiliary train a detector for visible pedestrian detection, such as CMT-CNN [39].

**Pedestrian detection in multispectral images.** Since visible and thermal images have complementary information, multispectral pedestrian detection is also an important problem. There are three main routes to address this problem. The first one is feature fusion, which uses CNN such as Konig et al. [7], CMPD [40], Kim et al.[9] and AR-CNN [41] or attention such as CIAN [11] and Dasgupta et al. [10] to combine both visible and thermal features. The second one is based on illumination-aware modules such as IAF R-CNN [12], Guan et al.[13] and MBNet [42], which address modality imbalance problems because different illumination conditions can affect the feature distribution of visible and thermal images. The final one considers unsupervised domain adaptation for multispectral pedestrian detection such as UMDA [43] and TS-RPN [44].

**Pedestrian detection in thermal images.** Due to insufficient light (nighttime) or bad weather (rain), many methods focus on pedestrian detection in thermal images. Thermal pedestrian detection contains three strategies. The first one is single domain methods such as Ghose et al. [15], GPCAnet [16], Kim et al. [17] and TCDet [18], which use only thermal images for training and testing. The second one is domain adaptation. They transfer the visible domain into the thermal domain by image adaptation such as Guo et al. [22] and Kieu et al. [21] or feature adaptation such as Herrmann [14], Kieu et al. [19], Kieu et al. [20] and Kim et al.[45]. The final one is unsupervised domain adaptation, such as Meta-UDA[46].

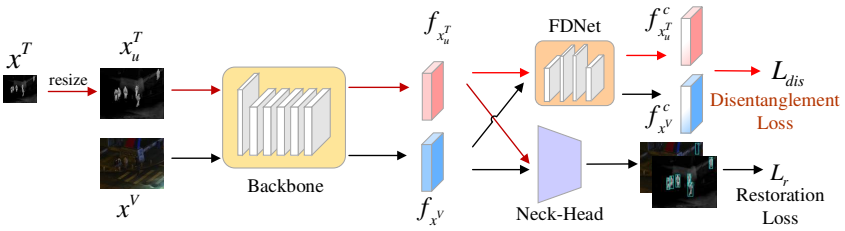


**Fig. 2** Our network structure based on YOLOv3. Backbone receives images and returns features. The feature disentanglement network (FDNet) receives features to extract content features. The Neck-Head network receives features to predict the boxes of pedestrians.

**Disentanglement Learning.** Disentanglement learning aims to decouple features, which has been widely used in domain adaptation for various computer vision tasks such as DDF [23], DDOD [24], DIDN [25] and Wu et al. [26]. Methods mostly focus on three areas. The first one is image classification such as ZSDA [47] and FDH-Net [48]. The second one is object detection such as Tang et al. [49] and VDD [50]. The final one is other tasks such as DRL-Net for person re-identification [51] and Lee et al. for image de-raining [52]. However, as far as we know, there are currently no disentanglement learning methods for thermal pedestrian detection.

# 3 The Proposed Method

## 3.1 Problem statement

Let $D_s = \{(x_{s,i}^V, x_{s,i}^T, y_{s,i}^V, y_{s,i}^T)\}_{i=1}^{N_s}$ denote the training dataset. $x_{s,i}^V \in \mathbf{R}^{W\times H\times c}$, denotes the $i$-th visible image, where $W$, $H$, and $c$ denote the width, height, and channel respectively. The paired $i$-th thermal image $x_{s,i}^T \in \mathbf{R}^{w\times h\times c}$, where $w$ and $h$ denote the width and height respectively. Here, $W > w$ and $H > h$. $y_{s,i}^V$ and $y_{s,i}^T$ denotes the annotation of the $i$-th paired visible-thermal images. Suppose the test dataset is $D_t = \{x_{t,j}^T\}_{j=1}^{N_t}$, our goal is to train an object detector on the thermal image with the help of visual image knowledge such that it can achieve better results on the $D_t$ data set than the object detector trained on the thermal images alone.

**Overall architecture.** The overall architecture of the proposed pedestrian detection framework is shown in Fig. 2. Our network is based on YOLOv3 [53]. The proposed DTR method consists of three steps: Initialization, Disentanglemet, and Restoration. At the initialization step, the up-sampled thermal image $x_u^T$, the corresponding high-resolution image $x^V$, and the up-down sampled image $x_{ud}^V$ are used to train an initial detector. The backbone network is Darknet53 and the neck-head network receives features and returns the detection results. Then, a feature disentanglement network is trained by requiring the content feature consistency between $x_{ud}^V$ and $x_u^T$. In the end, by using the frozen feature disentanglement network and requiring the content feature

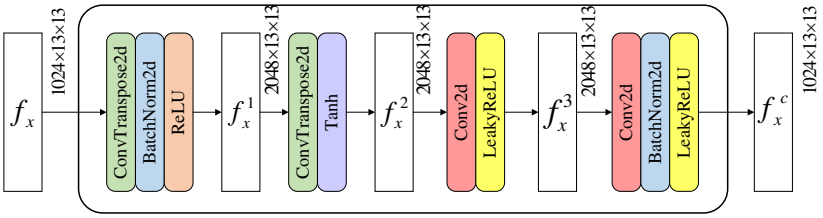


**Fig. 3** The proposed feature disentanglement network. The width and height of feature map are both 13. After the first ConvTranspose2d, the feature channel number is increased from 1024 to 2048. And after the first Conv2d, the channel of the feature is from 2048 to 1024.

consistency between $x^V$ and $x_u^T$, the backbone network learns to restore the complete content features for $x_u^T$.

## 3.2 Detector Initialization

In order to ensure that the detector can adapt to different modalities (thermal and visible) and image sharpness after scaling by different scales, the up-sampled thermal image $x_u^T$ and the paired visual image $x^V$, and the up-down sampled image $x_{ud}^V$ are used to initialize a YOLOv3 detector. The loss function is defined as follows,

$$L_I = L_D(y_u^T, \hat{y}_u^T) + \lambda_I(L_D(y^V, \hat{y}^V) + L_D(y_{ud}^V, \hat{y}_{ud}^V)), \tag{1}$$

where $y_u^T$, $y^V$ and $y_{ud}^V$ denote the detection ground truth of $x_u^T$, $x^V$, and $x_{ud}^V$, respectively. $\hat{y}_u^T$, $\hat{y}^V$ and $\hat{y}_{ud}^V$ denote the prediction results in $x_u^T$, $x^V$, and $x_{ud}^V$, respectively. Since thermal and visible images are paired, $x_u^T$, $x^V$, and $x_{ud}^V$ share the same ground truth after resizing. $L_D$ is the standard detection loss of YOLOv3. Here, we set $\lambda_I = 0.5$ such that the contributions from visual loss and thermal loss are equal.

The detector is updated once according to the loss $L_I$ after calculating the losses $L_D(y_u^T, \hat{y}_u^T)$, $L_D(y^V, \hat{y}^V)$, and $L_D(y_{ud}^V, \hat{y}_{ud}^V)$ respectively.

## 3.3 Feature Disentanglement

In this part, we aim to train a feature disentanglement network to disentangle the features into content features and modal features where the feature is from YOLOv3 backbone. Different from other disentanglement networks, the feature disentanglement network only has one branch rather than two or more branches, because the DTR method focus on restoring content features. Meanwhile, the existing disentanglement networks are usually based on Multi-Layer Perception (MLP) machine [54], however, different from the traditional classification models, the feature maps output from YOLOv3 backbone is huge whose size is $1024 \times 13 \times 13$. Using MLP requires a lot of computation. Therefore, our

feature disentanglement network is composed of convolution and transposed convolution layers [**?** ] instead of MLP which is shown in Fig.3.

The width and height of input feature maps are small, but the channel number is large, which is similar to the middle layer feature of U-net [55]. Therefore, we use the decoder-encoder architecture of U-net for feature transformation. The feature disentanglement network contains four parts. The first part consists of a transposed convolution layer, a 2D batch normalization, and a ReLU activation function which can be denoted as follows,

$$f_x^1 = (ReLU \circ BN \circ TransConv_{3\times3})(f_x), \tag{2}$$

where $f_x$ denotes the feature map from YOLOv3 backbone. The second part is composed of a transposed convolution layer and a Tanh activation function denoted as

$$f_x^2 = (Tanh \circ TransConv_{3\times3})(f_x^1). \tag{3}$$

The third part has a convolution layer and a LeakyReLU activation function as

$$f_x^3 = (LeakyReLU \circ Conv_{3\times3})(f_x^2), \tag{4}$$

where $f_x^3$ denotes the feature map from the third part. The last part consists of a convolution layer, a 2D batch normalization, and a LeakyReLU activation function as follows,

$$f_x^c = (LeakyReLU \circ BN \circ Conv_{3\times3})(f_x^3), \tag{5}$$

where $f_x^c$ denotes the content feature extracted from the feature disentanglement network.

Since $x_u^T$ and $x_{ud}^V$ have the same content features, the following loss function $L_{dis}$ is used to train the feature disentanglement network,

$$L_{dis} = L_D^c(y_u^T, \hat{y}_u^T) + L_D^c(y_{ud}^V, \hat{y}_{ud}^V) + \lambda_d L_a(f_{x_u^T}^c, f_{x_{ud}^V}^c), \tag{6}$$

where $L_D^c$ is the standard detection loss of YOLOv3 by using feature maps from the feature disentanglement network rather than YOLOv3 backbone. Using $L_D^c$ rather than $L_D$ aims to ensure that the feature disentanglement network returns content features rather than irrelevant features. $L_a(\phi, \psi) = \|\phi - \psi\|_2$, for any two features $\phi$ and $\psi$. The loss function $L_a$ is used to align the content features $f_{x_u^T}^c$ and $f_{x_{ud}^V}^c$.

We set $\lambda_d$ as 10 here. The low-value $\lambda_d$ causes failure to train the feature disentanglement network. While the high-value $\lambda_d$ causes the feature disentanglement network might extract irrelevant features.

## 3.4 Feature Restoration

After the feature disentanglement network is frozen, for the visual image $x^V$, without losing generality, we can assume the feature disentanglement network

**Table 1** Comparisons on the KAIST dataset at day and night in terms of log-average miss rate (lower is better). The best results are highlighted in **bold**.

| Method | Resolution | MR all | MR day | MR night |
|---|---|---|---|---|
| FasterRCNN-T [56] | High | 47.59 | 50.13 | 40.93 |
| TPIHOG [57] | High | – | – | 57.38 |
| SSD300 [14] | High | 69.81 | – | – |
| Guo et al. [22] | High | 46.30 | 53.37 | 31.63 |
| Guo et al. [22] | High | 42.65 | 49.59 | 26.70 |
| Kieu et al. [19] | High | 35.20 | 40.00 | 20.50 |
| Ours | Low | **34.23** | **39.47** | **19.68** |

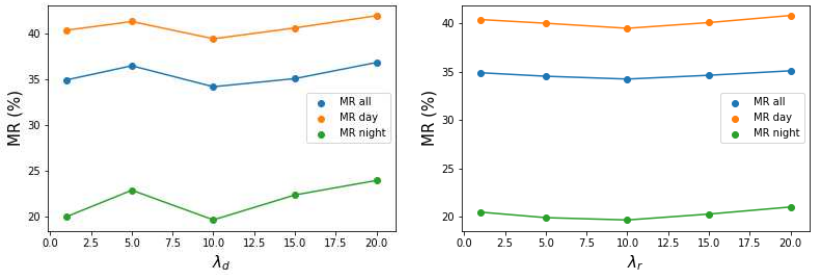


**Fig. 4** Hyperparameter analysis on the KAIST dataset. (a)$\lambda_d$ and (b)$\lambda_r$.

can extract complete content features; while for the up-sampled thermal image $x_u^T$, the feature disentanglement network can only get incomplete content features. But only the up-sampled $x_u^T$ is available at the inference stage. So we train YOLOv3 backbone such that it has the ability to extract the complete content features with the help of the visual image $x^V$. By requiring the content feature consistency between $x_u^T$ and $x^V$, the restoration loss function is defined as follows,

$$L_r = L_D(y_u^T, \hat{y}_u^T) + \lambda_I \cdot L_D(y^V, \hat{y}^T) + \lambda_I \cdot L_D(y_{ud}^V, \hat{y}_{ud}^V) + \lambda_r L_a(f_{x_u^T}^c, f_{x^V}^c), \quad (7)$$

where the loss function $L_a$ aims to restore content features by aligning two features $f_{x_u^T}^c$, $f_{x^V}^c$.

We set $\lambda_r$ as 10 here. The low-value $\lambda_r$ causes failure to restore content features; while the high-value $\lambda_r$ causes that YOLOv3 backbone returns only content features and ignores modal features.

# 4 Experiments

**Datasets.** Our experiments are conducted on the KAIST dataset and LLVIP dataset because these datasets ensure that thermal images and visible images are paired. The KAIST dataset consists of 59328 thermal-visible image pairs

**Table 2** Comparison with the YOLOv3 on the LLVIP dataset in terms of average precision (higher is better). ✓means that the corresponding data is used, and × means that the corresponding data is not used. The best results are highlighted in **bold**.

| | Training | | |
|---|---|---|---|
| Method | Visible | Thermal | AP |
| YOLOv3 | × | ✓ | 31.3 |
| YOLOv3 | ✓ | ✓ | 30.8 |
| Ours | ✓ | ✓ | **32.6** |

**Table 3** Ablation detection results showing the effect of the strategy. The best results are highlighted in **bold**.

| Disentanglement | Restoration | MR all | MR day | MR night |
|---|---|---|---|---|
| × | × | 38.01 | 43.36 | 23.92 |
| ✓ | × | 36.47 | 41.43 | 21.11 |
| ✓ | ✓ | **34.23** | **39.47** | **19.68** |

for training and 45156 for testing. As is common practice [56, 58–60], 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 [60] and testing annotations from [56]. The final dataset consists 7601 thermal-visible image pairs for training and 2252 for testing. The LLVIP dataset consists of 12025 thermal-visible image pairs for training and 3463 for testing. And the dataset is under low-light conditions [61].

**Evaluation Criteria.** As is common practice, for the KAIST dataset, the evaluation 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). For the LLVIP dataset, the evaluation is AP, which is the same as the evaluation of object detection.

**Implementation Details.** All of our networks are implemented in PyTorch on two NVIDIA GTX 1080Ti. The networks are pre-trained in the COCO dataset. We set high-resolution size as $416 \times 416$ and low-resolution size as $208 \times 208$. During training, we set aside 10% of the training images for validation. The batch size is 8. For initialization, we set the epochs as 5. The learning rate is 0.0001 and decays linearly to 94% for each epoch. For disentanglement, we set the epochs as 5. The learning rate is 0.0000001 and decays linearly to 94% for each epoch. For restoration, we set the epochs as 5. The learning rate is 0.0001 and decays linearly to 94% for each epoch.

**Table 4** Analysis of our method with the varying place locations of the feature disentanglement network on the KAIST dataset. The best results are highlighted in **bold**.

| Number | shape | MR all | MR day | MR night |
|---|---|---|---|---|
| 1 | $208 \times 208 \times 64$ | 38.94 | 43.77 | 27.93 |
| 2 | $104 \times 104 \times 128$ | 35.61 | 40.87 | 24.39 |
| 3 | $52 \times 52 \times 256$ | 37.11 | 41.28 | 25.47 |
| 4 | $26 \times 26 \times 512$ | 35.21 | 40.03 | 23.86 |
| 5 | $13 \times 13 \times 1024$ | **34.23** | **39.47** | **19.68** |

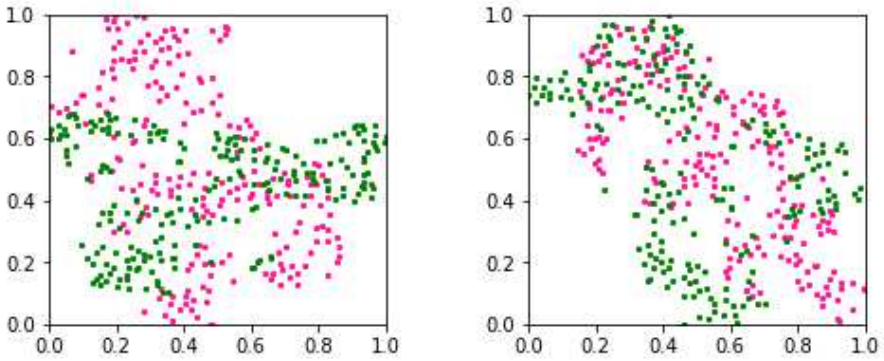


**Fig. 5** Visualization of the feature disentanglement network. The left and right figures represent the features before and after the feature disentanglement network respectively. Red and green points represent the thermal and visible features respectively.

## 4.1 Comparisons

**Experiments on KAIST dataset.** At present, there does not exist any experimental report on our problem setting. So we choose two categories of methods for comparison. The first category is single domain based method, which only uses thermal images for training and testing such as FasterRCNN-T [56], TPIHOG [57] and Guo et al. [22]. Another one is domain adaptation method, which transfers visible domain knowledge into thermal domain such as SSD300 [14] and Kieu et al. [19]. All these methods use high-resolution thermal images for training and testing. Table 1 shows the detection results on the KAIST dataset. Our method performs better by using low-resolution thermal images for testing than other methods using high-resolution thermal images for testing. Therefore, our method is able to achieve knowledge transfer at different resolutions of visible and thermal images.

**Experiments on LLVIP dataset.** We compare our DTR with the state-of-the-art method YOLOv3[53]. Table 2 shows the detection results on the LLVIP dataset. DTR boosts the AP by +1.3% (from 31.3% to 32.6%).

## 4.2 Further Analysis

**Ablation Study.** To explore the effectiveness of the strategy during adaptation, as shown in Table 3, we conduct an ablation study on the KAIST dataset. With the disentanglement strategy, DTR reduces the MR by -1.54% (from 38.01% to 36.47%), while for the restoration strategy, DTR reduces the MR by -2.24% (from 36.47% to 34.23%).

**Hyper-parameters Analysis.** We perform analysis on $\lambda_d$ and $\lambda_r$ on the KAIST dataset. (1) We modify $\lambda_d$ by changing $\{5, 7.5, 10, 12.5, 15\}$. As shown in Fig.4(a), when $\lambda_d = 10$, the performance on KAIST is the best. Our model can keep a relatively stable result in a wide range of $\lambda_d$. Besides, the low-value $\lambda_d$ means the feature disentanglement network cannot transform features into content features because of insufficient training for $L_a(\hat{f}_{x_u^T}, \hat{f}_{x_{ud}^V})$. The high-value $\lambda_d$ indicates that the loss function $L_{dis}$ only ensures the identical outputs of the feature disentanglement network. However, because of the low weight of loss for pedestrian detection, the output returned by the feature disentanglement network contains irrelevant features rather than only content features. (2) We modify $\lambda_r$ by changing $\{5, 7.5, 10, 12.5, 15\}$. As shown in Fig.4(b), when $\lambda_r = 10$, the performance on KAIST is the best. Our model can keep a relatively stable result in a wide range of $\lambda_r$. Besides, the low-value $\lambda_r$ means low-resolution thermal images cannot restore content features from high-resolution visible images because of insufficient training for $L_a(\hat{f}_{x_u^T}, \hat{f}_{x^V})$. The high-value $\lambda_r$ indicates that the loss function $L_r$ only ensures the identical outputs of the feature disentanglement network. However, because of the low weight of loss for pedestrian detection, the backbone only returns content features, which lack thermal modal features.

**Location of the feature disentanglement network.** Darknet of YOLOv3 contains five Residual Blocks. In our method, the feature disentanglement network is placed behind the last Residual Block. To demonstrate the effect of the location of the feature disentanglement network, we put the feature disentanglement network after each Residual Block separately. Since the features returned by different Residual blocks own different shapes, if the channel of the feature is $c$, the channel number will be $2c$ after the first ConvTranspose2d and $c$ after the last Conv2d respectively. The result is shown in Table 4. The performance is the best when the feature disentanglement network is placed behind the last Residual Block. This is because a deeper network can provide semantic features.

**Visualization of Features.** We randomly selected 10% of the sample visible-thermal image pairs in the KAIST dataset for the visualization of features. We use the backbone to get features of visible and thermal images. Then we use the feature disentanglement network to get content features. We use the TSNE algorithm [62] to reduce the dimensions of these features to two dimensions. As shown in Fig. 5, after the feature disentanglement network, the features of visible-thermal images are closer, which demonstrates that the feature disentanglement network successfully transforms features into content features.

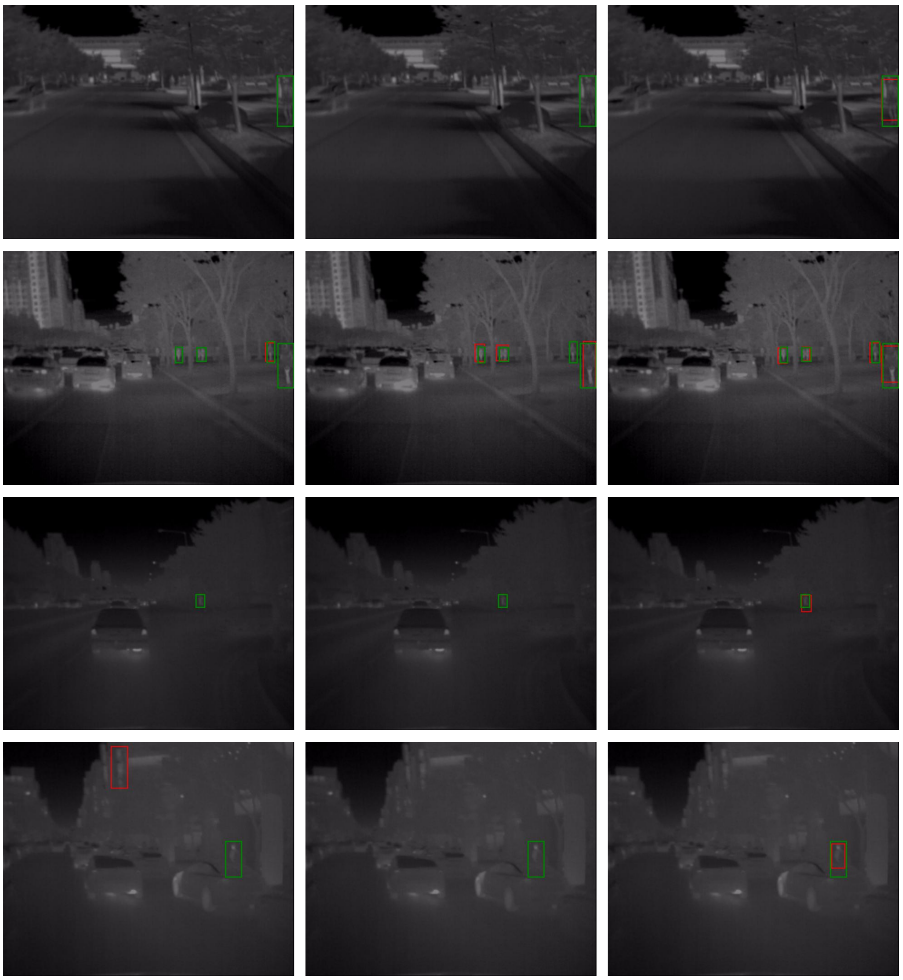

**Fig. 6** Pedestrian detection results on the KAIST dataset. The first two rows are daytime images and the last two are nighttime. The first column is the baseline method YOLOv3 trained by thermal images. The second column is the baseline method YoLOv3 trained by both thermal and visible images. The third column is our method. Green boxes are the ground truth and red boxes are the detection results. Best viewed in color.

## 4.3 Visualization comparison

In this section, we print the prediction results and annotations on some images and save them for visual comparison. We compare our method with YOLOv3. In Fig.6, we give some example detections from YOLOv3 and DTR on the KAIST. In Fig.7, example detections from YOLOv3 and DTR on the LLVIP are shown. As shown in Fig. 6 and Fig. 7, our method can reduce false positive predictions and improve true positive predictions. Through the visual analysis of these experiments, we can verify that our method can get better performance.

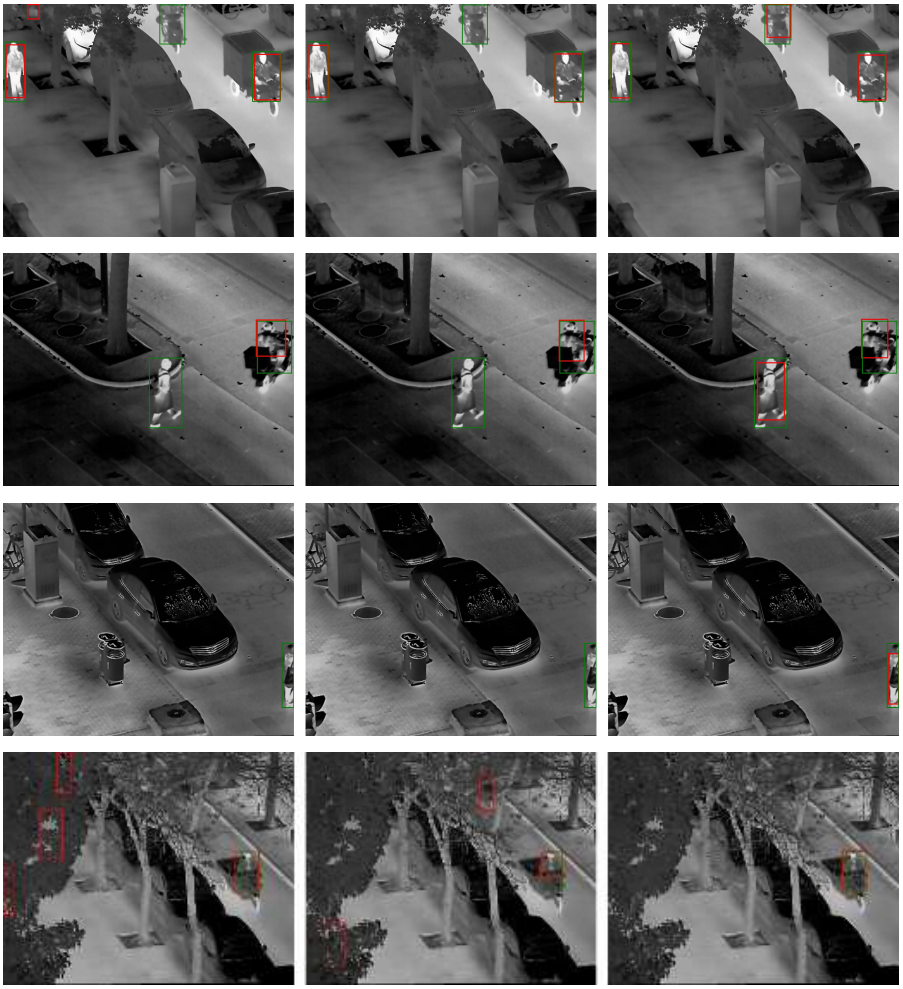

**Fig. 7** Pedestrian detection results on the LLVIP dataset. The setting is the same as Fig.6. Best viewed in color.

## 5 Conclusion

We proposed a novel strategy, dubbed as Disentanglement Then Restoration (DTR), which can solve the new problem that the resolution of thermal images is lower than the corresponding visible images. Based on this strategy, a feature disentanglement network is proposed. Compared with other disentanglement learning methods, our proposed feature disentanglement network can easily transform features into content features. The complete content features of thermal images can be restored by learning the changes of content features caused by different resolutions. Experiments confirm the effectiveness of our method. Moreover, not limited to YOLO, theoretically, the proposed method can also be applied to other detection frameworks, e.g. SSD or Faster RCNN.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62276048), Sichuan Science and Technology Program (2020YFG0476).

## 6 Conflict of Interest

There are no conflicts of interest.

## 7 Data Availability

The KAIST dataset analyzed during the current study is available at https://soonminhwang.github.io/rgbt-ped-detection/. The LLVIP dataset analyzed during the current study is available at https://bupt-ai-cz.github.io/LLVIP/.

## References

[1] Cao, J., Pang, Y., Xie, J., Khan, F.S., Shao, L.: From handcrafted to deep features for pedestrian detection: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 4913–4934 (2022)

[2] Tang, Y., Li, B., Liu, M., Chen, B., Wang, Y., Ouyang, W.: Autopedestrian: an automatic data augmentation and loss function search scheme for pedestrian detection. IEEE Transactions on Image Processing **30**, 8483–8496 (2021)

[3] Zhou, C., Wu, M., Lam, S.-K.: Enhanced multi-task learning architecture for detecting pedestrian at far distance. IEEE Transactions on Intelligent Transportation Systems **30**, 15588–15604 (2022)

[4] He, Y., Zhu, C., Yin, X.-C.: Occluded pedestrian detection via distribution-based mutual-supervised feature learning. IEEE Transactions on Intelligent Transportation Systems **23**, 10514–10529 (2021)

[5] Jiao, Y., Yao, H., Xu, C.: San: selective alignment network for cross-domain pedestrian detection. IEEE Transactions on Image Processing **30**, 2155–2167 (2021)

[6] Wu, J., Zhou, C., Yang, M., Zhang, Q., Li, Y., Yuan, J.: Temporal-context enhanced detection of heavily occluded pedestrians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13430–13439 (2020)

[7] Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M.: Fully convolutional region proposal networks for multispectral person detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 49–56 (2017)

[8] Chen, Z., Huang, X.: Pedestrian detection for autonomous vehicle using multi-spectral cameras. IEEE Transactions on Intelligent Vehicles **4**(2), 211–219 (2019)

[9] Kim, J.U., Park, S., Ro, Y.M.: Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1510–1523 (2022)

[10] Dasgupta, K., Das, A., Das, S., Bhattacharya, U., Yogamani, S.: Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. IEEE Transactions on Intelligent Transportation Systems **23**, 15940–15950 (2022)

[11] Zhang, L., Liu, Z., Zhang, S., Yang, X., Qiao, H., Huang, K., Hussain, A.: Cross-modality interactive attention network for multispectral pedestrian detection. Information Fusion **50**, 20–29 (2019)

[12] Li, C., Song, D., Tong, R., Tang, M.: Illumination-aware faster r-cnn for robust multispectral pedestrian detection. Pattern Recognition **85**, 161–171 (2019)

[13] Guan, D., Cao, Y., Yang, J., Cao, Y., Yang, M.Y.: Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion **50**, 148–157 (2019)

[14] Herrmann, C., Ruf, M., Beyerer, J.: Cnn-based thermal infrared person detection by domain adaptation. In: Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything, vol. 10643, p. 1064308 (2018). International Society for Optics and Photonics

[15] Ghose, D., Desai, S.M., Bhattacharya, S., Chakraborty, D., Fiterau, M., Rahman, T.: Pedestrian detection in thermal images using saliency maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)

[16] Xu, Z., Vong, C.-M., Wong, C.-C., Liu, Q.: Ground plane context aggregation network for day-and-night on vehicular pedestrian detection. IEEE Transactions on Intelligent Transportation Systems **22**(10), 6395–6406 (2020)

[17] Kim, J.U., Park, S., Ro, Y.M.: Robust small-scale pedestrian detection with cued recall via memory learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3050–3059 (2021)

[18] Kieu, M., Bagdanov, A.D., Bertini, M., Bimbo, A.d.: Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In: European Conference on Computer Vision, pp. 546–562 (2020). Springer

[19] Kieu, M., Bagdanov, A.D., Bertini, M., Bimbo, A.D.: Domain adaptation for privacy-preserving pedestrian detection in thermal imagery. In: International Conference on Image Analysis and Processing, pp. 203–213 (2019). Springer

[20] Kieu, M., Bagdanov, A.D., Bertini, M.: Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(1), 1–19 (2021)

[21] Kieu, M., Berlincioni, L., Galteri, L., Bertini, M., Bagdanov, A.D., Del Bimbo, A.: Robust pedestrian detection in thermal imagery using synthesized images. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 8804–8811 (2021). IEEE

[22] Guo, T., Huynh, C.P., Solh, M.: Domain-adaptive pedestrian detection in thermal images. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 1660–1664 (2019). IEEE

[23] Liu, D., Zhang, C., Song, Y., Huang, H., Wang, C., Barnett, M., Cai, W.: Decompose to adapt: Cross-domain object detection via feature disentanglement. IEEE Transactions on Multimedia (2022). https://doi.org/10.1109/TMM.2022.3141614

[24] Chen, Z., Yang, C., Li, Q., Zhao, F., Zha, Z.-J., Wu, F.: Disentangle your dense object detector. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4939–4948 (2021)

[25] Lin, C., Yuan, Z., Zhao, S., Sun, P., Wang, C., Cai, J.: Domain-invariant disentangled network for generalizable object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8771–8780 (2021)

[26] Wu, A., Han, Y., Zhu, L., Yang, Y.: Instance-invariant domain adaptive object detection via progressive disentanglement. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 4178–4193 (2022)

[27] Baek, J., Hyun, J., Kim, E.: A pedestrian detection system accelerated by kernelized proposals. IEEE Transactions on Intelligent Transportation Systems **21**(3), 1216–1228 (2019)

[28] Zhou, C., Wu, M., Lam, S.-K.: Group cost-sensitive boostlr with vector form decorrelated filters for pedestrian detection. IEEE Transactions on Intelligent Transportation Systems **21**(12), 5022–5035 (2019)

[29] Liu, X., Toh, K.-A., Allebach, J.P.: Pedestrian detection using pixel difference matrix projection. IEEE Transactions on Intelligent Transportation

Systems **21**(4), 1441–1454 (2019)

[30] Shen, J., Zuo, X., Zhu, L., Li, J., Yang, W., Ling, H.: Pedestrian proposal and refining based on the shared pixel differential feature. IEEE Transactions on Intelligent Transportation Systems **20**(6), 2085–2095 (2018)

[31] Yang, P., Zhang, G., Wang, L., Xu, L., Deng, Q., Yang, M.-H.: A part-aware multi-scale fully convolutional network for pedestrian detection. IEEE Transactions on Intelligent Transportation Systems **22**(2), 1125–1137 (2020)

[32] Cai, Z., Saberian, M., Vasconcelos, N.: Learning complexity-aware cascades for deep pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3361–3369 (2015)

[33] Cao, J., Pang, Y., Li, X.: Learning multilayer channel features for pedestrian detection. IEEE Transactions on Image Processing **26**(7), 3210–3220 (2017)

[34] Lin, C., Lu, J., Wang, G., Zhou, J.: Graininess-aware deep feature learning for pedestrian detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 732–747 (2018)

[35] Lin, C., Lu, J., Zhou, J.: Multi-grained deep feature learning for robust pedestrian detection. IEEE Transactions on Circuits and Systems for Video Technology **29**(12), 3608–3621 (2018)

[36] Pang, Y., Xie, J., Khan, M.H., Anwer, R.M., Khan, F.S., Shao, L.: Mask-guided attention network for occluded pedestrian detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4967–4975 (2019)

[37] Luo, Y., Zhang, C., Lin, W., Yang, X., Sun, J.: Sequential attention-based distinct part modeling for balanced pedestrian detection. IEEE Transactions on Intelligent Transportation Systems **23**, 15644–15654 (2022)

[38] Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6995–7003 (2018)

[39] Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5363–5371 (2017)

[40] Li, Q., Zhang, C., Hu, Q., Fu, H., Zhu, P.: Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. IEEE Transactions on Multimedia (2022). https://doi.org/10.1109/TMM.2022.3160589

[41] Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5127–5137 (2019)

[42] Zhou, K., Chen, L., Cao, X.: Improving multispectral pedestrian detection by addressing modality imbalance problems. In: European Conference on Computer Vision, pp. 787–803 (2020). Springer

[43] Guan, D., Luo, X., Cao, Y., Yang, J., Cao, Y., Vosselman, G., Ying Yang, M.: Unsupervised domain adaptation for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)

[44] Cao, Y., Guan, D., Huang, W., Yang, J., Cao, Y., Qiao, Y.: Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. Information Fusion **46**, 206–217 (2019)

[45] Kim, J.U., Park, S., Ro, Y.M.: Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory. In: 36th AAAI Conference on Artificial Intelligence (AAAI 22) (2022). Association for the Advancement of Artificial Intelligence

[46] VS, V., Poster, D., You, S., Hu, S., Patel, V.M.: Meta-uda: Unsupervised domain adaptive thermal object detection using meta-learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1412–1423 (2022)

[47] Jhoo, W.Y., Heo, J.-P.: Collaborative learning with disentangled features for zero-shot domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8896–8905 (2021)

[48] Lin, C.-C., Chu, H.-L., Wang, Y.-C.F., Lei, C.-L.: Joint feature disentanglement and hallucination for few-shot image classification. IEEE Transactions on Image Processing **30**, 9245–9258 (2021)

[49] Tang, L., Li, B., Zhong, Y., Ding, S., Song, M.: Disentangled high quality salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3580–3590 (2021)

[50] Wu, A., Liu, R., Han, Y., Zhu, L., Yang, Y.: Vector-decomposed disentanglement for domain-invariant object detection. In: Proceedings of the

IEEE/CVF International Conference on Computer Vision, pp. 9342–9351 (2021)

[51] Jia, M., Cheng, X., Lu, S., Zhang, J.: Learning disentangled representation implicitly via transformer for occluded person re-identification. IEEE Transactions on Multimedia (2022). https://doi.org/10.1109/TMM.2022.3141267

[52] Lee, Y., Yoo, H., Yu, J., Jeon, M.: Learning to see in the rain via disentangled representation. IEEE Robotics and Automation Letters (2021). https://doi.org/10.1109/LRA.2021.3117249

[53] Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)

[54] Peng, X., Huang, Z., Sun, X., Saenko, K.: Domain agnostic learning with disentangled representations. In: International Conference on Machine Learning, pp. 5102–5112 (2019). PMLR

[55] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[56] Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. arXiv preprint arXiv:1611.02644 (2016)

[57] Baek, J., Hong, S., Kim, J., Kim, E.: Efficient pedestrian detection at nighttime using a thermal camera. Sensors **17**(8), 1850 (2017)

[58] Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(4), 743–761 (2011)

[59] Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1037–1045 (2015)

[60] Li, C., Song, D., Tong, R., Tang, M.: Multispectral pedestrian detection via simultaneous detection and segmentation. arXiv preprint arXiv:1808.04818 (2018)

[61] Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3496–3504 (2021)

[62] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11), 2579–2605 (2008)