Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection

Lu Zhang[1,3], Xiangyu Zhu[2,3], Xiangyu Chen[5], Xu Yang[1,3], Zhen Lei[2,3], Zhiyong Liu[1,3,4*]
[1] SKL-MCCS, Institute of Automation, Chinese Academy of Sciences
[2] CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences [4] CEBSIT, Chinese Academy of Sciences
[5] Renmin University of China
{zhanglu2016,xu.yang,zhiyong.liu}@ia.ac.cn, {xiangyu.zhu,zlei}@nlpr.ia.ac.cn

Abstract

Multispectral pedestrian detection has shown great advantages under poor illumination conditions, since the thermal modality provides complementary information for the color image. However, real multispectral data suffers from the position shift problem, i.e. *the color-thermal image pairs are not strictly aligned, making one object has different positions in different modalities. In deep learning based methods, this problem makes it difficult to fuse the feature maps from both modalities and puzzles the CNN training. In this paper, we propose a novel Aligned Region CNN (AR-CNN) to handle the weakly aligned multispectral data in an end-to-end way. Firstly, we design a Region Feature Alignment (RFA) module to capture the position shift and adaptively align the region features of the two modalities. Secondly, we present a new multimodal fusion method, which performs feature re-weighting to select more reliable features and suppress the useless ones. Besides, we propose a novel RoI jitter strategy to improve the robustness to unexpected shift patterns of different devices and system settings. Finally, since our method depends on a new kind of labelling: bounding boxes that match each modality, we manually relabel the KAIST dataset by locating bounding boxes in both modalities and building their relationships, providing a new KAIST-Paired Annotation. Extensive experimental validations on existing datasets are performed, demonstrating the effectiveness and robustness of the proposed method. Code and data are available at* `https://github.com/luzhang16/AR-CNN`.

1. Introduction

Pedestrian detection is an important research topic in computer vision field with various applications, such as video surveillance, autonomous driving, and robotics. Although great progress has been made by the deep learning based methods (*e.g.* [42, 53, 54, 33]), detecting the pedestrian in adverse illumination conditions, occlusions and clutter background is still a challenging problem. Recently, many works in robot vision [4, 49], facial expression recognition [9], material classification [41], and object detection [45, 11, 20, 51] show that adopting a novel modality can improve the performance and offer competitive advantages over single sensor systems. Among the sensors, thermal camera is widely used in face recognition [3, 44, 27], human tracking [30, 46] and action recognition [59, 15] for its biometric robustness. Motivated by this, multispectral pedestrian detection [24, 52, 17, 39] has attracted massive attention and provides new opportunities for around-the-clock applications, mainly due to its superiority of complementary nature between color and thermal modalities.

Challenges A common assumption of multispectral pedestrian detection is that the color-thermal image pairs are geometrically aligned [24, 52, 34, 28, 32, 18, 55]. However, the modalities are just weakly aligned in practice, which means there is the position shift between modalities, making one object has different positions on different modalities, see Figure 1(a). This position shift problem can be caused by physical properties of different sensors (*e.g.* parallax, mismatched resolutions and field-of-views), imperfection of alignment algorithms, external disturbance, and hardware aging. Moreover, the calibration for color-thermal cameras are tortuous, generally require particular hardware as well as special heated calibration board [26, 23, 24, 47, 8].

The position shift problem degrades the pedestrian detector in two aspects. First, features from different modalities are mismatched in the corresponding positions, which puzzles the inference. Second, it is difficult to cover the objects in both modalities with a single bounding box. In existing datasets [24, 17], the bounding box is either labelled on single modality (color or thermal) or a big bounding box is labelled to cover the objects on both modalities. This la-

*Corresponding author

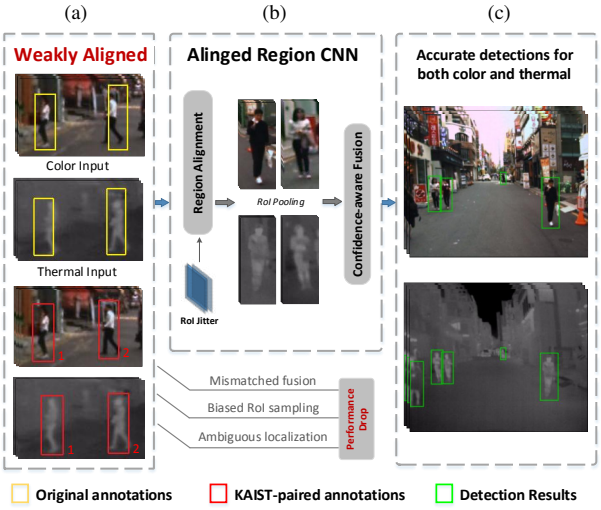


Figure 1. Overview of our framework. (a) The color-thermal pair and its annotations, the yellow boxes denote original KAIST annotations, which have position shift between two modalities; the red boxes are the proposed KAIST-Paired annotations, which have independent labelling to match each modality. (b) Illustration of the proposed AR-CNN model. (c) Detection results of both modalities under the position shift problem.

bel bias will give bad supervision singals and degrade the detector especially for CNN based methods [40, 35], where the intersection over union (IoU) overlap is used for foreground/background classification. Therefore, how to robustly localize each individual person on weakly aligned modalities remains to be a critical issue for multispectral pedestrian detectors.

Our Contributions (1). To the best of our knowledge, this is the first work that tackles the position shift problem in multispectral pedestrian detection. In this paper, we analyse the impacts of the position shift problem and propose a novel detection framework to merge the information from both modalities. Specifically, a novel RFA module is presented to shift and align the feature maps from two modalities. Meanwhile, the RoI jitter training strategy is adopted to randomly jitter the RoIs of the sensed modality, improving the robustness to the patterns of position shift. Furthermore, a new confidence-aware fusion method is presented to effectively merge the modality, which adaptively performs feature re-weighting to select more reliable features and depress the confusing ones. Figure 1 depicts an overview of the proposed approach.

(2). To realize our method, we manually relabel the KAIST dataset and provide a novel KAIST-Paired annotation. We first filter the image pairs with original annotations and obtain $20,025$ valid frames. Then $59,812$ pedestrians are carefully annotated by locating the bounding boxes in both modalities and building their relationships.

(3). The experimental results show that the proposed approach reduces the degradation from position shift problem and makes full use of both modalities, achieving the state-of-the-art performance on the challenging KAIST and CVC-14 dataset.

2. Related Work

Multispectral Pedestrian Detection As an essential step for various applications, pedestrian detection has attracted massive attention from the computer vision community. Over the years, extensive features and algorithms have been proposed, including both traditional detectors [14, 12, 38, 56] and the lately dominated CNN-based detectors [37, 22, 1, 50, 58]. Recently, multispectral data have shown great advantages, especially for the all-day vision [25, 7, 8]. Hence the release of large-scale multispectral pedestrian benchmarks [24, 17] is encouraging the research community to advance the state-of-the-art by efficiently exploiting multispectral input data. Hwang *et al*. [24] propose an extended ACF method, leveraging aligned color-thermal image pairs for around-the-clock pedestrian detection. With the recent development of deep learning, the CNN-based methods [52, 48, 6, 55, 19, 32] significantly improve the multispectral pedestrian detection performances. Liu *et al*. [34] adopt the Faster R-CNN architecture and analyze different fusion stages within the CNN. König *et al*. [28] adapt the Region Proposal Network (RPN) and Boosted Forest (BF) framework for multispectral input data. Xu *et al*. [52] design a cross-modal representation learning framework to overcome adverse illumination conditions.

However, most existing methods are employed under the full alignment assumption, hence directly fuse features of different modalities in the corresponding pixel position. This not only hinders the usage of the weakly aligned dataset (*e.g*. CVC-14 [17]), but also restricts the further development of multispectral pedestrian detection, which is worthy of attention but still exhibits a lack of study.

Weakly Aligned Image Pair Weakly aligned image pair is a common phenomenon in multispectral data, since images from different modalities are usually collected and processed independently. A common paradigm to address this problem is to conduct image registration (*i.e*. spatial alignment) [60, 2, 10, 36] as preprocessing. It geometrically aligns two images: the *reference* and *sensed* images, which can be considered as an image-level solution for the position shift problem. The standard registration includes four typical processes: feature detection, mapping function design, feature matching, and image transformation and resampling. Though well-established, the image registration mainly focuses on the low-level transformation of the whole image, which actually introduces time-consuming preprocessing and disenables the CNN-based detector to be trained in an end-to-end way.

(a) (b)

Figure 2. The visualization examples of ground truth annotations in the KAIST (boxes in yellow) and CVC-14 (boxes in orange) dataset. Image patches are cropped on the same position of color-thermal image pairs.

3. Motivation

To provide insights into the position shift problem in weakly aligned image pairs, we start with our analysis of the KAIST [24] and CVC-14 [17] multispectral pedestrian dataset. Then we experimentally study how the position shift problem impacts the detection performance.

3.1. Important Observations

From the multispectral image pairs and the corresponding annotations in the KAIST and CVC-14 dataset, several issues can be observed.

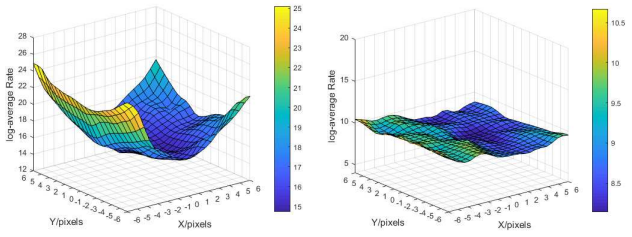
Weakly Aligned Features As illustrated in Figure 2(a), the weakly aligned color-thermal image pairs suffer from position shift problem, which makes it unreasonable to directly fuse the feature maps of different modalities.

Localization Bias Due to the position shift problem, the annotation must be refined to match both modalities, see Figure 2(a). One way is to adopt larger bounding boxes, encompassing pedestrians of both modalities, but generating too big bounding boxes for each modality. Another remedy is to only focus on one particular modality, while introducing bias for another modality.

Unpaired Objects Since the image pair of two modalities may have different field-of-views due to bad camera synchronization and calibration, some pedestrians exist in one modality but are truncated/lost in another, see Figure 2(b). Specifically, 12.5% $(2,245$ of $18,017)$ of bounding boxes are unpaired in the CVC-14 [17] dataset.

3.2. How the Position Shift Impacts?

To quantitatively evaluate how the position shift problem influences the detection performance, we conduct experiments on the relatively well-aligned KAIST dataset by manually simulating the position shift.



(a) (b)

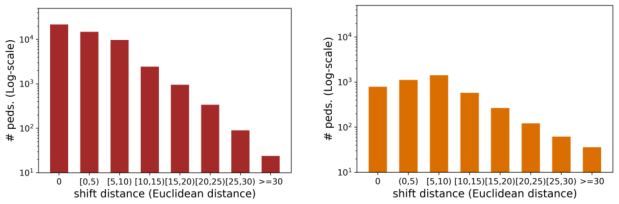
Figure 3. Surface plot of the detection performances within the position shift experiments. (a) Baseline detector. (b) The proposed approach. Horizontal coordinates indicate different step sizes by which sensed images are shifted along the x-axis and y-axis. Vertical coordinates denote the log-average miss rates (MR) measured on the reasonable test set of KAIST dataset, lower is better.

Baseline Detector We build our baseline detector based on the adapted Faster R-CNN [57] framework and adopt the halfway fusion settings [34] for multispectral pedestrian detection. To mitigate the negative effect of harsh quantization on localization, we use RoIAlign [21] instead of the standard RoIPool [40] for the region feature pooling process. Our baseline detector is solid: it has 15.18 MR on the KAIST reasonable test set, 10.97 better than the 26.15 reported in [34, 32].

Robustness to Position Shift In the testing phase, we fix the thermal image but spatially shift the color image along x-axis and y-axis. The shift pixel is selected in $\{(\Delta x, \Delta y) \mid \Delta x, \Delta y \in [-6, 6]; \Delta x, \Delta y \in Z\}$, which contains a total of 169 shift modes. As shown in Figure 3(a) and Table 1, the performance dramatically drops as the absolute shift values increase. Especially, the worst case $(\Delta x, \Delta y) = (-6, 6)$ suffers $\sim$ **65.3%** relative performance decrement, *i.e.* from 15.18 MR to 25.10 MR. Interestingly, when the image is shifted to a specific direction ($\Delta x = 1$, $\Delta y = -1$), a better result is achieved (15.18 MR to 14.68 MR), which indicates that we can improve the performance by appropriately handling the position shift problem.

ΔMR (%)		Δx				
		-6	-4	-1	0	1
Δy	1	$\downarrow$ 6.55	$\downarrow$ 2.62	$\downarrow$ 0.31	$\downarrow$ 0.14	$\downarrow$ 0.49
	0	$\downarrow$ 6.32	$\downarrow$ 2.41	$\downarrow$ 0.21	$0_{(15.18)}$	$\downarrow$ 0.14
	-1	$\downarrow$ 7.19	$\downarrow$ 2.58	$\downarrow$ 0.19	$\uparrow$ 0.25	$\uparrow$ 0.50
	-4	$\downarrow$ 8.27	$\downarrow$ 4.01	$\downarrow$ 0.84	$\uparrow$ 0.18	$\downarrow$ 0.22
	-6	$\downarrow$ 9.92	$\downarrow$ 5.27	$\downarrow$ 1.79	$\downarrow$ 1.37	$\downarrow$ 1.21

Table 1. Numerical results of the position shift experiments. The scores are corresponding with the results in Figure 3(a). Result in the origin is highlighted in blue, $\downarrow$ refers to performance drop and $\uparrow$ on the contrary.



(a) KAIST (b) CVC-14

Figure 4. The statistics of bounding box shift within color-thermal image pairs in KAIST and CVC-14 dataset.

4. The Proposed Approach

This section introduces the proposed KAIST-Paired annotation (Section 4.1) and Aligned Region CNN. The architecture of AR-CNN is shown in Figure 5, which consists of the region feature alignment module (Section 4.2), the RoI jitter training strategy (Section 4.3) and the confidence-aware fusion step (Section 4.4).

4.1. KAIST-Paired Annotation

In order to address the position shift problem, we first manually annotate the color-thermal bounding boxes pairs on each modality by the following principles:

- Localizing both modalities. The pedestrians are localized in both color and thermal images, aiming to clearly indicate the object locations on each modality.
- Adding relationships. A unique index is assigned to each pedestrian, indicating the pairing information between modalities.
- Labelling the unpaired objects. The pedestrians that only appear in one modality are labelled as "unpaired" to identify such situation.
- Extreme case. If the image quality of one modality is beyond human vision, *e.g.* color image under extremely bad illumination, we make the bounding box of color modality consistent with that in the thermal modality.

Statistics of KAIST-Paired From the new KAIST-Paired annotation, we can get the statistics information of shift distance in the original KAIST dataset. As illustrated in Figure 4(a), more than half of the bounding boxes have the position shift problem, and the shift distance mostly ranges from 0 to 10 pixels.

4.2. Region Feature Alignment

In this subsection, we propose the Region Feature Alignment (RFA) module to predict the shift between two modalities. Note that the position shift is not simply affine transformation and depends on the cameras. Furthermore, the shift distance varies from pixels to pixels, always small in the center and large in the edge. As a result, the shift prediction and alignment process is performed in a region-wise way.

Reference and Sensed Modality We introduce the concept of the *reference* and *sensed* [60, 2] image into the multispectral setting. In our implementation, we select the thermal image as the reference modality and the color image as the sensed modality. During training, we fix the reference modality and perform the learnable feature-level alignment and RoI jitter process on the sensed one.

Proposals Generation As illustrated in Figure 5, we utilize the Region Proposal Network (RPN) to generate numerous proposals. We aggregate the proposals from both reference feature map (Conv4_r) and sensed feature map (Conv4_s) to keep a high recall rate.

Alignment Process The concrete connection scheme of the RFA module is shown in Figure 6. Firstly, given several proposals, this module enlarges contextual RoIs to encompass sufficient information of regions. For each modality, the contextual region features are pooled into a small feature map with a fixed spatial extent of $H \times W$ (*e.g.* 7×7). Secondly, the feature map from each modality is then concatenated to get the multimodal representation. From this representation, two consecutive fully connected layers are used to predict the shift targets (*i.e.* t_x and t_y) of this region, so that the new coordinates of the sensed region is predicted. Finally, we re-pool the sensed feature map on the new region to get aligned feature representation with the reference modality. Since we have access to the annotated bounding boxes pairs on both modalities, the ground truth shift targets of the two region features can be calculated as follow:

$$t_x^* = (x_s - x_r)/w_r \quad t_y^* = (y_s - y_r)/h_r \tag{1}$$

In Equation 1, x, y denote the center coordinates of the box, w and h indicate the width and height of the box . Variables x_s, x_r are for the sensed and reference ground truth box respectively, t_x^* is the shift target for x coordinate, and likewise for y.

Multi-task Loss Similar to Fast R-CNN [16], we use the smooth L1 loss as the regression loss to measure the accuracy of predicted shift targets, *i.e.*,

$$L_{shift}(\{p_i^*\}, \{t_i\}, \{t_i^*\}) = \frac{1}{N_{shift}} \sum_{i=1}^{n} p_i^* \text{smoothL}_1(t_i - t_i^*) \tag{2}$$

where i is the index of RoI in a mini-batch, t_i is the predicted shift target, p_i^* and t_i^* are the associated ground truth class label (pedestrian $p_i^* = 1$ *vs.* background $p_i^* = 0$) and shift target of the i-th sensed RoI. N_{shift} is the total number of ground truth objects to be aligned (*i.e.* $N_{shift} = n$). For each training example, we minimize an objective function

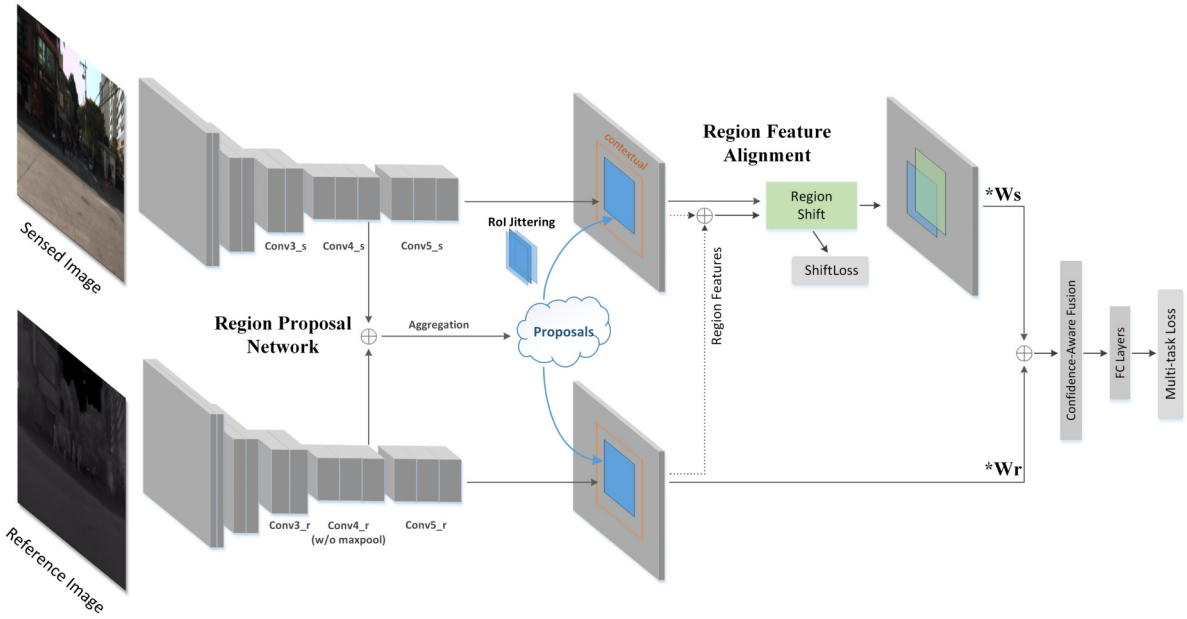


Figure 5. The network structure of Aligned Region CNN (AR-CNN). We adopt the two-stream framework to deal with color-thermal inputs. Given a pair of images, numerous proposals are generated and aggregated by the Region Proposal Network, then the Region Feature Alignment module is introduced to align the region features. After alignment, the region features of color and thermal feature maps are pooled respectively, then the confidence-aware fusion method is performed.

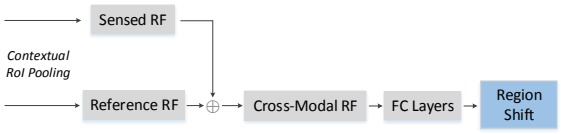


Figure 6. Connection scheme of the RFA module. RF denotes region feature and $\oplus$ refers to channel concatenation. The cross-modal region feature is fed into two fully-connected layers to predict this region's shift between two modalities.

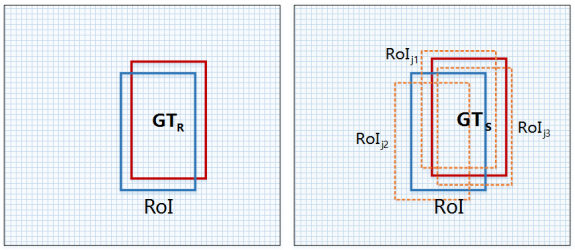
of Fast R-CNN which is defined as follow:

$$L(\{p_i\},\{t_i\},\{g_i\},\{p_i^*\},\{t_i^*\},\{g_i^*\}) = L_{cls}(\{p_i\},\{p_i^*\}) + \lambda L_{shift}(\{p_i^*\},\{t_i\},\{t_i^*\}) + L_{reg}(\{p_i^*\},\{g_i\},\{g_i^*\}) \quad (3)$$

where p_i and g_i are the predicted confidence and coordinates of the pedestrian, p_i^* and g_i^* are the associated ground truth label and the reference ground truth coordinates. Here the two terms L_{shift} and L_{reg} are weighted by a balancing parameter λ. In our current implementation, we set $\lambda = 1$, and thus the two terms are roughly equally weighted. For the RPN module, the loss function is defined as in the literature [40].

4.3. RoI Jitter Strategy

In reality, the shift patterns are unexpected due to the changes of devices and system settings. To improve the robustness to shift patterns, we propose a novel RoI jitter strategy to augment the shift modes. Specifically, the RoI jitter



(a) Reference image (b) Sensed image

Figure 7. Illustration of the RoI jitter strategy. Red boxes denote the ground truths, GT_R and GT_S stand for the reference and sensed modality respectively. Blue boxes represents the RoIs, *i.e.* the proposals, which are shared by both modalities. RoI_{j1}, RoI_{j2}, and RoI_{j3} are three feasible proposal instances after jitter.

introduces stochastic disturbances to the sensed RoIs and shifts the targets of RFA accordingly, which enriches the patterns of position shift in the training process, as shown in Figure 7.

The jitter targets are randomly generated from a normal distribution,

$$t_x^j,\ t_y^j \sim N(0,\sigma_0^2;0,\sigma_1^2;0) \quad (4)$$

where t^j denotes jitter targets of x-axis and y-axis, and σ is the hyperparameter of the radiation extent of jitter. After, the RoI jitters to the RoI_j by using the inverse process of bounding box transformation of Equation 1.

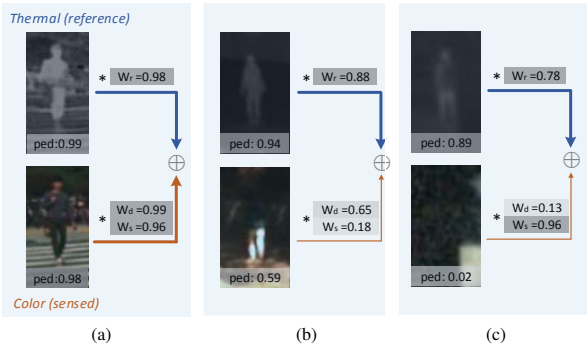


Figure 8. Illustration of the confidence-aware fusion method. There are three typical situations: (a) at day time, the color and thermal features are consistent and complementary. (b) under poor illumination, it is difficult to distinguish the pedestrian in color modality, hence we pay more weight on the thermal modality. (c) the pedestrian only exists in the thermal modality due to the position shift, so we depress the color feature.

Mini-batch Sampling While training the CNN-based detector, a small set of samples is randomly selected. We consistently define the positive and negative examples with respect to the reference modality, since the RoI jitter process is only performed on the sensed modality. Specifically, the RoI pair is treated as positive if the reference RoI has the IoU overlap with reference ground truth box greater than 0.5, and negative if the IoU is between 0.1 and 0.5.

4.4. Confidence-Aware Fusion

In around-the-clock operation, modalities provide variational qualities of information: the color data is discriminable at day time but fades at night; the thermal data presents clear human shape throughout the day and night while loses fine visual details (*e.g.* clothing). The naive fusion of features from different modalities is not appropriate since we want the detector to pay more attention to reliable modality. To this end, we propose a confidence-aware fusion method to make full use of the characteristics between two different modalitites via re-weighting their features, and select the more informative features while suppressing less useful ones.

As shown in Figure 8, the confidence-aware module has a two-stream architecture and fuses the feature maps from different modalities. This module adds a branch for each modality, which is composed of two fully connected layers for the confidence prediction. We calculate two confidence weights: $\mathrm{W}_r = |p_r^1 - p_r^0|$, $\mathrm{W}_s = |p_s^1 - p_s^0|$, in which p^1 and p^0 denote the probability of pedestrian and background, r and f refer to the reference and sensed modality, respectively. Then, we use multiplication to perform feature re-weighting (see Figure 8) on the input feature maps to select more reliable features for estimation.

Unpaired Objects During training, since the unpaired objects only exist in one modality, treating them as either background or foreground will lead to ambiguous classification. To mitigate this feature ambiguity, we calculate a disagreement weight, $\mathrm{W}_d = 1 - |p_r^1 - p_s^1| = 1 - |p_r^0 - p_s^0|$, and perform re-weighting on the sensed features, *i.e.* the sensed feature will be depressed if it provides a contradictory prediction with the reference modality.

5. Experiments

In this section, we conduct several experiments on the KAIST [24] and CVC-14 [17] dataset. We set the more reliable thermal input as the reference image and color input as the sensed one, the opposite configuration is discussed in the supplementary material. All methods are evaluated on the "reasonable" setup [13].

5.1. Dataset

KAIST The popular KAIST dataset [24] contains $95,328$ color and thermal image pairs with $103,128$ dense annotations and $1,182$ unique pedestrians. It is recorded in various scenes at day and night to cover the changes in light conditions. Detection performance is evaluated on the test set, which consists of $2,252$ frames sampled every 20th frame from videos.

CVC-14 The CVC-14 dataset [17] contains visible (grayscale) plus thermal video sequences, captured by a car traversing the streets at 10 FPS during day and night time. The training and testing set contains $7,085$ and $1,433$ frames, respectively. Note that even with post-processing, the cameras are still not well calibrated. As a result, annotations are individually provided in each modality. It is worth noting that the CVC-14 dataset has a more serious position shift problem, see Figure 4(b), which makes it difficult for state-of-the-art methods to use the dataset [52, 32, 18, 55, 5].

5.2. Implementation Details

Our AR-CNN detector uses VGG-16 [43] as the backbone network, which is pre-trained on the ILSVRC CLS-LOC dataset [29]. We set the σ_0 and σ_1 of RoI jitter to 0.05 by default, which can be adjusted to handle wider or narrower misalignment. All the images are horizontally flipped for data augmentation. We train the detector for 2 epochs with the learning rate of 0.005 and decay it by 0.1 for another 1 epoch. The network is optimized using the Stochastic Gradient Descent (SGD) algorithm with 0.9 momentum and 0.0005 weight decay. Multi-scale training and testing are not applied to ensure fair comparisons with other methods.

As for evaluation, the log miss rate averaged over the false positives per image (FPPI) range of $[10^{-2}, 10^{0}]$ (MR) is calculated to measure the detection performance, the lower score indicates better performance. Since there are

Method	MR			MR^C			MR^T		
	Day	Night	All	Day	Night	All	Day	Night	All
ACF+T+THOG (optimized) [24]	29.59	34.98	31.34	29.85	36.77	32.01	30.40	34.81	31.90
Halfway Fusion [34]	24.88	26.59	25.75	24.29	26.12	25.10	25.20	24.90	25.51
Fusion RPN [28]	19.55	22.12	20.67	19.69	21.83	20.52	21.08	20.88	21.43
Fusion RPN+BF [28]	16.49	15.15	15.91	16.60	15.28	15.98	17.56	14.48	16.52
Adapted Halfway Fusion	15.36	14.99	15.18	14.56	15.72	15.06	15.48	14.84	15.59
IAF-RCNN [32]	14.55	18.26	15.73	14.95	18.11	15.65	15.22	17.56	16.00
IATDNN+IAMSS [18]	14.67	15.72	14.95	14.82	15.87	15.14	15.02	15.20	15.08
CIAN [55]	14.77	11.13	14.12	15.13	12.43	14.64	16.21	9.88	14.68
MSDS-RCNN [31]	10.60	13.73	11.63	9.91	14.21	11.28	12.02	13.01	12.51
AR-CNN (Ours)	**9.94**	**8.38**	**9.34**	**8.45**	**9.16**	**8.86**	**9.08**	**7.04**	**8.26**

Table 2. Comparisons with the state-of-the-art methods on the KAIST dataset. Besides the MR protocol, we also evaluate the detectors on MR^C and MR^T in the KAIST-Paired annotation.

	Method	MR		
		Day	Night	All
Visible	SVM [17]	37.6	76.9	-
	DPM [17]	25.2	76.4	-
	Random Forest [17]	26.6	81.2	-
	ACF [39]	65.0	83.2	71.3
	Faster R-CNN [39]	43.2	71.4	51.9
Visible+Thermal	MACF [39]	61.3	48.2	60.1
	Choi *et al*. [6]	49.3	43.8	47.3
	Halfway Fusion [39]	38.1	34.4	37.0
	Park *et al*. [39]	31.8	30.8	31.4
	AR-CNN (Ours)	**24.7**	**18.1**	**22.1**

Table 3. Pedestrian detection results on the CVC-14 dataset. MR is used to compare the performance of detectors. The first column refers to input modalities of the approach. We use the reimplementation of ACF, Faster R-CNN, MACF, and Halfway Fusion in literature [39].

some problematic annotations in the original test set of the KAIST benchmark, we use the widely adopted improved test set annotations provided by Liu *et al*. [34]. Besides, based on our annotated KAIST-Paired, we propose the MR^C and MR^T which indicate the log-average miss rate on color and thermal modality.

5.3. Comparison Experiments

KAIST. We evaluate our approach and conduct comparisons with other published methods, as illustrated in Table 2. Our proposed approach achieves 9.94 MR, 8.38 MR, and 9.34 MR on the reasonable day, night, and all-day subset respectively, better than other available competitors (*i.e*. [24, 34, 28, 32, 18, 55]). Besides, in consideration of the position shift problem, we also evaluate the state-of-the-art methods with the KAIST-Paired annotation, *i.e*. log-average miss rate associated with color modality (MR^C) and thermal modality (MR^T). From Table 2 we can see that our AR-CNN detector has greater advantages, *i.e*. 8.86 *vs*. 11.28 MR^C and 8.26 *vs*. 12.51 MR^T, demonstrating the superiority of the proposed approach.

CVC-14. We follow the protocol in [39] to conduct the experiments. Table 3 shows that the AR-CNN outperforms all state-of-the-art methods, especially on the night subset (18.1 *vs*. 30.8 MR). This validates the contribution of thermal modality, and demonstrates the performance can be significantly boosted by correctly utilizing the weakly aligned modalities.

5.4. Robustness to Position Shift

Following the settings in Section 3.2, we test the robustness of AR-CNN to position shift by evaluating the MR^T on KAIST dataset. Figure 3(b) depicts the visual results by a surface plot. Compared to the baseline results in Figure 3(a), it can be observed that the robustness to position shift is significantly enhanced, with the overall performance improved. To further evaluate the robustness, we design four metrics[1]: S^{0°, S^{45°, S^{90°, S^{135°, where the 0°-135° indicate the shift directions. For each shift direction, we have 21 shift modes, which range from -10 to 10 pixels. The mean and standard deviation of those 21 results are calculated and

[1] Δx and Δy denote the shift pixels, which are selected in the following sets:

$S^{0^\circ} : \{(\Delta x, \Delta y) \mid \Delta y = 0, \Delta x \in [-10, 10]; \Delta x \in Z\}$

$S^{45^\circ} : \{(\Delta x, \Delta y) \mid \Delta x = \Delta y, \Delta x \in [-10, 10]; \Delta x \in Z\}$

$S^{90^\circ} : \{(\Delta x, \Delta y) \mid \Delta x = 0, \Delta y \in [-10, 10]; \Delta y \in Z\}$

$S^{135^\circ} : \{(\Delta x, \Delta y) \mid \Delta x = -\Delta y, \Delta y \in [-10, 10]; \Delta y \in Z\}$

Method				S^{0°			S^{45°		S^{90°		S^{135°	
				O	μ	σ	μ	σ	μ	σ	μ	σ
Halfway Fusion [34]				25.51	33.73	7.17	36.25	9.66	28.30	2.26	36.71	9.87
Fusion RPN [28]				21.43	30.12	7.13	31.69	10.60	24.48	1.97	34.02	10.64
Adapted Halfway Fusion				15.59	23.44	7.49	26.91	11.55	17.95	2.03	27.26	11.18
CIAN [55]				14.68	23.64	7.69	24.07	11.50	15.07	1.35	23.98	11.57
MSDS-RCNN [31]				12.51	20.96	7.87	24.43	11.74	14.42	1.34	24.23	10.99
	RFA	RoIJ	CAF									
AR-CNN				12.94	21.05	7.10	15.80	9.77	13.46	1.04	16.18	6.91
	✓			10.90	11.91	2.81	12.38	2.59	11.00	0.21	12.34	2.27
	✓	✓		9.87	11.17	1.20	11.84	1.71	10.27	0.17	11.50	1.34
	✓	✓	✓	**8.26**	**9.34**	**0.95**	**9.73**	**1.24**	**8.91**	**0.43**	**9.79**	**1.04**

Table 4. Quantitative results of the robustness of detectors to position shift on the KAIST dataset. O denotes the MR^T score at the origin, μ, σ represents the mean and standard deviation of MR^T scores respectively. In this testing, we reimplement the ACF+T+THOG, Halfway Fusion and Fusion RPN, and use the model provided in [55] and [31] for CIAN and MSDS-RCNN.

shown in Table 4. It can be observed that our AR-CNN detector achieves the best mean performance and the smallest standard deviation on all metrics, demonstrating the robustness of the proposed approach under diverse position shift conditions.

5.5. Ablation Study

In this section, we perform ablation experiments on the KAIST dataset for a detailed analysis of our AR-CNN detector. All the ablated detectors are trained using the same setting of parameters.

Region Feature Alignment Module. To demonstrate the contribution of the RFA module, we evaluate the performance with and without RFA in Table 4. We find the RFA remarkably reduces the MR^T and the standard deviation under diverse position shift conditions. Specifically, for S^{45°, the standard deviation is reduced by a significant 8.53 (from 9.77 to 1.24), and consistent reduction is also observed on the other three metrics.

RoI Jitter Strategy. Based on the RFA module, we further add the RoI jitter strategy and evaluate its contribution. As shown in Table 4, the RoI jitter further reduces the mean and standard deviation of results and achieves 9.87 MR^T at the origin. Besides, we can see that RoI jitter works more on standard deviation than the performance, which demonstrates that it improves the robustness to shift patterns.

Confidence-Aware Fusion Method. To validate the effectiveness of the confidence-aware fusion method, we compared performance with and without it. As shown in Table 4, the newly added confidence-aware fusion method slightly depresses the standard deviation, and further reduces MR^T at the origin by 1.61. This demonstrates that the detection performance can be further improved by confidence-aware fusion, since it helps the network to select more reliable features for adaptive fusion.

6. Conclusion

In this paper, a novel Aligned Region CNN method is proposed to alleviate the negative effects of position shift problem in weakly aligned image pairs. Specifically, we design a new region feature alignment module, which predicts the position shift and aligns the region features between modalities. Besides, an RoI jitter training strategy is adopted to further improve the robustness to random shift patterns. Meanwhile, we present a novel confidence-aware fusion method to enhance the representation ability of fused feature via adaptively re-weighting the features. To realize our method, we relabel the large-scale KAIST dataset by locating the bounding boxes in both modalities and building their relationships. Our model is trained in an end-to-end fashion and achieves state-of-the-art accuracy on the challenging KAIST and CVC-14 dataset. Furthermore, the detector robustness to position shift is improved with a large margin. It is also worth noting that our method is a generic solution for multispectral object detection rather than only the pedestrian problem. In the future, we plan to explore the generalization of the AR-CNN detector and extend it to other tasks, considering this weakly aligned characteristic is widespread and hard to completely avoid when multimodal inputs are required.

Acknowledgments. This work was supported by the National Key Research and Development Plan of China under Grant 2017YFB1300202, the NSFC under Grants U1613213, 61627808, 61876178, and 61806196, the Strategic Priority Research Program of Chinese Academy of Science under Grant XDB32050100.

References

[1] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection & segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4950–4959, 2017.

[2] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4):325–376, 1992.

[3] Pradeep Buddharaju, Ioannis T Pavlidis, Panagiotis Tsiamyrtzis, and Mike Bazakos. Physiology-based face recognition in the thermal infrared spectrum. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(4):613–626, 2007.

[4] Frantisek Burian, Petra Kocmanova, and Ludek Zalud. Robot mapping with range camera, ccd cameras and thermal imagers. In *Methods and Models in Automation and Robotics (MMAR), 2014 19th International Conference On*, pages 200–205, 2014.

[5] Yanpeng Cao, Dayan Guan, Yulun Wu, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150:70–79, 2019.

[6] Hangil Choi, Seungryong Kim, Kihong Park, and Kwanghoon Sohn. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In *2016 23rd IEEE International Conference on Pattern Recognition (ICPR)*, pages 621–626, 2016.

[7] Yukyung Choi, Namil Kim, Soonmin Hwang, and In So Kweon. Thermal image enhancement using convolutional neural network. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 223–230, 2016.

[8] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[9] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38(8):1548–1568, 2016.

[10] Suma Dawn, Vikas Saxena, and Bhudev Sharma. Remote sensing image registration techniques: A survey. In *International Conference on Image and Signal Processing*, pages 103–112, 2010.

[11] Zhuo Deng and Longin Jan Latecki. Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in rgb-depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5762–5770, 2017.

[12] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(8):1532–1545, 2014.

[13] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(4):743–761, 2012.

[14] Piotr Dollr, Zhuowen Tu, Pietro Perona, and Serge Belongie. Integral channel features. In *British Machine Vision Conference (BMVC)*, 2009.

[15] Chenqiang Gao, Yinhe Du, Jiang Liu, Jing Lv, Luyu Yang, Deyu Meng, and Alexander G Hauptmann. Infar dataset: Infrared action recognition at different times. *Neurocomputing*, 212:36–47, 2016.

[16] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[17] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M López. Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, 16(6):820, 2016.

[18] Dayan Guan, Yanpeng Cao, Jun Liang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019.

[19] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Christel-Loic Tisse. Exploiting fusion architectures for multispectral pedestrian detection and segmentation. *Applied Optics*, 57(18):D108–D116, 2018.

[20] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Proceedings of the European conference on Computer Vision (ECCV)*, pages 345–360, 2014.

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[22] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4073–4082, 2015.

[23] Soonmin Hwang, Yukyung Choi, Namil Kim, Kibaek Park, Jae Shin Yoon, and In So Kweon. Low-cost synchronization for multispectral cameras. In *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 435–436. IEEE, 2015.

[24] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015.

[25] Namil Kim, Yukyung Choi, Soonmin Hwang, and In So Kweon. Multispectral transfer network: Unsupervised depth estimation for all-day vision. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[26] Namil Kim, Yukyung Choi, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, and In So Kweon. Geometrical calibration of multispectral calibration. In *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 384–385. IEEE, 2015.

[27] Seong G Kong, Jingu Heo, Faysal Boughorbel, Yue Zheng, Besma R Abidi, Andreas Koschan, Mingzhong Yi, and Mongi A Abidi. Multiscale fusion of visible and thermal ir images for illumination-invariant face recognition. *International Journal of Computer Vision*, 71(2):215–233, 2007.

[28] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 243–250, 2017.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[30] Alex Leykin, Yang Ran, and Riad Hammoud. Thermal-visible video fusion for moving target tracking and pedestrian classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[31] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference (BMVC)*, September 2018.

[32] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.

[33] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 20(4):985–996, 2018.

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

[35] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD:single shot multibox detector. In *Proceedings of the European conference on Computer Vision (ECCV)*, pages 21–37, 2016.

[36] JB Antoine Maintz and Max A Viergever. A survey of medical image registration. *Medical image analysis*, 2(1):1–36, 1998.

[37] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3127–3136, 2017.

[38] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 424–432, 2014.

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

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[41] Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, and Chandra Kambhamettu. Material classification with thermal imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4649–4656, 2015.

[42] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3626–3633, 2013.

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

[44] Diego A Socolinsky, Andrea Selinger, and Joshua D Neuheisel. Face recognition with visible and thermal infrared imagery. *Computer Vision and Image Understanding*, 91(1-2):72–114, 2003.

[45] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 808–816, 2016.

[46] Atousa Torabi, Guillaume Massé, and Guillaume-Alexandre Bilodeau. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding*, 116(2):210–221, 2012.

[47] Wayne Treible, Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Michael O'Neal, Brian Phelan, Kelly Sherbondy, and Chandra Kambhamettu. Cats: A color and thermal stereo benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2961–2969, 2017.

[48] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 509–514, 2016.

[49] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3343–3352, 2019.

[50] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7774–7783, 2018.

[51] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 244–253, 2018.

[52] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5363–5371, 2017.

[53] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li. Convolutional channel features. In *Proceedings of the IEEE Interna-*

tional Conference on Computer Vision (ICCV), pages 82–90, 2015.

[54] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *Proceedings of the European conference on Computer Vision (ECCV)*, pages 443–457, 2016.

[55] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019.

[56] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Filtered channel features for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1751–1760, 2015.

[57] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3221, 2017.

[58] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018.

[59] Yu Zhu and Guodong Guo. A study on visible to infrared action recognition. *IEEE Signal Processing Letters*, 20(9):897–900, 2013.

[60] Barbara Zitova and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.