Weakly Aligned Feature Fusion for Multimodal Object Detection

Lu Zhang*, Zhiyong Liu*, Senior Member, IEEE, Xiangyu Zhu*, Member, IEEE, Zhan Song, Xu Yang*, Zhen Lei*, Senior Member, IEEE, and Hong Qiao*, Fellow, IEEE

Abstract—To achieve accurate and robust object detection in the real-world scenario, various forms of images are incorporated, such as color, thermal, and depth. However, multimodal data often suffer from the position shift problem, i.e., the image pair is not strictly aligned, making one object has different positions in different modalities. For the deep learning method, this problem makes it difficult to fuse multimodal features and puzzles the convolutional neural network (CNN) training. In this article, we propose a general multimodal detector named aligned region CNN (AR-CNN) to tackle the position shift problem. First, a region feature (RF) alignment module with adjacent similarity constraint is designed to consistently predict the position shift between two modalities and adaptively align the cross-modal RFs. Second, we propose a novel region of interest (RoI) jitter strategy to improve the robustness to unexpected shift patterns. Third, we present a new multimodal feature fusion method that selects the more reliable feature and suppresses the less useful one via feature reweighting. In addition, by locating bounding boxes in both modalities and building their relationships, we provide novel multimodal labeling named KAIST-Paired. Extensive experiments on 2-D and 3-D object detection, RGB-T, and RGB-D datasets demonstrate the effectiveness and robustness of our method.

Index Terms—Deep learning, feature fusion, multimodal object detection, pedestrian detection.

I. INTRODUCTION

ROBUST object detection is a crucial ingredient of many computer vision applications in the real world, such as robotics, surveillance, and autonomous driving. Nowadays, smart devices are usually equipped with different sensors (e.g., cameras, thermal cameras, time-of-flight, structured light, LiDAR, and radars), which provides different image forms. Compared to pure RGB data [1]–[3], integrating the extra modality is proven to be effective for recognition: infrared camera provides biometric information for around-the-clock pedestrian detection [4]–[6] and face recognition [7], [8], and depth channel provides 3-D information for object detection [9]–[11] and pose estimation [12], [13], and so on. Motivated by this, multimodal object detection has attracted massive attention, not only toward the improvement of accuracy but also for a better robustness promise.

Most existing multimodal methods [4], [11], [14] detect objects under the alignment assumption, i.e., the image pairs from different modalities are well aligned and have strong pixel-to-pixel correspondence. However, this assumption does not always hold in practice, as shown in Fig. 1, which poses great challenges to existing methods. To fill the gap between the alignment assumption and real-world applications, we discuss three main reasons behind the challenges.

First, the alignment quality of the image pair is limited by the physical properties of different sensors, e.g., geometrical disparity due to relative displacement [15], mismatched resolutions, and field-of-views. For example, the border areas of color images need to be sacrificed since the color camera usually has a higher resolution and larger field view than the thermal camera [4]. This will naturally lead to imperfect calibration and result in alignment error. Second, the calibration and recalibration processes are important, while tortuous generally requires particular hardware, such as a special heated calibration board [16]–[18], making recalibration difficult when the device is in operation. Moreover, the alignment
is easily impacted by external disturbance (e.g., mechanical vibrations and temperature variation) and hardware aging, which is hard to be avoided when a system starts to operate.

As a result, the multimodal images are often weakly aligned in practice; thus, there is a position shift between one object in different modalities, i.e., the position shift problem. The position shift problem degrades the convolutional neural network (CNN)-based detectors mainly in two aspects. First, multimodal features in the corresponding position are spatially mismatched, which will puzzle the CNN training. Second, the position shift problem makes it difficult to match objects from both input modalities by using a shared bounding box. As a workaround, existing multimodal datasets [4], [5] use larger bounding boxes to encompass objects from both modalities or annotate objects on a single input modality. Such label bias can lead to bad supervision signals and is harmful to the training process, especially for modern detectors [19]–[22], since they generally use the intersection over union (IoU) overlap for foreground/background classification.

Thus, how to robustly locate the object on weakly aligned modalities remains to be a crucial issue for multimodal object detection, while it is barely touched or studied in the literature. Besides, due to different devices, deployment environments, and operation duration, the priorities of modality information and the patterns of position shifts are ever-changing. Hence, it is a very important challenge to cope with the position shift problem using a single model.

In this article, a novel framework is proposed to tackle the position shift problem in an end-to-end fashion. Specifically, we present a novel region feature alignment (RFA) module to shift and align the to-be-fused region features (RFs) from input modalities. With this module, the region-based alignment process is inserted in the network and works in a learnable way. To further enhance the robustness to different patterns of position shift, we propose the region of interest (RoI) jitter training strategy, which augments the RoIs of the sensed modality via random jitter.

To validate the effectiveness of our method, we apply the proposed framework on the typical RGB-T (multispectral) pedestrian detection. To better meet the around-the-clock demands, we design a new confidence-aware fusion (CAF) module that selects the more reliable feature and suppresses the less useful one via adaptive feature reweighting. Besides, we provide a novel KAIST-Paired1 annotation by locating the bounding boxes in both modalities and building their relationships. We also collect a drone-based RGB-T dataset, which includes more object categories (e.g., car, bus, truck, and cyclist) to validate the generalization to RGB-T object detection. Moreover, we extend the proposed method to the RGB-D-based object detection task. Experiments on 3-D object detection are conducted on the standard NYUv2 dataset. To further validate the proposed framework, we build a dataset containing different kinds of indoor objects, called SL-RGBD, in which image pairs are labeled with multimodal bounding boxes. Extensive experimental results demonstrate that the proposed method significantly improves the robustness to position shift problem and takes full advantage of multimodal inputs by effective feature fusion, thus achieving state-of-the-art results on the challenging KAIST and CVC-14 datasets. For the general 2-D and 3-D object detection tasks, our method also achieves the best robustness performance on the NYUv2 and SL-RGBD datasets.

This article is an extension of our previous work [23] mainly in the following aspects: 1) we improve our method with the adjacent similarity constraint for the region shift prediction and extend the approach to a general multimodal object detection task, including RGB-thermal and RGB-depth detections, and 2-D and 3-D object detection; 2) the experimental upper bound of position shift for the weakly aligned image pair is introduced to specify the definition of the weakly aligned; and 3) additional experiments and analyses are conducted to better show the effectiveness and generalization ability of the proposed method.

The remainder of this article is organized as follows. We review the related work in Section II. Section III discusses the position shift problem in weakly aligned image pairs and presents our motivation. Section IV describes the proposed framework in detail. Extensive experimental results of RGB-T pedestrian detection are given in Section V. In Section VI, we further report the results of the proposed method on RGB-D-based 2-D and 3-D object detection. We conclude this article in Section VII.

II. RELATED WORK

Recent intelligence systems have access to various forms of images, such as RGB, depth, and infrared. Compared to the RGB-based method, the adoption and fusion of novel modality greatly improve the performance and robustness, thus enabling many practical applications, e.g.,

1available at https://github.com/luzhang16/AR-CNN
A. Multispectral (RGB-T) Pedestrian Detection

Pedestrian detection is an essential step for many applications and recently gets increasing attention. In previous years, many algorithms and features have been proposed, including the traditional detectors [24]–[27] and the lately dominated CNN-based detectors [28]–[32]. With the recent release of large-scale multispectral pedestrian benchmarks, efficiently exploiting the multispectral data has shown great advantages on pedestrian detection, especially for the around-the-clock operation [33]–[38]. Hwang et al. [4] propose the KAIST multispectral dataset and extend the ACF method to make full use of the aligned color–thermal image pairs. With the success of deep learning, several CNN-based methods [39]–[42] are proposed in recent years. Liu et al. [43] build a two-stream detector and experimentally analyze different fusion timings. König et al. [44] fuse features of the region proposal network (RPN) and introduce the boosted forest (BF) framework as the classifier. Xu et al. [45] propose a cross-modality learning framework that can cope with missing thermal data at testing time. Zhang et al. [40] utilize the cross-modality attention mechanism to recalibrate the channel responses of halfway feature maps. However, most previous approaches conduct multimodal fusion under the full alignment assumption. This not only hinders the use of weakly aligned datasets (such as CVC-14 [5]) but also limits the further development of multispectral pedestrian detection, which is worthy of attention but still lacks research.

B. RGB-D Object Detection

We briefly introduce the 2.5-D approaches in RGB-D images, in which the depth images are generally treated in a similar fashion as RGB images. In [46], the detector combines the histogram of oriented gradients (HOG) features of RGB data and histogram of oriented depths (HOD) of dense depth data in a probabilistic way. With the recent dominance of CNNs, Gupta et al. [9] build a CNN-based detector on the top of precomputed object segmentation candidates. They extend the R-CNN [47] to utilize depth information with HHA encoding (horizontal disparity, height above ground, and angle of local surface normal w.r.t. gravity direction). However, the outputs of detectors are still limited to 2-D ones. In order to infer the 3-D bounding box, Deng and Latecki [11] propose the Amodal3Det that further explores the strong CNN to infer 3-D dimensions in RGB-D data efficiently. Besides, the single-shot 3-D-SSD [48] is designed to achieve high-speed inference. Rahman et al. [49] downsample the RoIs to get better 3-D initialization. Different from them, we consider the position shift problem in RGB-D pairs for 3-D object detection.

C. Weakly Aligned Image Pair

Though multimodal object detection is extensively studied, few detection methods touch the position shift problem in a weakly aligned image pair. Since this problem is unavoidable in realistic scenarios due to both hardware and environmental factors, before the detection methods, it can be independently mitigated in two ways: preprocess and postprocess.

The preprocess aims to solve the camera calibration problem and elaborate a delicate modality-aligned system. However, this hardware promise is easily impacted by unavoidable external disturbance. Also, the recalibration [16], [50] process can be difficult due to the characteristics of additional modalities. For example, the calibration of RGB and thermal images generally need particular hardware and the special heated calibration board [16]–[18].

A common paradigm of the postprocess is to conduct image registration (i.e., spatial alignment) [51]–[53]. As an image-level solution for the position shift problem, it geometrically aligns the reference and sensed images. This task mainly consists of four processes: feature detection, mapping function design, feature matching, image transformation, and resampling. Although the image registration is well established, the low-level transformation on the whole image is often time-consuming and restricts end-to-end training of the CNN-based detector. Moreover, since different modalities present different appearances, the registration process faces great challenges because the correspondence of key points is harder to determine.

III. Motivation

In this section, we present our analysis of the position shift problem in the weakly aligned image pair. To provide insights into this problem, we first analyze the popular KAIST [4] and CVC-14 [5] RGB-T pedestrian datasets and the structured-light-based RGB-D system. Then, we experimentally study how the position shift deteriorates the detection performance.

A. Position Shift Problem

The position shift is defined as the spatial displacement between two images of different cameras. The image pair is taken in the same scene and time, generally including the same objects. However, caused by the position shift problem, the pixel-to-pixel relationship does not hold, and two corresponding image patches can be located in different positions.

1) RGB-T Pair: From the image pairs in KAIST and CVC-14 RGB-T pedestrian datasets and the drone-based object dataset, we can observe several issues.

1) Misaligned Features: Fig. 2(a) illustrates the position shift between modalities; this problem makes it difficult to directly fuse such misaligned features. Besides, the shift varies in different positions even in a single image. In Fig. 3, we present the statistics of position shift in two datasets, which show that collection devices and operation environments also influence the shift pattern.

2) Localization Bias: As shown in Fig. 2(a), the annotations need to be adjusted to match the weakly aligned image pair. One way to remedy is using larger bounding boxes to encompass objects from both modalities but also including the unnecessary background information. Another workaround is only focusing on one particular modality, but this could introduce bias for another modality.
Fig. 2. Some instances of original annotations in the KAIST (yellow boxes), CVC-14 (orange box), and drone-based (white boxes) datasets. We crop the image patches in the same position, but the objects pose poor alignment. Thus, such a shared bounding box for both modalities cannot precisely locate the object. (a) Poor alignment. (b) Unpaired objects.

3) **Unpaired Objects:** In practice, color and thermal cameras often have different field-of-views; the situation is even worse when the synchronization and calibration are not good. Therefore, as shown in Fig. 2(b), the objects appear in one modality but are truncated or lost in another. Specifically, ∼12.5% (2245 of 18,017) of pedestrians are unpaired in CVC-14 [5].

2) **RGB-D Pair:** Around different types of 3-D sensing systems, the projector-camera-based structured light methods have been more and more important for 3-D vision-related applications. A typical structured light system is composed of one projector and one industry camera. Since patterns used by the structured light systems are usually monochrome or binary, monochrome cameras are preferred. In real applications, to obtain high-quality color texture for the reconstructed 3-D models, an extra high-resolution RGB camera, such as a digital single-lens reflex (DSLR) camera, is usually added to the system [54]. As a result, alignment of the high-resolution color image and the low-resolution point cloud or depth image is essential for any structured light systems that used an extra color texture camera.

To reveal the realistic position shift problem of RGB-D data, we build such an RGB-D data collection system with a pair of color and structural light cameras, as shown in Fig. 4. We can see that, even though the system is well-calibrated, time and external disturbance will degrade the quality of the calibration. As demonstrated in Fig. 4, the main problem is position shift, along with milder deformation and rotation.

**IV. OUR METHOD**

In this section, we introduce the aligned region CNN (AR-CNN) framework and KAIST-Paired annotation (see Section IV-C1). The architecture is illustrated in Fig. 6, containing the RFA module (see Section IV-A) and the RoI.
We utilize the RPN [19] for the proposal generation process. Fig. 6. Architecture of the proposed AR-CNN. We use color–thermal input as an example; after the two-stream feature extractor, numerous proposals are generated by the fused RPN. Then, the RFA module is utilized to predict and align the shift of regions, and an extra shift loss is calculated by utilizing the proposed KAIST-Paired annotation. Meanwhile, we perform the RoI Jitter strategy to enrich the shift patterns during training. After that, we can pool the aligned RFs from the feature map of each modality. For the all-day RGB-T pedestrian detection, we utilize the CAF method to pay attention to the more reliable modality, and $W_r$ and $W_s$ denote the confidence weight of reference and sensed modality, respectively. For the 3-D RGB-D object detection, we calculate the multitask loss based on the 3-D box initialization and regression process.


drolley training strategy (see Section IV-B). We also present the CAF step (see Section IV-C1) for all-day RGB-T pedestrian detection and the adaption to RGB-D-based 3-D object detection task (see Section IV-D). Algorithm 1 describes the whole pipeline.

A. Region Feature Alignment

To predict and align the shift between two modalities, we present the RFA module in this section. In practice, the position shift is not a simple affine transformation due to the different properties of camera systems. As a result, the shift varies in different regions. It is usually large in regions near the image edge and relatively small near the image center. Therefore, the RFA module performs the shift prediction and feature alignment at the regional level.

1) Reference and Sensed Modality: In image registration [51], [52], the reference and sensed images refer to the stationary and transformed images, respectively. We introduce this concept into our multimodal setting. In the training phase, the reference modality is fixed, and the feature alignment and RoI jitter strategy are performed on the sensed modality.

2) Proposals’ Generation: To keep more potential objectness regions, we aggregate both the reference and sensed feature maps (Conv4_r and Conv4_s) to generate the proposals. We utilize the RPN [19] for the proposal generation process.

3) Feature Alignment: Fig. 7 illustrates the concrete connection scheme of the RFA module. Specifically, the RFA module first enlarges the proposals (RoIs) to encompass the contextual information. Then, we pool the contextual RFs of each modality to small $H \times W$ (such as $7 \times 7$) feature maps. Then, the pooled RFs are concatenated to obtain the cross-modal representation. Based on this representation, we use two consecutive fully connected layers to predict the shift (i.e., $t_x$ and $t_y$) of each region. In this way, we obtain new coordinates of the sensed region and repool on this new region to get the aligned sensed features.

Since the proposed KAIST-Paired annotation (more details in Section IV-C1) provides multimodal bounding boxes, we can calculate the targets of ground-truth shift as follows:

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

where $x$ and $y$ are the center coordinates of the box, $w$ and $h$ refer to the width and height, $x_r$ and $x_s$ denote the reference and sensed ground-truth box, and $t_x^*$ and $t_y^*$ indicate the shift target for $x$ and $y$ coordinates, respectively.

4) Adjacent Similarity Constraint: For natural images, the position shift is spatially smooth, i.e., the shift targets for adjacent regions tend to be similar. Base on this, we add the adjacent similarity constraint to stabilize the training process of the RFA module. Specifically, we first randomly sample one of four nearest pixels (with feature stride) to the pooled features maps of the RoI. Then, the same alignment targets are assigned to the RoI and its neighbor; thus, it encourages the RFA module to predict similar for adjacent regions.

5) Position Shift Loss: To measure the accuracy of predicted shift targets, we calculate the position shift loss as follows:

$$ L_{\text{shift}}(\{p_i^*\}, \{t_i\}, \{t_i^*\}) = \frac{1}{N_{\text{shift}}} \sum_{i=1}^n p_i^* \text{smooth} L_1(t_i - t_i^*). \quad (2) $$

The subscript $i$ denotes the index of RoIs in minibatch, $p_i^*$ is the class label (1 for the pedestrian and 0 for the background) of the RoI, $t_i$ indicates the predicted shift, and $t_i^*$
Algorithm 1 Framework of the Multimodal Object Detection System

Input:
The image of reference modality, \( I_r \);
The image of sensed modality, \( I_s \);
The output threshold \( \tau \);

Output:
A set of detection results \( D_r \) on reference images;
1: Extracting the feature maps \( \text{Conv}_r \) and \( \text{Conv}_s \) from the reference and sensed modality;
2: Aggregating \( \text{Conv}_r \) and \( \text{Conv}_s \) to generate a set of 2D proposals \( P \);
3: Extracting region-wise features \( \mathcal{F}_r \) and \( \mathcal{F}_s \) of \( P \) by using RoI pooling;
4: Predicting the position shift for \( P \) using the RFA module and re-extracting aligned region-wise features \( \mathcal{F}_r^j \);
5: Fusing region-wise features \( \mathcal{F}_r \) and \( \mathcal{F}_s \) with confidence-aware module to obtain \( \mathcal{F}_f \);
6: Conducting region-wise classification and (3D) bounding boxes regression to obtain confidence \( C^i \) and coordinates \( B^i \) for the \( F^i_f \in \mathcal{F}_f \) of each proposal \( P^i \in P \);
7: \( D_r \leftarrow \emptyset \);
8: for each \( P^i \in P \) do
9: \quad if \( C^i > \tau \) then
10: \quad \quad \( D_r \leftarrow D_r \cup \{ [B^i, C^i] \} \);
11: \quad end if
12: end for
13: return \( D_r \);

is the ground-truth shift target. \( N_{\text{shift}} \) denotes the number of to-be-aligned RoIs.

6) Multitask Loss: Then, the multitask loss function is defined as follows:

\[
L((p_i), \{t_i\}, \{g_i\}, \{p^*_i\}, \{t^*_i\}, \{g^*_i\}) = L_{\text{cls}}((p_i), \{p^*_i\}) + \lambda_1 L_{\text{shift}}(\{p^*_i\}, \{t_i\}, \{t^*_i\}) + \lambda_2 L_{\text{asc}}(\{p^*_i\}, \{\hat{t}_i\}, \{t^*_i\}) + L_{\text{reg}}(\{p^*_i\}, \{g_i\}, \{g^*_i\}).
\]

\( L_{\text{cls}} \) and \( L_{\text{reg}} \) are similar to the loss in Fast R-CNN [57]. \( L_{\text{shift}} \) is the loss of the adjacent similarity constraint, and \( \hat{t}_i \) is the predicted shift of the four-neighborhood. Variable \( p_i \) refers to the predicted class confidence for the \( i \)th RoI, \( g_i \) is the predicted refinement, and \( p^*_i \) and \( g^*_i \) are associated ground truths. We balance \( L_{\text{shift}} \) and \( L_{\text{reg}} \) by setting \( \lambda \) as the weighting parameter. To weight the two terms, we use \( \lambda_1 = 0.75 \) and \( \lambda_2 = 0.25 \). Apart from this multitask loss, the definition of RPN loss follows the literature [19].
in the daytime yet fade in the nighttime; thermal images provide the pedestrian silhouette around the clock but lack visual details, such as the clothing. To make full use of the characteristics between two different modalities, we propose the CAF method to select the more reliable feature while suppressing the less useful one via feature reweighting.

As illustrated in Fig. 9, the CAF module uses two confidence weights \( W_r \) and \( W_f \). To calculate the two weights, we add a branch for each modality to obtain classification score \( p \). Then, the confidence weights for each modality are calculated as \( W_r = |p^r_1 - p^b_1| \) and \( W_f = |p^f_1 - p^b_1| \), where \( p^1 \) and \( p^b \) are the probability of foreground and background, \( r \) denotes the reference modality, and \( s \) is the sensed one. Then, this module performs the multiplication to pay more attention to the reliable modality.

**Unpaired Objects:** To mitigate the ambiguous classification for unpaired objects, we use the disagreement weight, \( W_d = 1 - |p^r_1 - p^b_1| = 1 - |p^f_1 - p^b_1| \). If the sensed modality provides a contradictory prediction with the reference, its feature will be suppressed.

### D. 3-D RGB-D Object Detection

To better understand the physical 3-D world, object detection also aims to predict the 3-D location and its full extent in 3-D space. In this article, we have RGB-D image pairs as inputs and leverage the 2.5-D representation to predict 3-D bounding boxes.

1) **2-D RoI Proposals:** Since RGB and depth images contain complementary information, we use both modalities to generate the proposals in 2-D space. As in Section IV-A, the RPN is adapted to generate 2-D proposals in an end-to-end fashion.

2) **3-D Bounding Box Initialization and Regression:** Given several 2-D proposals, we use some basic transformations to initialize the related 3-D bounding boxes. Following the settings in [11], each 3-D bounding box is parameterized into one seven-entry vector \([x_{cam}, y_{cam}, z_{cam}, l, w, h, \theta]\). \([x_{cam}, y_{cam}, z_{cam}]\) is the centroid under camera coordinate system, and \([l, w, h]\) is the 3-D dimension. The parameter \( \theta \in [-\pi/2, \pi/2] \) is the angle between the principal axis and its orientation vector under the tilt coordinate system, in which the point clouds are aligned with gravity direction. Hence, this rotation is only around the gravity direction.

For initialization, the median of depth value in each individual region is used as the initialization of the \( z \)-axis, noted as \( z_{ini} \), and then, \( x_{ini} \) and \( y_{ini} \) can be calculated as follows:

\[
\begin{align*}
x_{ini} &= z_{ini} \times (c_x - o_x) / f \\
y_{ini} &= z_{ini} \times (c_y - o_y) / f
\end{align*}
\]

where \( f \) is the focal length of RGB camera, \((o_x, o_y)\) is the principal point, and \((c_x, c_y)\) is the center coordinate of 2-D box proposal. The 3-D size \([l, w, h]\) is set as the classwise average dimension, which is inspired by the familiar size in human 3-D perception [11], [60], [61]. The angle \( \theta \) is set to 0 by default.

With the seven-entry targets prediction between initialized vector and targeted vector, i.e., \([v_x, v_y, v_z, v_l, v_w, v_h, v_\theta]\), the 3-D regression loss can be calculated as follows:

\[
L_{3dreg}(p^*, v_{3d}, v^*_{3d}) = \left[p^* > 1\right] \text{smoothL1}(v_{3d} - v^*_{3d})
\]

where \( p^* \) is the ground-truth class index, \( v^*_{3d} \) are the ground-truth regression offsets, and \( v_{3d} \) are the predicted regression targets.

### V. Experiments on Case 1: RGB-T Object Detection

In this section, we first conduct several experiments on the challenging KAIST [4] and CVC-14 [5] RGB-T pedestrian datasets. We evaluate all methods on the “reasonable” setup in which pedestrians are not or partially occluded and the heights are larger than 55 pixels. Then, to verify the generalization of our method on object detection, we collect a drone-based RGB-T dataset that includes four categories of objects: car, truck, bus, and cyclist.

#### A. Dataset

1) **KAIST:** The KAIST multispectral pedestrian dataset [4] includes 95 328 color–thermal frame pairs with 103 128 annotations. To cover diverse illumination conditions, the dataset is captured in different scenes, through day and night. For the testing, we use the test set containing 2252 image pairs sampled from the video with 20-frame skips.

2) **CVC-14:** The CVC-14 dataset [5] uses both visible and FIR cameras to record color–thermal video sequences during day and night time. It contains 8518 frames, among which 7085 frames are for training and 1433 for testing. The two datasets are not in a well-calibrated condition. In consequence, this dataset suffers more from the position shift problem, as shown in Fig. 3(b). This problem makes it difficult for state-of-the-art multimodal detectors to apply to the CVC-14 dataset.

3) **Drone-Based Dataset:** The drone-based dataset is collected from day to night using a DJI MATRICE 300 RTK drone with precalibrated RGB-T cameras, covering different scenes such as city, suburbs, and countryside. It contains 831 pairs of all-day images with 6763 pairs of dense annotations.
for four object categories: car, truck, bus, and cyclist. Though the RGB-T cameras are precalibrated, there is still the position shift problem due to the different physical properties of cameras and external disturbance in operation.

### B. Implementation Details

We use ImageNet [65] pretrained VGG-16 [66] as the backbone network. By default, $\sigma_0$ and $\sigma_1$ in the RoI jitter are set to 0.05, which can be adjusted to tackle a wider or narrower position shift. We horizontally flip all the images for data augmentation. We set the initial learning rate to 0.0005 for 9 epochs and 0.0005 for 1 more epoch. For the drone-based dataset, we use an initial learning rate of 0.0005 for 9 epochs and 0.005 for 3 epochs and 0.0005 for 1 more epoch. For the drone-based dataset, we use an initial learning rate of 0.0005 for 9 epochs and decay it by 0.1 for another 3 epochs.

We utilize the log-average MR to measure the detection performance, a lower score indicates better performance. We plot the MR averaged against the false positives per image (FPPI) over the log-space range of $[10^{-2}, 10^0]$. For the test set, we use the widely adopted improved annotation, which is proposed by Liu et al. [43]. Besides, based on our KAIST-Paired annotation, we use $MR_C$ and $MR_T$ to denote the MR for color and thermal modality, respectively. For the drone-based dataset, we use mean Average Precision (mAP) to evaluate the performance of detectors.

### C. Comparison Experiments

1) KAIST: The proposed AR-CNN is evaluated and compared to other available competitors [4], [40], [42]–[44], [62]. As shown in Fig. 10, our method achieves 9.79, 8.07, and 9.03 MR on a reasonable day, night, and all-day subset, respectively, achieving the state-of-the-art performance. Besides, in consideration of the position shift problem, the methods are also evaluated with the KAIST-Paired annotation, i.e., $MR_C$ and $MR_T$. As shown in Fig. 10, the proposed method has significant advantages, i.e., 8.52 versus 11.28 $MR_C$ and 8.01 versus 12.51 $MR_T$, demonstrating the superiority of our method.

2) CVC-14: Our experiments on the CVC-14 dataset are following the protocol in [58]. As shown in Table I, compared to other competitors, the proposed method achieves the best performance. The greater advantage on the night subset (18.1 versus 30.8 MR) validates the contribution of thermal modality and shows that the detection accuracy can be significantly improved by appropriately handling the position shift problem.

3) Drone-Based Dataset: As shown in Table II, compared to the available competitors, our approach achieves the best 36.4 mAP, which demonstrates the generalization ability from pedestrian detection to the wider object detection.

### D. Robustness to Position Shift

Most multimodal systems suffer from the position shift problem, but the degrees of the shift are varying with different devices and settings. To clearly set the experiments, we introduce the empirical upper bound of position shift for the weak aligned image pair.

Since the pedestrian detection and object detection tasks use evaluation metric, i.e., log average MR (lower is better) and mAP (higher is better), we, respectively, define the performance degradation rate $R_d$ as follows:

$$R_d = (MR_{\text{degraded}} - MR_{\text{original}})/MR_{\text{original}} \quad \text{(7)}$$

$$R_d = (\text{mAP}_{\text{original}} - \text{mAP}_{\text{degraded}})/\text{mAP}_{\text{original}}. \quad \text{(8)}$$

In this article, we set the shift of 50% performance degradation ($R_d = 0.5$) as the weak aligned bound. For instance, if shifting 10 and $-15$ pixels along the $x$-axis leads to 50% performance degradation, then the two upper bounds $B_{a1}$ and $B_{a2}$ on this direction are 10 and $-15$. In experiments, we calculate this bound based on the Halfway Fusion [43] model.

First, the thermal modality is set as the reference since it usually provides consistent images throughout the day. We use $MR_T$ to evaluate the detector. Following the settings in Section III-B, the visual results in a surface plot are depicted in Fig. 5(b). As illustrated in Fig. 5, compared to the baseline method, the proposed detector significantly enhances the robustness to position shift, and the performance at the origin is also improved. Besides, we design four metrics to quantitatively evaluate the robustness, i.e., $S^0$, $S^{45}$, $S^{90}$.

---

**Table I**

| Method       | MR     |
|--------------|--------|
|              | Day    | Night | All   |
| SVM [5]      | 37.6   | 76.9  | -     |
| DPM [5]      | 25.2   | 76.4  | -     |
| Random Forest [5] | 26.6 | 81.2  | -     |
| ACF [58]     | 65.0   | 83.2  | 71.3  |
| Faster R-CNN [58] | 43.2 | 71.4  | 51.9  |
| MACF [58]    | 61.3   | 48.2  | 60.1  |
| Choi et al. [59] | 49.3 | 43.8  | 47.3  |
| Halfway Fusion [58] | 38.1 | 34.4  | 37.0  |
| Park et al. [58] | 31.8 | 30.8  | 31.4  |
| AR-CNN (Ours) | 24.3  | 18.1  | 22.0  |

**Table II**

| Method       | mAP   | car | bus | truck | cyclist |
|--------------|-------|-----|-----|-------|---------|
| SSD [20]     | 23.6  | 40.6| 31.7| 21.3  | 6.8     |
| Faster R-CNN [19] | 26.7 | 44.7| 30.7| 21.9  | 9.5     |
| Halfway Fusion [43] | 31.9 | 52.6| 37.7| 25.0  | 12.4    |
| CIAN [40]    | 34.8  | 57.0| 40.2| 28.7  | 13.3    |
| AR-CNN (Ours) | 36.4  | 59.8| 40.6| 29.9  | 15.1    |
Fig. 10. Comparisons of the MR curves reported on the KAIST dataset. The performance is evaluated by the MR protocol, and also MR\textsuperscript{C} and MR\textsuperscript{T} using the KAIST-Paired annotation. (a) Day, MR. (b) Night, MR. (c) All-day, MR. (d) Day, MR\textsuperscript{C}. (e) Night, MR\textsuperscript{C}. (f) All-day, MR\textsuperscript{C}. (g) Day, MR\textsuperscript{T}. (h) Night, MR\textsuperscript{T}. (i) All-day, MR\textsuperscript{T}.

TABLE III
D\textsc{ETECTION PERFORMANCES AND ROBUSTNESS TO POSITION SHIFT ON THE KAIST D\textsc{ATASET}.} O \textsc{Refer}s to MR\textsuperscript{T} S\textsc{CORES AT THE ORIGIN, AND µ AND σ \textsc{Denote}} the mean and standard deviation of MR\textsuperscript{T}. \textsc{W}e \textsc{use the reimplemented model of the literature [43], [44] and the provided model of the literature [40], [64].}

| Method                  | \( S^0 \)  | \( S^{45} \) | \( S^{90} \)  | \( S^{135} \) |
|-------------------------|------------|-------------|-------------|-------------|
|                         | \( O \) \( \mu \) \( \sigma \) | \( \mu \) \( \sigma \) | \( \mu \) \( \sigma \) | \( \mu \) \( \sigma \) |
| Halfway Fusion [43]     | 25.51      | 35.72       | 35.19       | 34.07       |
| Fusion RPN [44]         | 21.43      | 31.57       | 30.42       | 29.96       |
| Adapted Halfway Fusion  | 15.59      | 25.02       | 25.85       | 22.85       |
| CIAN [40]               | 14.68      | 24.50       | 22.97       | 20.01       |
| MSDS-RCNN [64]          | 12.51      | 22.19       | 22.32       | 20.10       |
| RFA RoI CAF             |            |             |             |             |
| AR-CNN                  | 12.94      | 22.27       | 14.76       | 19.34       |
|                        | 10.90      | 12.20       | 11.85       | 16.87       |
|                        | 9.87       | 11.74       | 11.01       | 15.92       |
|                        | 8.26       | 9.41        | 9.52        | 9.66        |
| + ASC \textsc{S} \textsuperscript{135} | \textbf{8.01} | \textbf{9.17} | \textbf{9.82} | \textbf{9.66} |
TABLE IV
DETECTION ROBUSTNESS TO THERMAL POSITION SHIFT (i.e., WE FIX THE COLOR IMAGE WHILE SHIFTING THE THERMAL IMAGE) ON THE KAIST DATASET. MR^C IS USED TO EVALUATE THE DETECTION PERFORMANCE

| Method                     | $S_0^\circ$ | $S_{45}^\circ$ | $S_{90}^\circ$ | $S_{135}^\circ$ |
|----------------------------|-------------|----------------|----------------|-----------------|
|                            | O | µ  | σ  | µ  | σ  | µ  | σ  | µ  | σ  | µ  | σ  |
| Halfway Fusion [43]        | 25.10 | 34.20 | 7.98 | 34.80 | 7.87 | 32.79 | 7.36 | 33.85 | 8.16 |
| Fusion RPN [44]            | 20.52 | 30.22 | 9.07 | 29.15 | 9.23 | 25.12 | 6.95 | 29.97 | 9.80 |
| Adapted Halfway Fusion     | 15.06 | 24.39 | 8.19 | 25.91 | 11.32 | 21.34 | 7.25 | 26.58 | 11.52 |
| CIAN [40]                  | 14.64 | 23.86 | 8.74 | 24.87 | 11.08 | 19.16 | 6.22 | 25.20 | 10.46 |
| MSDS-RCNN [64]             | 11.28 | 20.06 | 7.95 | 20.81 | 9.54 | 16.61 | 6.36 | 21.73 | 10.14 |
| Ours [23]                  | 8.86  | 12.51 | 1.60 | 11.52 | 2.07 | 11.27 | 1.41 | 11.83 | 2.22 |
| + ASC                      | 8.52  | 11.79 | 1.28 | 10.58 | 1.90 | 11.31 | 1.47 | 11.03 | 2.16 |

metrics, demonstrating the robustness to diverse position shift conditions.

Experiments on the Color Reference: Then, we fix the color image as the reference modality. Table IV shows that the proposed method still achieves the best performance and smallest standard deviation, further validating the effectiveness of the AR-CNN. In addition, compared to the thermal reference, the color reference configuration performs at a lower level in experiments. This validates our intuition: the modality with stable imagery is more suitable to serve as the reference one.

E. Ablation Study

To analyze our model in more detail, we conduct ablation studies on the KAIST dataset.

1) Region Feature Alignment: Table III shows detection results with and without the RFA module. It can be observed that the MR and $\sigma$ under various position shift conditions are remarkably reduced by the RFA module. Specifically, for $S_{45}^\circ$, the RFA reduces $\sigma$ by a significant 6.84 (i.e., from 8.92 to 2.08). For the other three metrics, consistent reductions are also observed. Moreover, some visualizations of proposals and detection results are shown in Fig. 11. For pedestrians with the position shift problem, proposals of the sensed (color) modality are adjusted to an aligned position. This phenomenon demonstrates that the RFA module can predict the regionwise position shift of two modalities and adaptively adjust the sensed proposals, thus enabling the modality-aligned feature fusion process for better classification and localization.

2) RoI Jitter Strategy: Then, we demonstrate the contribution of the RoI jitter. Table III shows that the strategy further improves the detector’s robustness. Specifically, for $S_0^\circ$, $\mu$ is reduced from 12.20 to 11.74, and $\sigma$ is decreased from 2.92 to 1.29. Meanwhile, it can be observed that this strategy is more helpful for the $\sigma$ than the MR at the origin, which demonstrates that the main contribution of this strategy is boosting the robustness to the position shift. Table V further shows the detection results with varying hyperparameters $\sigma_0$ and $\sigma_1$. When $\sigma_0$ and $\sigma_1$ are set to 0.05, we can obtain the best overall score. When varying the hyperparameter, the proposed method exhibits stable performances that are better than the baseline.

3) Confidence-Aware Fusion: We also compare performances with and without the CAF module to validate the effectiveness of our multimodal fusion scheme. Table III shows that this module significantly reduces MR^C at the origin by 1.61 yet slightly suppresses $\sigma$. This illustrates that the main contribution of the CAF module is improving the detection performance since it conducts adaptive fusion by paying more attention to reliable features.

Besides, we also visualize some failed cases in Fig. 12 to bring insights for the error analysis and further improvement of the detector. Generally, the far-scale tiny people are easier to be missed since it is difficult to obtain sufficient features after the CNN pooling. We also notice that the detection performance will degrade if the environment is complex, e.g., with a cluttered background and a cluster of objects, due to disturbed or incomplete features. In the future, we would like...
Fig. 11. Qualitative results of the proposed approach. The first row illustrates the reference proposals whose confidence score (in the range $[0, 1.0]$) is greater than 0.6, and the second row shows the corresponding sensed proposals. To demonstrate the effectiveness of the RFA module, some proposal instances are shown in the third row: orange dotted boxes denote the reference proposals, which are good ones in the reference image but suffer the position shift problem in the sensed modality; red bounding boxes refer to the adjusted sensed proposals after the RFA process. In the last two rows, green bounding boxes show the final detection results whose confidence score is greater than 0.6.

Fig. 12. Visualization of the failed cases on the KAIST test set. Red boxes represent the ground truth, yellow boxes denote ignored objects, and green boxes refer to detected objects whose confidence scores are greater than 0.3.

to work on the occluded person by enhancing the classification and localization heads to further improve the detection performance.

VI. EXPERIMENTS ON CASE 2: RGB-D OBJECT DETECTION

In this section, we conduct experiments of 2-D object detection on the SL-RGBD dataset and 3-D object detection on the NYUv2 [68] dataset.

A. Dataset

1) NYUv2: The NYU Depth V2 dataset [68] is a popular yet challenging dataset for indoor scene understanding, which contains 1449 RGB-D pairs with 19 indoor object classes. To achieve 3-D object detection, Song et al. [69] extended the NYUv2 dataset with extra 3-D bounding boxes and refine the depth map by integrating multiple frames of raw video data. In [11], the 3-D annotations are further refined by addressing issues about inconsistency and inaccuracy.
Fig. 13. Data instances and detection results in the SL-RGBD dataset. The first row shows RGB images, and the second row depicts corresponding depth images that suffer from different degrees of the position shift problem.

Fig. 14. Visualization of results in weakly aligned image pairs. The first and second rows are aligned RGB-D image pairs in which the depth images are refined by integrating multiple frames from the raw video. The third row shows different patterns of the position shift problem. The last three rows depict 3-D bounding boxes from ground truth, Halfway Fusion, and AR-CNN, respectively. We show the boxes whose confidence score (in the range $[0, 1.0]$) is greater than 0.6. For better visualization, the point cloud is generated by aligned RGB and depth images.
C. Performance and Robustness

For training and testing, the improved 3-D annotation provided by Deng and Latecki [11] is used. Following previous works [11], [48], [49], we use the true positive if the IoU is greater than 0.5 and 0.25, respectively. Following previous works [11], 2-D and 3-D object detection, the detected box is treated as pedestrian detection. We use the ImageNet pretrained model and fine-tune the same training schedule as that in the all-day RGB-T pedestrian detection experiments on our SL-RGBD dataset. As shown in the third row, our model achieves better performance than its Halfway Fusion counterpart. For example, the performances of our model and Halfway fusion are 32.1 versus 23.3 mAP when the shift pattern is (50, −50) on the image plane.

B. Implementation Details

The hyperparameter is set the same as in RGB-T pedestrian detection. We use the ImageNet pretrained model and fine-tune it on the NYUv2 and SL-RGBD datasets, respectively. We use the same training schedule as that in the all-day RGB-T pedestrian detection.

We use mAP to evaluate the performance of detectors. For 2-D and 3-D object detection, the detected box is treated as true positive if the IoU is greater than 0.5 and 0.25, respectively. Following previous works [11], [48], [49], we use the improved 3-D annotation provided by Deng and Latecki [11] for training and testing.

C. Performance and Robustness

1) 2-D Object Detection: We conduct the 2-D object detection experiments on our SL-RGBD dataset. As shown in Table VI, incorporating the depth modality significantly improves the detection performance (58.7 versus 52.9 mAP), and our AR-CNN model achieves the best 66.3 mAP results since the position shift problem is taken into consideration.

2) 3-D Object Detection: To further validate the effectiveness of our model, we test the performance and robustness of AR-CNN on the RGB-D-based 3-D object detection task. Similar to that for 2-D pedestrians, we alternatively fix the color and depth image and then manually shift the other. Table VII shows the results on the standard NYUv2 dataset, which uses color images as the reference modality. Our AR-CNN achieves comparative performance and best robustness (33.5 versus 30.2 mean performance and 3.8 versus 6.7 standard variances in $S^{90}$), which demonstrates the validness and generalization ability of the proposed method.

In Fig. 14, we use some examples to further show the model’s robustness to the position shift problem. When the system suffers from the position shift problem, as shown in the third row, our model achieves better performance than its Halfway Fusion counterpart. For example, the performances of our model and Halfway fusion are 32.1 versus 23.3 mAP when the shift pattern is (50, −50) on the image plane.

2) SL-RGBD: The SL-RGBD dataset is built for indoor recognition, which contains 214 RGB-D image pairs with 1297 object annotations; examples are shown in Fig. 13. The object classes are selected from a subset of the YCB benchmark, consisting of lemon, banana, strawberries, orange, scissors, plastic bolt and nut, and keys. Various patterns of position shift are included to show the characteristics of weakly aligned RGB-D pairs. The details of the collection system are in Section III-A.

| Method                  | $S^{90}$ | $S^{135}$ |
|-------------------------|----------|-----------|
|                          | $O$ | $\mu$ | $\sigma$ | $O$ | $\mu$ | $\sigma$ | $O$ | $\mu$ | $\sigma$ |
| Amodal3Det (RGB only) [11] | 30.0 | - | - | 31.5 | 6.9 | 31.0 | 7.1 |
| DSS [67]                 | 36.3 | - | - | 30.7 | 7.0 | 28.7 | 7.5 |
| 3D-SSD [48]              | 39.7 | - | - | 30.1 | 7.0 | 28.7 | 7.5 |
| Rahman et al. [49]       | 43.1 | 30.2 | 7.3 | 31.3 | 7.1 | 29.8 | 7.0 |
| Amodal3Det [11]          | 40.9 | 29.2 | 7.0 | 30.7 | 7.0 | 28.7 | 7.5 |
| Halfway Fusion [43]      | 41.0 | 29.0 | 6.7 | 30.1 | 7.1 | 29.8 | 7.2 |
| AR-CNN (ours)            | 41.2 | 33.5 | 3.8 | 34.3 | 4.0 | 34.2 | 3.9 |

VII. Conclusion

To handle the practical position shift problem in multimodal object detection, we propose a novel framework to improve the robustness of the detector when using weakly aligned image pairs. First, we design a novel RFA module to predict the shift and align the multimodal features. Second, to further enrich the shift patterns, we propose an RoI-level augmentation strategy named RoI jitter. For the all-day RGB-T pedestrian detection, we present new multimodal labeling named KAIST-Paired and propose the CAF method to pay attention to the more reliable modality. The proposed method achieves the best performance and robustness on CVC-14 and KAIST datasets. For the RGB-D-based 2-D and 3-D detection, we extend our method by aligning the features of 2-D proposals and then conducting a more reliable 3-D bounding boxes regression. Compared to the state of the art, the proposed method achieves comparative performance and better robustness on NYUv2 and SL-RGBD datasets. In the real-world scenario, the weakly aligned characteristic usually exists when using the multimodal system. The proposed method provides a generic solution for multimodal object detection, especially when it faces the challenges of the position shift problem.

REFERENCES

[1] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014, pp. 740–755.
[2] G.-S. Xia et al., “DOTA: A large-scale dataset for object detection in aerial images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3974–3983.
[3] Y. Pang, J. Cao, Y. Li, J. Xie, H. Sun, and J. Gong, “TJU-DHD: A diverse high-resolution dataset for object detection,” IEEE Trans. Image Process., vol. 30, pp. 207–219, 2021.
et al.

[28] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, “What can help pedestrian detection?”, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3127–3136.

[29] J. Hosang, M. Omran, R. Benenson, and B. Schiele, “Taking a deeper look at pedestrians?”, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 4073–4082.

[30] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, “Reputation loss: Detecting pedestrians in a crowd?”, in Proc./IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7774–7783.

[31] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, “Occlusion-aware R-CNN: Detecting pedestrians in a crowd?”, in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 657–653.

[32] J. Xie, Y. Pang, H. Cholakkal, R. Anwer, F. Khan, and L. Shao, “PSC-net: Learning part spatial co-occurrence for occluded pedestrian detection,” Sci. China Inf. Sci., vol. 64, no. 2, pp. 1–13, Feb. 2021.

[33] N. Kim, Y. Choi, S. Hwang, and I. S. Kweon, “Multispectral transfer network: Unsupervised depth estimation for all-day vision,” in Proc. Adv. Neural Conf. Artif. Intell. (ICONIP), vol. 32, no. 1, pp. 6983–6991.

[34] Y. Choi, N. Kim, S. Hwang, and I. So Kweon, “Thermal image enhancement using convolutional neural network,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2016, pp. 223–230.

[35] Y. K. Choi, N. I. Kim, K. B. Park, S. M. Hwang, J. S. Yoon, and I. S. Kweon, “All-day visual place recognition: Benchmark dataset and baseline,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRw), Jun. 2015, pp. 1–9.

[36] K. Zhou, L. Chen, and X. Cao, “Improving multispectral pedestrian detection by addressing modality imbalance problems,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 1–19.

[37] A. Wolpert, M. Teutsch, M. S. Sarfraz, and R. Stiefelhagen, “Anchor-free small-scale multispectral pedestrian detection,” in Proc. Brit. Mach. Vis. Conf. (BMVC), 2020, pp. 1–14.

[38] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, “Multispectral fusion for object detection with cyclic fuse-and-refine blocks,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2020, pp. 276–280.

[39] J. Wagner, V. Fischer, M. Herman, and S. Behnke, “Multispectral pedestrian detection using deep fusion convolutional neural networks,” in Proc. 24th Eur. Symp. Artif. Neural Netw. Comput. Intell. Mach. Learn. (ESANN), 2016, pp. 509–514.

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

[41] D. Guan, Y. Cao, J. Yang, Y. Cao, and C.-L. Tisse, “Exploiting fusion architectures for multispectral pedestrian detection and segmentation,” Appl. Opt., vol. 57, no. 18, pp. D108–D116, 2018.

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

[43] J. Liu, S. Zhang, S. Wang, and D. Metaxas, “Multispectral deep neural networks for pedestrian detection,” in Proc. Brit. Mach. Vis. Conf., 2016, pp. 73.1–73.13.

[44] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, “Fully convolutional region proposal networks for multispectral pedestrian detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 243–250.

[45] D. Wu, W. Ouyang, S. Cui, and S. Liu, “Learning cross-modal deep representations for robust pedestrian detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5363–5371.

[46] L. Spinello and K. O. Arras, “People detection in RGB-D data,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., Sep. 2011, pp. 3838–3843.

[47] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[48] Q. Luo, H. Ma, L. Tang, Y. Wang, and R. Xiong, “3D-SDS: Learning hierarchical features from RGB-D images for amodal 3D object detection,” Neurocomputing, vol. 378, pp. 364–374, Feb. 2020.

[49] M. M. Rahman, Y. Tan, J. Xue, L. Shao, and K. Lu, “3D object detection: Learning 3D bounding boxes from scaled down 2D bounding boxes in RGB-D images,” Sci. China Inf. Sci., vol. 47, pp. 145–158, Feb. 2019.

[50] S. Vidas, R. Lukemond, S. Denman, C. Fookes, S. Sridharan, and T. Windridge, “A mask-based approach for the geometric calibration of thermal-infrared cameras,” IEEE Trans. Instrum. Meas., vol. 61, no. 6, pp. 1625–1635, Jun. 2012.

[51] B. Zitová and J. Flusser, “Image registration methods: A survey,” in Vis. Comput., vol. 21, no. 6, pp. 520–530, 1998.
Z. Song, R. Chung, and X.-T. Zhang, “An accurate and robust strip-edge-based structured light means for shiny surface micromeasurement in 3-D,” IEEE Trans. Ind. Electron., vol. 60, no. 3, pp. 1023–1032, Mar. 2013.

Z. Song, R. Benenson, and B. Schiele, “CityPersons: A diverse dataset for pedestrian detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3213–3221.

K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.

R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.

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

H. Choi, S. Kim, K. Park, and K. Sohn, “Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks,” in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), Dec. 2016, pp. 621–626.

J. Fredebon, “The role of instructions and familiar size in absolute judgments of size and distance,” Perception Psychophys., vol. 51, no. 4, pp. 344–354, Jul. 1992.

A. Kar, S. Tulsiani, J. Carreira, and J. Malik, “Amodal completion and size constancy in natural scenes,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 127–135.

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

Y. Cao, D. Guan, Y. Wu, J. Yang, Y. Cao, and M. Y. Yang, “Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection,” ISPRS J. Photogramm. Remote Sens., vol. 150, pp. 70–79, Apr. 2019.

C. Li, D. Song, R. Tong, and M. Tang, “Multispectral pedestrian detection via simultaneous detection and segmentation,” in Proc. Brit. Mach. Vis. Conf. (BMVC), Sep. 2018, pp. 1–12.

O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

S. Song and J. Xiao, “Deep sliding shapes for amodal 3D object detection in RGB-D images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 808–816.

T. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from RGB-D images,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2012, pp. 746–760.

S. Song, S. P. Lichtenberg, and J. Xiao, “Sun RGB-D: A RGB-D scene understanding benchmark suite,” in Proc. IEEE Conf. Comput. Vis. pattern Recognit., Jun. 2015, pp. 567–576.

Lu Zhang received the B.E. degree in automation from Shandong University, Jinan, China, in 2016, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2021. She is currently an Assistant Professor with Chinese Academy of Sciences. Her research interests include computer vision, pattern recognition, and robotics.

Zhiyong Liu (Senior Member, IEEE) received the B.E. degree in electronic engineering from Tianjin University, Tianjin, China, in 1997, the M.E. degree in control engineering from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2000, and the Ph.D. degree in computer science from The Chinese University of Hong Kong, Hong Kong, in 2003. He is currently a Professor with the Institute of Automation, CASIA. His current research interests include machine learning, pattern recognition, computer vision, and bioinformatics.

Xiangyu Zhu (Member, IEEE) received the B.E. degree from Sichuan University (SCU), Chengdu, China, in 2012, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Associate Professor with Chinese Academy of Sciences. His research interests include pattern recognition and computer vision, in particular, 3-D morphable model and face analysis.

Zhan Song received the Ph.D. degree in mechanical and automation engineering from The Chinese University of Hong Kong, Hong Kong, in 2008. He is currently a Professor with the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS), Beijing, China. His current research interests include structured light-based sensing, image processing, 3-D face recognition, and human–computer interaction.

Xu Yang received the B.E. degree in electronic engineering from the Ocean University of China, Qingdao, China, in 2009, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2014. He is currently an Associate Professor with CASIA. His current research interests include computer vision, graph algorithms, and robotics.

Zhen Lei (Senior Member, IEEE) received the B.S. degree in automation from the University of Science and Technology of China, Hefei, China, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010. He is currently a Professor with Chinese Academy of Sciences. He has published over 100 articles in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular.

Hong Qiao (Fellow, IEEE) received the B.S. and M.S. degrees in engineering from Xi’an Jiaotong University, Xi’an, China, in 1986 and 1989, respectively, the Ph.D. degree in robotics control from De Montfort University, Leicester, U.K., in 1995, and the M.Phil. degree from the University of Strathclyde, Glasgow, U.K., in 1997. She held teaching and research positions with universities in the U.K. and Hong Kong from 1990 to 2004. She is currently a “100-Talents Project” Professor with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her current research interests include robotic manipulation, robotic vision, bioinspired intelligent robot, and brain-like intelligence.