Multi-modal pedestrian detection with misalignment based on modal-wise regression and multi-modal IoU

Napat Wanchaitanawong,[a,*] **Masayuki Tanaka**,[a] **Takashi Shibata**,[b] **and Masatoshi Okutomi**[a]

[a]Tokyo Institute of Technology, Tokyo, Japan
[b]NTT Corporation, Atsugi, Japan

Abstract. Multi-modal pedestrian detection, which integrates visible and thermal sensors, has been developed to overcome many limitations of visible-modal pedestrian detection, such as poor illumination, cluttered background, and occlusion. By adopting the combination of multiple modalities, we can efficiently detect pedestrians even with poor visibility. Nevertheless, the critical assumption of multi-modal pedestrian detection is that multi-modal images are perfectly aligned. In general, however, this assumption often becomes invalid in real-world situations. Viewpoints of the different modal sensors are usually different. Then, the positions of pedestrians on the different modal images have disparities. We proposed a multi-modal faster-RCNN specifically designed to handle misalignment between two modalities. The faster-RCNN consists of a region proposal network (RPN) and a detector. We introduce position regressors for both modalities in the RPN and the detector. Intersection over union (IoU) is one of the useful metrics for object detection but is defined only for a single-modal image. We extend it into multi-modal IoU to evaluate the preciseness of both modalities. Our experimental results with the proposed evaluation metrics demonstrate that the proposed method has comparable performance with state-of-the-art methods and outperforms them for data with significant misalignment. [DOI: 10.1117/1.JEI.32.1.013025]

Keywords: deep-learning; convolution neural network; pedestrian detection; multi-modal object detection.

Paper 221007G received Sep. 22, 2022; accepted for publication Jan. 18, 2023; published online Feb. 7, 2023.

1 Introduction

Pedestrian detection is one of the active research topics in computer vision field, with several crucial applications, such as autonomous driving[1] and video surveillance systems.[2] Only visible images (e.g., RGB images) were used in this task. Pedestrian detections with only visible images have several issues. Detection accuracy is significantly degraded under poor lighting conditions.[3–7] To overcome those issues, various approaches have been proposed to combine multiple modalities (e.g., visible and far-infrared)[8,9] and utilize the highly apparent regions of these modalities together. Those methods simply combine features of both modalities directly, namely, typical two-stream faster region-based convolutional neural network (R-CNN).[10–13] The fundamental assumption for this two-stream approach is that alignment between two modalities is perfect. In general, however, this assumption often breaks down due to lack of time synchronization, inaccurate calibration, or the effects of disparity for stereo.[14,15] For instance, MSDS-RCNN,[10] which combines detection and semantic segmentation tasks to optimize the model; however, without any consideration about misalignment, is very sensitive to misalignment and can only precisely locate pedestrians in visible modality, as shown in Fig. 1(a), their detection bounding boxes are only for visible modality.

Recently, several methods have been proposed to address the misalignment problem. For example, aligned region-convolutional neural network (AR-CNN)[16] proposed incorporating

*Address all correspondence to Napat Wanchaitanawong, wnapat@ok.sc.e.titech.ac.jp

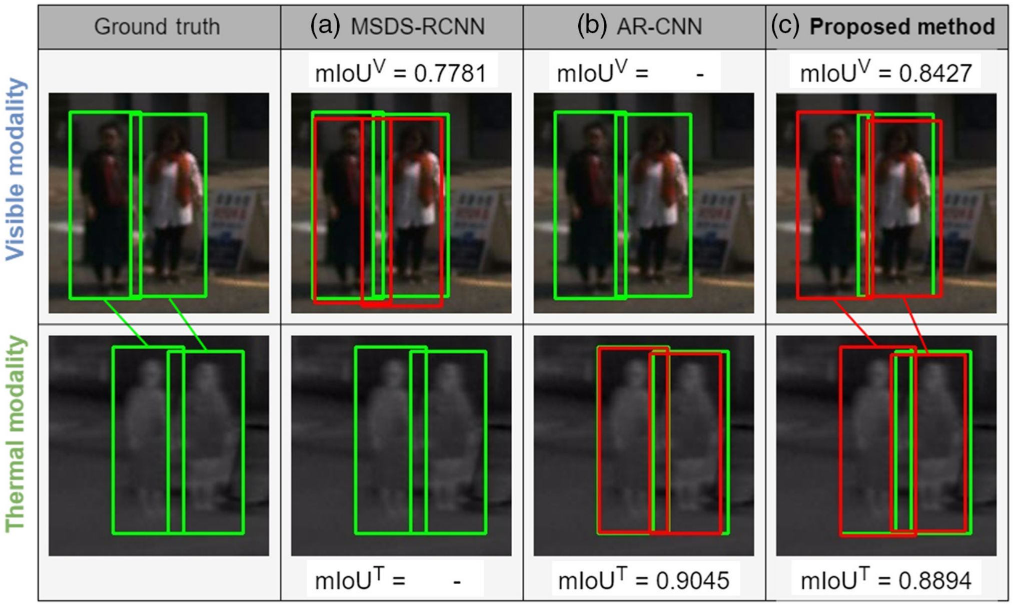
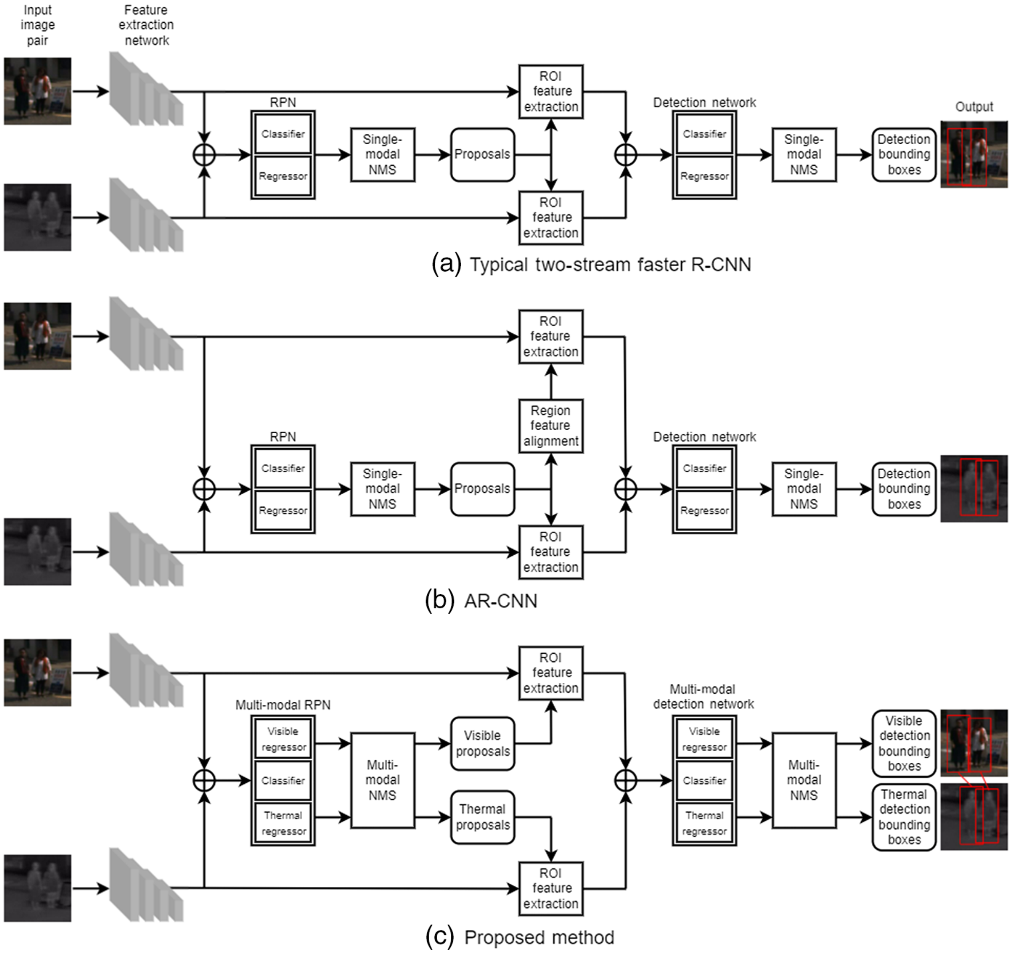


Fig. 1 Visualization examples of ground truth annotations by[16] (boxes in green), detection results (boxes in red), and overlap area between them, measured by mean visible IoU (mIoU^V) and mean thermal IoU (mIoU^T) of MSDS-RCNN,[10] AR-CNN,[16] and the proposed method. Image patches are cropped from visible-thermal image pairs in the same position from KAIST multispectral pedestrian detection dataset[8] with large misalignment. (a) MSDS-RCNN;[10] (b) AR-CNN;[16] and (c) proposed method.

an alignment module inside the faster-RCNN. MBNet[17] proposed an illumination-aware feature alignment module to adaptively align features between two modalities. These explicitly designed methods have improved performance and robustness against misalignment. Still, to the best of our knowledge, existing methods that consider misalignment only output one coordinate for each object to represent its position in both modalities, completely neglecting that each object can have different positions in different modalities due to misalignment. As shown in Fig. 1(b), despite the prediction of shift distances of objects between modalities, AR-CNN[16] only detects pedestrians in thermal modality (their implementation only produces bounding boxes for thermal modality). Even though detection bounding boxes of MSDS-RCNN[10] can accurately locate objects in visible modality, and the same goes for AR-CNN in thermal modality, those bounding boxes are far off from their corresponding objects in another modality due to the large disparity between them. For this reason, when a significant misalignment is present, the existing methods can not locate each object in both modalities, forcing us to utilize only their reliable modalities and ignore the others. In summary, multi-modal pedestrian detection with large misalignment still has ongoing problems; one of them is the ability to accurately locate objects for both modalities amid large misalignment.

To tackle the problems mentioned above, we propose a multi-modal faster-RCNN that is robust against misalignment. We use several novel strategies for the proposed multi-modal detection, including (1) modal-wise position regressor, (2) multi-modal mini-batch sampling, and (3) multi-modal non-maximum suppression (NMS). The proposed method detects each object as a pair of bounding boxes with different coordinates in each modality, as illustrated in Fig. 1(c). It is noteworthy that despite the differences in the position of detection bounding boxes between modalities, all bounding boxes have paired relations between modalities; each pair indicates the same object in both modalities. Consequently, the proposed method can accurately pinpoint all objects in both modalities and match them regardless of displacement caused by misalignment. Figure 2 shows the different faster R-CNN-based approaches to multi-modal pedestrian detections. As shown in Fig. 2(a), a typical two-stream faster R-CNN fuses features from both modalities directly without handling the disparity between each object. They can only output detection bounding boxes for either visible or thermal modality. Explicitly addressing the misalignment



(a) Typical two-stream faster R-CNN

(b) AR-CNN

(c) Proposed method

Fig. 2 Comparison of multi-modal pedestrian detection frameworks based on faster R-CNN. (a) Typical two-stream faster R-CNN; (b) AR-CNN;[16] and (c) proposed method.

problem, AR-CNN integrates region feature alignment to align each visible region with its counterpart thermal region before the detection network. Still, their method only outputs detection bounding boxes according to the position of objects in thermal modality alone, as shown in Fig. 2(b). Our proposed method, on the contrary, installed with a dual-regressor for both region proposal network (RPN) and detector and newly-introduced multi-modal NMS, can output pairs of bounding boxes, which accurately locate objects in both modalities, as shown in Fig. 2(c).

We also introduce new evaluation metrics to analyze the performance of multi-modal pedestrian detection networks against misalignment based on the precision of detection bounding boxes in both modalities, namely, multi-modal IoU ($\mathrm{IoU}^{\mathrm{M}}$) and multi-modal log-average miss rate (MR^{M}). Our experiments show that the proposed method's performance significantly outperforms state-of-the-art methods for large misalignment data.

Overall, the main contributions of this article are as follows: (1) we introduce a new problem for multi-modal pedestrian detection with large misalignment to precisely locate objects in both modalities and correctly match them. We also introduce new evaluation metrics, $\mathrm{IoU}^{\mathrm{M}}$ and MR^{M}, to evaluate the performance on these tasks; (2) we introduce new training strategies for our multimodal pedestrian detection network to deal with misalignment: modal-wise regression, multimodal mini-batch sampling, and multi-modal NMS. Our strategies are applicable to other multimodal detection tasks. (3) We experiment on KAIST[8] multispectral pedestrian dataset with our experimental setting and evaluation metric to productively analyze the performance of multimodal detectors against misalignment. Our proposed method achieves comparable performance with state-of-the-art methods and outperforms them when misalignment is large.

This article is an extended version of our previous conference paper,[18] mainly in the following points: (1) we improve the performance of our method with a new implementation for NMS and parameters fine-tuning; (2) additional explanation and analysis of misalignment problem and proposed model are portrayed in detail; and (3) more experiments are conducted to demonstrate the advantage of our methods compared to state-of-the-art (SOTA) methods and ablation study to carefully inspect the impact of each component of the proposed model.

2 Related Work

In this section, related works are reviewed. First, we review single-modal pedestrian detection, i.e., pedestrian detection that only uses visible images. Second, we review multi-modal pedestrian detection, i.e., pedestrian detection that uses both visible and thermal images, dividing into naive feature fusion and adaptive feature fusion.

2.1 *Single-Modal Pedestrian Detection*

Pedestrian detection has improved dramatically since the traditional hand-crafted features-based methods, such as histogram of oriented gradient[19] and integral channel features (ICF),[20] which were made obsolete by superior deep-learning-based methods.[3–7] The most notable pedestrian dataset of visible images is the Caltech Pedestrian Dataset.[21] However, limited to only information from the visible channel, many pedestrians are still very difficult to recognize, even with human perception. The detection performance was degraded by many challenges,[4] such as low image resolution, occlusion, adverse illumination, cluttered background, and inconsistent pattern of humans.

2.2 *Multi-Modal Pedestrian Detection*

2.2.1 *Naive feature fusion*

KAIST multispectral pedestrian detection (KAIST) dataset[8] has been widely used in the research field of multi-modal pedestrian detection, making it progress steadily. Despite non-CNN-based approaches such as aggregate channel features[22] in the early days, the CNN-based approach is mainstream in this field currently.[10–13,23–29] The main challenge in the early days was how to combine and make use of information from both modalities.[30–33] Most existing methods assume that visible-thermal image pairs are geometrically aligned. Those methods fuse both modalities' features directly in corresponding pixel positions, as shown in Fig. 2(a). Although many geometric calibration and image alignment methods for multi-modal cameras have been proposed,[34–37] accurate and dense alignment for each pixel is still an open problem. As a result, their detectors suffer dramatically worse performance in poorly ARs.

2.2.2 *Adaptive feature fusion*

AR-CNN[16] is the first work that immensely tackles the misalignment issue in multi-modal CNN-based pedestrian detection. They proposed AR-CNN, considering the disparity between multi-modalities. They also provided KAIST paired annotation, which includes annotated bounding boxes for each modality. Their method predicts the shift distance between modalities for each region of interest (RoI), relocates the visible region into the thermal area, and then aligns them together, as shown in Fig. 2(b). They successfully improved performance from previous methods that do not consider misalignment, revealing the influence of misalignment. MBNet[17] also proposes a method that takes modality imbalance into account. However, those methods assume that the misalignment is weak, which leads to inaccurate detection of bounding boxes in one (or both) modality when the misalignment is significant. To tackle this problem, we introduce the modal-wise regressor to detect each object in a pair of bounding boxes with different coordinates in each modality, as shown in Fig. 2(c), resulting in more accurate object localization in both modalities.

2.3 *Evaluation Metric for Pedestrian Detection*

In object detection, there are several evaluation metrics. The most fundamental metric is intersection over union (IoU), which measures the overlap between two bounding boxes. We predefine an IoU threshold (usually 0.5) between the predicted bounding box and ground truth bounding box to classify whether that bounding box is true positive or false positive or between each bounding box to discard the low score one with NMS. One of the most common metrics to measure object detection accuracy is average precision (AP). For PASCAL VOC dataset,[38] AP is the average precision for recall over 0 to 1 with IoU threshold of 0.5, which then average over all object categories. Meanwhile, for COCO dataset,[39] mAP (or just AP) is the average AP for IoU threshold over 0.5 to 0.95 with a step size of 0.05 to measure the precision of detection with varying restrictions, not just IoU threshold of 0.5. Despite that, MR has been used as the primary evaluation metric for pedestrian detection since we only focus on one class in pedestrian detection, pedestrian. Moreover, pedestrian detection is closely related to real-life applications such as autonomous driving cars, any false negative in detection could cause a severe accident. At first, the Caltech Pedestrian Dataset[21] plotted MR over false positives per image (FPPI) in log-log scale and used MR at 10^0 FPPI as a common reference point to compare performances. KAIST dataset,[8] however, stepped up to log-average MR over 10^{-2} to 10^0 (MR) following the suggestion by Dollar et al.[40] Since then, MR has been the primary evaluation metric for multi-modal pedestrian detection.

3 Proposed Method

This section explains our proposed method and evaluation metrics in detail. First, we describe each evaluation metric we propose for multi-modal pedestrian detection. Second, we describe our proposed multi-modal faster R-CNN network designed for misalignment problems.

3.1 *Proposed Evaluation Metrics*

We propose evaluation metrics that we use in our training and performance testing. First, multi-modal IoU ($\mathrm{IoU}^{\mathrm{M}}$) is introduced. Second, multi-modal MR (MR^{M}) is introduced.

3.1.1 *Multi-modal IoU*

Traditionally, we use IoU in object detection tasks to evaluate the overlap between ground truth and detection bounding boxes. The IoU is defined as

$$\mathrm{IoU} = \frac{\mathrm{GT} \cap \mathrm{DT}}{\mathrm{GT} \cup \mathrm{DT}}, \tag{1}$$

where GT and DT denote ground truth and detection bounding boxes, respectively. $\mathrm{GT} \cap \mathrm{DT}$ represents the area of intersection of ground truth and detection bounding boxes, $\mathrm{GT} \cup \mathrm{DT}$ represents the area of union of ground truth and detection bounding boxes. However, when there is a misalignment between modalities, the coordinates of each object in both modalities are not the same. If we are only concerned about the precision of one modality, another modality will have poor precision. To measure the ability to handle both modalities, especially when the level of misalignment is high, we introduce a new evaluation metric, which we call "multi-modal IoU ($\mathrm{IoU}^{\mathrm{M}}$)" defined as

$$\mathrm{IoU}^{\mathrm{M}} = \frac{(\mathrm{GT}^V \cap \mathrm{DT}^V) + (\mathrm{GT}^T \cap \mathrm{DT}^T)}{(\mathrm{GT}^V \cup \mathrm{DT}^V) + (\mathrm{GT}^T \cup \mathrm{DT}^T)}, \tag{2}$$

where GT^V and GT^T denote paired ground-truth bounding boxes referring to the same object of visible and thermal modalities, respectively. DT^V and DT^T denote paired detection bounding boxes referring to the same object from visible and thermal modality, respectively. $\mathrm{IoU}^{\mathrm{M}}$ can be used to determine the precision of detection bounding boxes in both modalities.

3.1.2 *Multi-modal MR*

Following the traditional evaluation of object detection, we categorize detection bounding boxes and ground-truth bounding boxes into true positives, false positives, and false negatives to evaluate detection results. The traditional way to do that is the greedy matching algorithm. Matched detection bounding boxes will become true positives, unmatched detection bounding boxes will become false positives, and unmatched ground truth bounding boxes will become false negatives. In pedestrian detection, we value false negatives the most since miss detection could be crucial in real-life applications. The lower the false negatives, the better. In multi-modal pedestrian detection, performance is traditionally measured by log-average MR suggested by Dollar et al.[40] MR is defined by geometrical mean of MRs at specific FPPI evenly divided in log space, which can be formulated as

$$\text{Log} - \text{Average Miss Rate (MR)} = \left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} = \exp\left[\frac{1}{n}\sum_{i=1}^{n} \ln a_i\right], \tag{3}$$

where $a_1, a_2, \ldots, a_n$ are MRs at n different FPPI evenly spaced in log space. MR is the proportion of false negative results to total objects, and FPPI is the proportion of false positive results to total images. Traditionally, we use nine MRs at evenly spaced FPPI over $[10^{-2}, 10^{0}]$ in log space $(10^{-2}, 10^{-1.75}, 10^{-1.5}, \ldots, 10^{0})$ to calculate MR, at which we call MR^{-2}. The lower the MR, the better.

The original KAIST dataset only had a single common annotation for each object in both modalities, despite misalignment between them. Their annotation also has many errors, such as imprecise localization, misclassification, and misARs.[10] Aware of the issue, many researchers relabeled KAIST annotation to solve the above errors. Liu et al.[11] provided improved annotation for the testing, which has become the standard annotation for performance evaluation. Li et al.[10] provided sanitized annotation for the training and demonstrated the effects caused by different kinds of annotation errors. Zhang et al.[16,41] provided revolutionary KAIST-paired annotation, which carefully localizes pedestrians in both modalities and builds their relationships. They also evaluated the detection performance by MR^{V} and MR^{T}, which denote MR evaluating by visible annotation and thermal annotation, respectively. However, those evaluations were performed separately, and their detection results have no relationship between visible and thermal bounding boxes, which makes To evaluate the precision of detection results in both modalities pairwise, we change the criteria of the greedy matching algorithm from IoU to IoU^{M}, which represents MR based on IoU^{M}, multi-modal MR (MR^{M}). To use this metric, the detection results must be pairs of bounding boxes; each pair locates the same object in both modalities, which could have different coordinates due to misalignment. Not only is MR^{M} able to measure the precision of bounding boxes in both modalities simultaneously, but it also measures the ability to correctly match objects between modalities with misalignment since the detection bounding box pair can mismatch with other nearby objects, which can potentially become false negative, resulting in lower MR^{M}. We experiment using MR^{M} as an evaluation metric to demonstrate its effectiveness in measuring the detection performance against misalignment.

3.2 *Proposed Model*

We adopt faster R-CNN[42] architecture and extend it into a two-stream network for multi-modal imaging, which consists of multi-modal RPN, multi-modal detector, and multi-modal NMS. Moreover, our multi-modal mini-batch sampling strategy is introduced. An overview of our network structure is shown in Fig. 3.

3.2.1 *Multi-modal RPN*

The proposed multi-modal RPN has regressors for both modalities, enabling proposals from each modality to adjust their sizes and positions independently. After receiving channel-wise concatenated features from backbone networks, the proposed multi-modal RPN will generate proposal pairs as its output via classifier and dual-regressor, predicting each proposal pair's

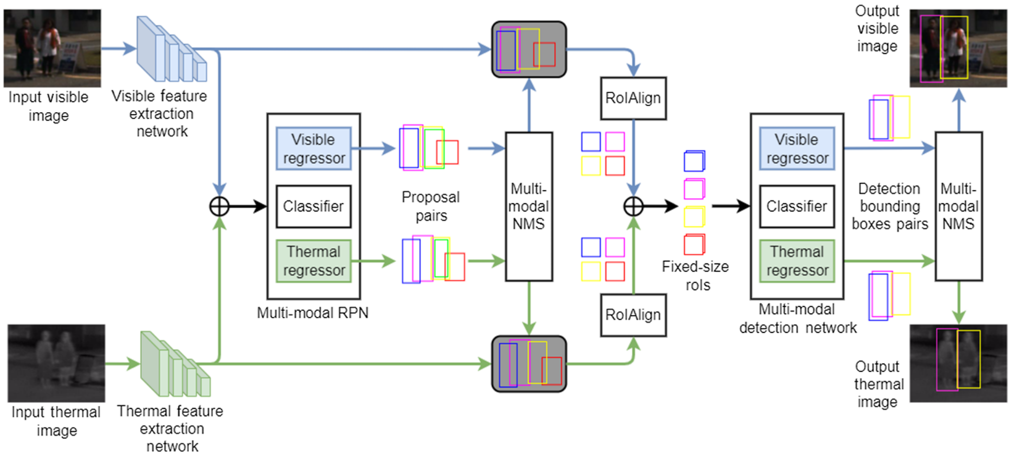


Fig. 3 The overall architecture of our network. We extend Faster R-CNN into a two-stream network to take visible-thermal image pairs as input, then return pairs of detection bounding boxes as output for both modalities. Blue and green blocks/paths represent properties of visible and thermal modalities, respectively. RoIs and bounding boxes with the same color represent their paired relations. $\oplus$ denotes channel-wise concatenation.

confidence score and regressing each proposal individually. We use multi-modal NMS (Sec. 3.2.3) to filter the best 300 out of many redundant proposal pairs to keep paired relations of proposals. All remaining proposals will be applied with RoIAlign[43] operation to extract their feature maps into the exact size of 7×7 before returning to channel-wise concatenate with their corresponding pairs, resulting in well-aligned RoI for the detector. While single-modal regressor returns proposal pair with the same position for both bounding boxes, multi-modal regressor returns proposal pair with different positions for both bounding boxes, which gives more accurate RoI for detector in case of significant misalignment. We adopt the loss function of RPN from faster R-CNN[42] and add one more regression loss to optimize the precision of both modalities, which is defined as

$$L(\{p_i\},\{\mathbf{t}_i^V\},\{\mathbf{t}_i^T\},\{p_i^*\},\{\mathbf{t}_i^{V*}\},\{\mathbf{t}_i^{T*}\}) = \frac{1}{N_{\mathrm{cls}}}\sum_i L_{\mathrm{cls}}(p_i,p_i^*) + \frac{\lambda}{N_{\mathrm{reg}}}\sum_i p_i^*[L_{\mathrm{reg}}^V(\mathbf{t}_i^V,\mathbf{t}_i^{V*}) + L_{\mathrm{reg}}^T(\mathbf{t}_i^T,\mathbf{t}_i^{T*})], \tag{4}$$

where i is the index of the anchor, p_i is the predicted probability of anchor i being an object. p_i^* is ground truth label of anchor i, which equals 1 if anchor i is positive (overlaps with an object above the high threshold) and 0 if anchor i is negative (overlaps with an object below the low threshold). L_{cls} is a cross-entropy over object and not object classes, which is defined as

$$L_{\mathrm{cls}}(p,p^*) = -(p^* \log(p) + (1-p^*)\log(1-p)). \tag{5}$$

$\mathbf{t}_i^V$ and $\mathbf{t}_i^T$ are vectors representing parameterized coordinates of predicted bounding box pairs as $\mathbf{t} = (t_x, t_y, t_w, t_h)$ that associate with anchor i in visible and thermal modalities, respectively, and $\mathbf{t}_i^{V*}$, $\mathbf{t}_i^{T*}$ are that of the ground truth bounding box pairs. Regression losses L_{reg}^V, L_{reg}^T are smooth L_1 loss for visible and thermal modalities, respectively, which are only activated for positive anchors ($p_i^* = 1$), defined as

$$L_{\mathrm{reg}}(\mathbf{t},\mathbf{t}^*) = \sum_{j\in\{x,y,w,h\}} \mathrm{smooth}_{L1}(t_j - t_j^*), \tag{6}$$

in which

$$\mathrm{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}, \tag{7}$$

N_{cls} is mini-batch size and N_{reg} is number of anchor locations. Following Girshick et al.,[44] the parameterized coordinates $\mathbf{t} = (t_x, t_y, t_w, t_h)$ for regression are scale-invariant translation and log-space shift relative, defined as

$$\begin{aligned} t_x &= (x_p - x_a)/w_a, \quad t_x^* = (x^* - x_a)/w_a, \\ t_y &= (y_p - y_a)/h_a, \quad t_y^* = (y^* - y_a)/h_a, \\ t_w &= \log(w_p/w_a), \quad t_w^* = \log(w^*/w_a), \\ t_h &= \log(h_p/h_a), \quad t_h^* = \log(h^*/h_a), \end{aligned} \tag{8}$$

where x and y denote the bounding box's center coordinates and w and h denote its width and height. Variables p, a, and * denote coordinates of prediction, anchor, and ground truth bounding boxes, respectively. We set $\lambda = 1$ for all experiments.

3.2.2 *Multi-modal detector*

Similar to RPN, the proposed multi-modal detector network has regressors for both modalities to independently adjust bounding boxes' positions and one classifier to predict each bounding box pair's confidence score. Multi-modal NMS (Sec. 3.2.3) is also applied to eliminate vague overlapping bounding box pairs. We will have detection result as pairs of bounding boxes for both modalities, which have different sizes and positions in different modalities, resulting in detection bounding boxes that are precise for both modalities and keep their paired relations. We adopt the loss function of the detector from fast R-CNN[45] and add one more regression loss, which is defined as

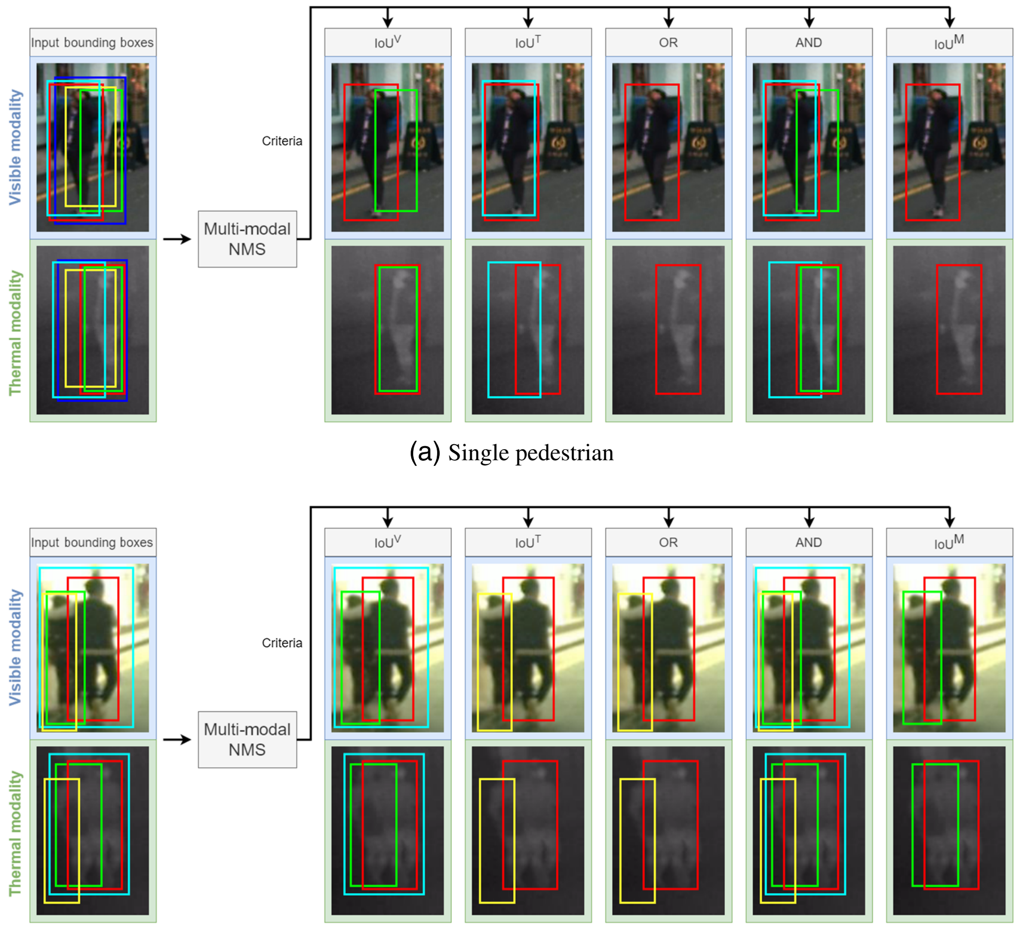
$$\begin{aligned} L(\{p_i\}, \{\mathbf{t}_i^V\}, \{\mathbf{t}_i^T\}, \{p_i^*\}, \{\mathbf{t}_i^{V*}\}, \{\mathbf{t}_i^{T*}\}) &= \sum_i L_{cls}(p_i, p_i^*) \\ &+ \lambda \sum_i p_i^* [L_{reg}^V(\mathbf{t}_i^V, \mathbf{t}_i^{V*}) + L_{reg}^T(\mathbf{t}_i^T, \mathbf{t}_i^{T*})], \end{aligned} \tag{9}$$

where i is the index of a bounding box pair, L_{cls} is a cross-entropy for class probability p_i and true class p_i^* [Eq. (5)] of bounding box pair i, since we only consider pedestrian class, there are only two classes, pedestrian and non-pedestrian, in which p_i^* equals 1 and 0, respectively. Regression losses L_{reg}^V, L_{reg}^T are smooth L_1 loss [Eqs. (6) and (7)] over predicted regression offsets $\mathbf{t}_i^V$, $\mathbf{t}_i^T$ and regression targets $\mathbf{t}_i^{V*}$, $\mathbf{t}_i^{T*}$ of bounding box pair i for visible and thermal modalities, respectively, which are also parameterized as Eq. (8). We set $\lambda = 1$ for all experiments.

3.2.3 *Multi-modal NMS*

NMS is a technique for selecting one entity out of many overlapping entities, usually using IoU as a suppression criterion, i.e., when the IoU between bounding boxes exceeds the threshold, the bounding box with a lower prediction score is suppressed. Since CNN-based methods generate many dense bounding boxes mostly detecting the same objects, the detection results are cluttered with unnecessary bounding boxes. Therefore, we use NMS to remove those lower-quality bounding boxes, keeping only the best bounding boxes to locate the objects. However, since we need to keep paired relations of bounding boxes between visible and thermal modalities in this procedure, we must select and suppress bounding boxes in a pairwise approach, or paired relations would be lost in the suppression process. For this purpose, we attempt various criteria that can select and suppress bounding boxes in a pairwise manner for our NMS.

As a naive extension, we can use either $\mathrm{IoU}^{\mathrm{V}}$ or $\mathrm{IoU}^{\mathrm{T}}$ as multi-modal NMS criteria, i.e., making one modality a dictator. In AR-CNN,[16] $\mathrm{IoU}^{\mathrm{T}}$ is used for NMS criteria, neglecting information of visible modality. Accordingly, we can use the logical operation of $\mathrm{IoU}^{\mathrm{V}}$ and $\mathrm{IoU}^{\mathrm{T}}$ as



(a) Single pedestrian

(b) Two pedestrians with partial occlusion

Fig. 4 Examples of multi-modal NMS processes with different criteria on the scenes where large misalignment is present. Bounding boxes with the same color reflect paired relations between them. Left side shows bounding box pairs prior to multi-modal NMS. Right side shows results of multi-modal NMS with different criteria. (a) Single pedestrian and (b) two pedestrians with partial occlusion.

criteria. For OR operation, if either IoU^V or IoU^T exceeds the threshold, the bounding box pair with a lower score will be suppressed. For AND operation, if both IoU^V and IoU^T exceed the threshold, the bounding box pair with a lower score will be suppressed. Lastly, we use the proposed IoU^M as criteria, where the proportion of intersection and union from both modalities is considered. Examples of proposed multi-modal NMS results with different criteria are shown in Fig. 4. Several detection bounding box pairs around objects are shown before and after the multimodal NMS process.

From Fig. 4(a), IoU^V and IoU^T clearly show weakness, considering only one modality, these criteria can not get rid of all ambiguous bounding boxes around the same object, which is also the same for logical operator AND. On the contrary, logical operator OR and IoU^M can suppress needless bounding boxes correctly. In case of multiple pedestrians with partial occlusion and misalignment, which is not uncommon in the real situation, correct bounding boxes could be removed if not handled properly. As shown in Fig. 4(b), IoU^V and logical operator AND, while preserving correct bounding boxes, fail to remove poor bounding boxes around the objects. IoU^t and logical operator OR remove most poor bounding boxes, including correct green bounding boxes. As a result, less precise yellow bounding boxes remain. Meanwhile, IoU^M can suppress and keep all correct bounding boxes precisely. IoU^M is our best candidate for NMS criteria since it considers the overlaps between bounding boxes in both modalities, unlike other criteria.

The experiment to demonstrate performance comparison between all NMS criteria is also conducted (Sec. 4.5.2).

3.2.4 *Multi-modal mini-batch sampling*

We follow sampling strategies from Faster R-CNN[42] and Fast R-CNN.[45] Although, the dual-regressor of our method requires that we select training samples as anchor pairs and RoI pairs for RPN and detector, respectively. Same as NMS, we can use $\mathrm{IoU}^{\mathrm{V}}$, $\mathrm{IoU}^{\mathrm{T}}$, logical operation of $\mathrm{IoU}^{\mathrm{V}}$ and $\mathrm{IoU}^{\mathrm{T}}$, or $\mathrm{IoU}^{\mathrm{M}}$ as criteria. For this purpose, we use $\mathrm{IoU}^{\mathrm{M}}$ as selection criteria instead of traditional IoU to consider the location of objects in both modalities. For RPN, we assign positive labels to anchors that overlap with any ground truth bounding box pair higher than the high $\mathrm{IoU}^{\mathrm{M}}$ threshold and assign negative labels to anchors that overlap with all ground truth bounding box pairs lower than the low $\mathrm{IoU}^{\mathrm{M}}$ threshold, for a total of 256 anchor pairs, whereas positive labels can take up to 128 anchor pairs. For detector, we assign positive labels to RoI pairs that overlap with any ground truth bounding box pair higher than the high $\mathrm{IoU}^{\mathrm{M}}$ threshold and assign negative labels to RoI pairs that overlap with all ground truth bounding box pairs lower than the high $\mathrm{IoU}^{\mathrm{M}}$ threshold but higher than the low $\mathrm{IoU}^{\mathrm{M}}$ threshold, for a total of 128 RoI pairs, whereas positive labels can take up to 32 RoI pairs.

4 Experiments

First, we describe the dataset we used in our experiments. Second, the details of our implementation are clarified. Third, we indicate evaluation details of our experiments, which include the explanation of simulated disparity of misalignment experiment. Fourth, we illustrate and discuss the results of our experiments compared with existing methods, divided into performance comparison and qualitative comparison. Finally, we conducted ablation experiments to verify the effectiveness of our network's components, multi-modal regressor, and multi-modal NMS.

4.1 *Dataset*

KAIST dataset[8] was used in our experiments. It is one of the widely used multi-modal pedestrian datasets, with more than 90,000 frames recorded both day and night to consider changes in light conditions. It was initially assumed to be geometrically aligned. However, the annotations have many errors,[10] such as imprecise localization, misclassification, and misARs. Many researchers constructed their improved version of annotations to use instead of the original. Improved annotations provided by Liu et al.[11] has officially been used as standard annotations for performance benchmark. Zhang et al.[16] carefully analyzed the misalignment problem of KAIST dataset and were the first to provide paired annotations for KAIST dataset, locating objects for each modality individually and building their paired relations. Since we focus on the misalignment problem, we adopted their annotations to use in our work for training and testing.

4.2 *Implementation Details*

We adopt VGG-16[46] pre-trained on ImageNet[47,48] as our two-stream backbone networks as in AR-CNN.[16] We train the network for three epochs with a learning rate of 0.005 and one additional epoch with a learning rate of 0.0005 by stochastic gradient descent optimizer with 0.9 momentum and 0.0005 weight decay. We select 8892 images from the training set containing informative pedestrians for the training. Image resolution is fixed to 640×512. All images are horizontally flipped and append to original training data for data augmentation. Since we utilize $\mathrm{IoU}^{\mathrm{M}}$ as batch sampling and NMS thresholds instead of traditional IoU, we conduct fine-tune experiments to get the suitable values for those thresholds. For multi-modal RPN's mini-batch sampling, we set $\mathrm{IoU}^{\mathrm{M}}$ of high and low thresholds at 0.63 and 0.3, respectively. For multi-modal detector's mini-batch sampling, we set $\mathrm{IoU}^{\mathrm{M}}$ of high threshold and low thresholds at 0.5 and 0.1, respectively. For the first NMS following RPN, we set $\mathrm{IoU}^{\mathrm{M}}$ threshold at 0.7 in the training to generate proposals with more variation in precision, which can benefit the training of the detector, and we set $\mathrm{IoU}^{\mathrm{M}}$ threshold at 0.55 in the testing to generate bounding boxes with higher precision. For the second NMS following detector, we set $\mathrm{IoU}^{\mathrm{M}}$ threshold at 0.53.

4.3 *Evaluation Details*

To thoroughly evaluate the effectiveness of our method against misalignment, we introduce simulated disparity of misalignment between modalities as an experiment set up by shifting thermal images by 2, 4, 6, 8, and 10 pixels horizontally in both directions to imitate the misalignment, which mainly occurs in the horizontal direction. There is no change to visible images. However, in case of any pedestrian goes over the image border as a result of shifting, That pedestrian will be ignored from the evaluation. Subsequently, we will have 11 subsets of different misalignments as test data for each horizontal shift. Mean and standard deviation (SD) of MR^{M} over all subsets are also calculated to evaluate the overall performance over different disparities and robustness to misalignment. For performance comparison, detection performance was measured by MR^{M} over the range of $[10^{-2}, 10^{0}]$ FPPI with $\mathrm{IoU}^{\mathrm{M}}$ threshold of 0.5 (MR^{M}_{50}) and 0.75 (MR^{M}_{75}), respectively, for all simulated disparity distances. Additionally, MR curves are plotted by using the mean and the worst MR of all simulated disparity distances over the range of $[10^{-2}, 10^{0}]$ FPPI with $\mathrm{IoU}^{\mathrm{M}}$ threshold of 0.5 (MR^{M}_{50}) and 0.75 (MR^{M}_{75}), respectively. Moreover, to see the performance of each modality independently, traditional MR of visible (MR^{V}) and thermal (MR^{T}) modalities are also evaluated for all simulated disparity distances, using $\mathrm{IoU}^{\mathrm{V}}$ for visible and $\mathrm{IoU}^{\mathrm{T}}$ for thermal, respectively.

For qualitative comparison, detection performance was measured by mean multi-modal IoU ($\mathrm{mIoU}^{\mathrm{M}}$) between all ground truth bounding boxes and detection bounding boxes with the highest $\mathrm{mIoU}^{\mathrm{M}}$ overlap in each scene. For ablation study, detection performance was measured by MR^{M} over the range of $[10^{-2}, 10^{0}]$ FPPI with $\mathrm{IoU}^{\mathrm{M}}$ threshold of 0.5 (MR^{M}_{50}) for all simulated disparity distances. All experiments were performed under reasonable configuration,[8] i.e., only pedestrians taller than 55 pixels under partial or no occlusion are considered. Only 2252 frames sampled from the test set with 20-frame skips were used in the performance test as traditional.

4.4 *Comparison with Existing Methods*

We selected three existing methods for our experiments, MSDS-RCNN[10] is representative of methods without misalignment consideration, AR-CNN[16] and MBNet[17] are methods that consider misalignment, trained by KAIST-paired annotations provided by Zhang et al.[16] For methods that do not have paired detection bounding boxes as their outputs, we substituted paired detection bounding boxes with their detection bounding boxes from one modality.

4.4.1 *Performance comparison*

As given in Table 1, MSDS-RCNN, the only method not considering misalignment, has much poorer performance than other methods, especially when the simulated disparity is significant. As for the proposed method, we have mediocre performance when there is no simulated disparity, and so as shift distance of −2, doing worse than MBNet[17] by about 8%. It is noteworthy that AR-CNN[16] and MBNet[17] have better performance at shift distance of −2 than without simulated disparity. This demonstrates that the dataset has some misalignment from the beginning, and simulated disparity could align objects in certain circumstances. Our method, however, has a noticeable performance improvement, achieving the best MR^{M}_{50} when disparities are larger than 4 pixels, demonstrating the effectiveness against misalignment of our method. Our method also achieved the best performance at all disparities when $\mathrm{IoU}^{\mathrm{M}}$ threshold is 0.75, as shown in Table 2, i.e., our method performs the best when the requirement of bounding boxes' precision is strict, indicating the superior precision of the proposed method's detection bounding boxes. We also have the lowest mean and SD for both MR^{M}_{50} and MR^{M}_{75}, demonstrating our best overall performance over all misalignment situations and most robust against misalignment.

The MR plots by mean MR (solid lines) and worse miss (dashed lines) rate over all simulated disparity distances is shown in Fig. 5. We can see MR plot of proposed method's mean MR at the

Table 1 Comparison with state-of-the-art methods on KAIST dataset, with simulated disparity of misalignment in the horizontal direction by MR_{50}^{M}, including their mean and SD over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Bold values indicate the best performance.

	Thermal images' horizontal shift distance (px)												
Methods	−10	−8	−6	−4	−2	0	2	4	6	8	10	Mean	SD
MSDS-RCNN[10]	27.06	18.76	15.93	12.74	12.58	11.09	11.72	13.25	15.06	21.38	27.48	17.00	5.94
AR-CNN[16]	21.61	14.65	10.43	8.67	8.22	8.79	8.68	10.10	11.02	14.65	19.84	12.42	4.69
MBNet[17]	23.14	15.31	11.02	8.92	**7.70**	**7.76**	8.64	9.88	11.17	14.87	21.70	12.74	5.43
Ours	**15.46**	**11.60**	**10.21**	**8.51**	8.43	8.28	**8.50**	**9.14**	**10.31**	**12.51**	**15.87**	**10.80**	**2.77**

Table 2 Comparison with state-of-the-art methods on KAIST dataset, with simulated disparity of misalignment in the horizontal direction by MR_{75}^{M}, including their mean and SD over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Bold values indicate the best performance.

	Thermal images' horizontal shift distance (px)												
Methods	−10	−8	−6	−4	−2	0	2	4	6	8	10	Mean	SD
MSDS-RCNN[10]	93.05	89.46	83.55	76.78	71.70	70.10	70.97	77.29	84.71	91.35	95.46	82.22	9.35
AR-CNN[16]	94.25	91.64	87.64	78.70	69.57	61.77	65.13	71.56	81.81	90.29	94.79	80.65	12.06
MBNet[17]	90.89	87.50	80.33	70.81	63.63	58.82	63.15	71.33	80.75	89.27	93.87	77.30	12.39
Ours	**63.30**	**59.94**	**56.67**	**55.87**	**55.45**	**55.07**	**55.39**	**56.35**	**57.01**	**59.72**	**62.58**	**57.94**	**2.96**

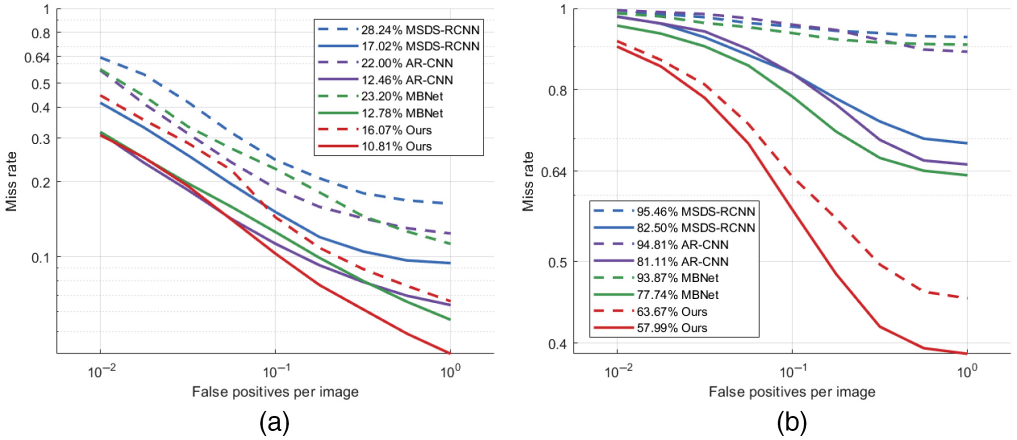


Fig. 5 Comparison of state-of-the-art methods' performance on KAIST dataset by mean MR (solid lines) and worst MR (dashed lines) of eleven different simulated disparities to FPPI curves. Numbers in legend show geometric mean of mean MR and geometric mean of worst MR over FPPI in the range of $[10^{-2}, 10^{0}]$ of each method. (a) $\mathrm{IoU}^{\mathrm{M}}$ threshold of 0.5 and (b) $\mathrm{IoU}^{\mathrm{M}}$ threshold of 0.75.

Table 3 Comparison with state-of-the-art methods on KAIST dataset, with simulated disparity of misalignment in the horizontal direction by MR^V_{50} and MR^T_{50}, including their mean and SD over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Bold values indicate the best performance.

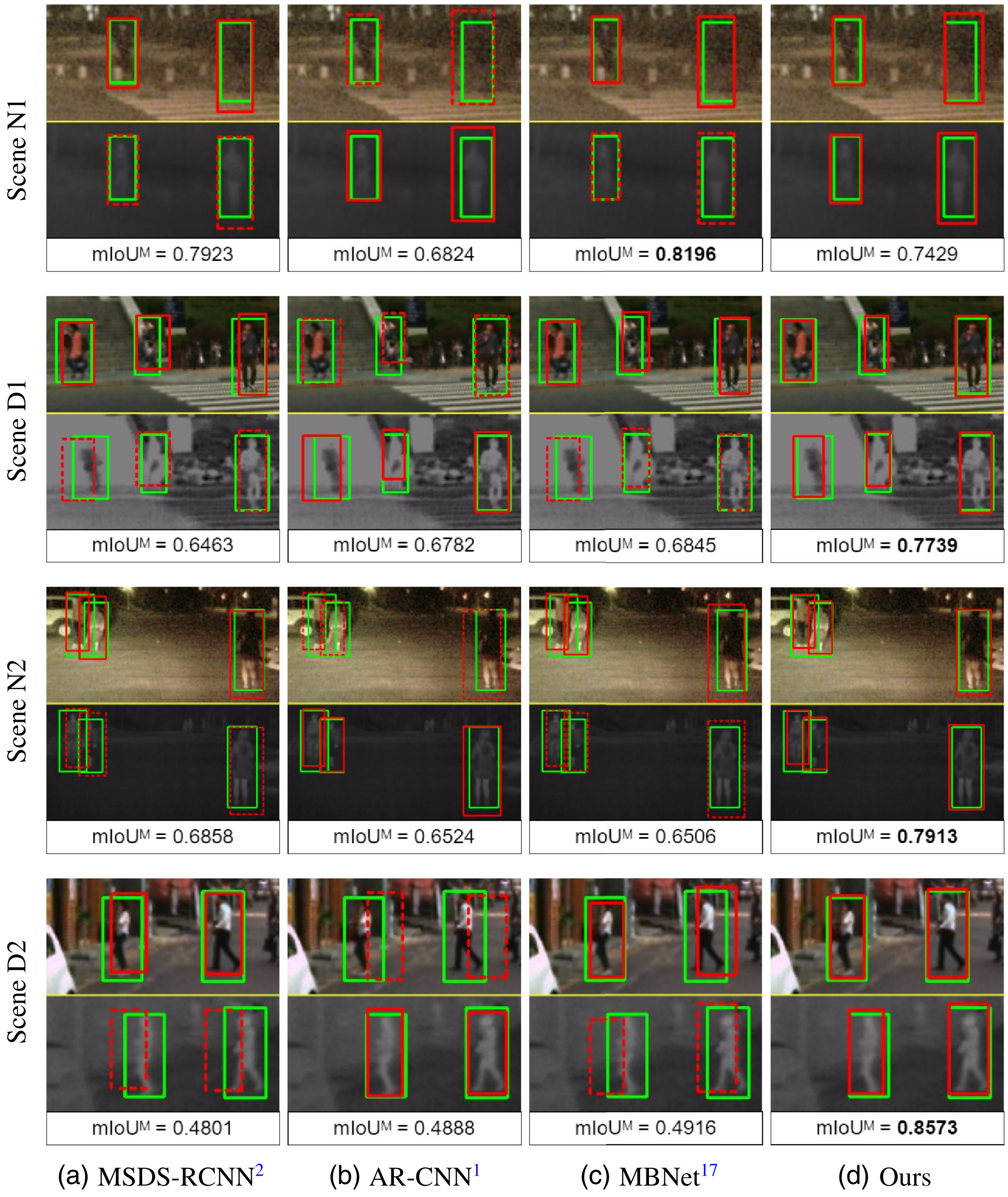
		Thermal images' horizontal shift distance (px)												
Methods		−10	−8	−6	−4	−2	0	2	4	6	8	10	Mean	SD
MSDS-RCNN[10]	MR^V_{50}	30.43	21.90	16.91	12.98	12.24	11.28	12.36	14.30	17.92	24.73	33.16	18.93	7.65
	MR^T_{50}	38.89	29.52	22.02	15.92	14.35	12.51	13.92	16.28	20.17	27.71	34.88	22.38	9.09
AR-CNN[16]	MR^V_{50}	70.83	57.92	37.70	19.14	11.40	9.12	13.41	17.54	27.80	43.26	61.39	33.59	22.05
	MR^T_{50}	**11.23**	**9.64**	**8.82**	**8.09**	**8.04**	9.05	**8.16**	**8.22**	**8.97**	**10.13**	**10.23**	**9.14**	**1.05**
MBNet[17]	MR^V_{50}	38.56	26.42	16.25	**10.27**	**8.82**	**7.89**	**8.68**	10.38	13.81	19.19	29.05	17.21	10.11
	MR^T_{50}	26.06	18.82	13.14	10.04	8.99	**8.12**	9.89	12.35	16.05	22.84	28.60	15.9	7.21
Ours	MR^V_{50}	**19.33**	**16.43**	**12.86**	10.83	10.10	9.32	9.63	**10.09**	**11.87**	**13.69**	**17.99**	**12.92**	**3.53**
	MR^T_{50}	12.92	11.31	10.41	8.83	8.45	8.55	8.46	9.40	9.85	11.41	12.87	10.22	1.69

bottom, with MR plot of the worst case being comparable to MR plots of other SOTA methods from Fig. 5(a). Our method also achieved the best MR^M_{50}, which is calculated by geometric mean of MR and over $[10^{-2}, 10^0]$ FPPI, for both mean MR and worst MR, lower than AR-CNN[16] by 13% and 27%, respectively. From Fig. 5(b), our proposed method clearly outshines other SOTA methods, separating at the bottom for both mean and worst case plots, demonstrating the superior robustness to misalignment and bounding box precision in both modalities. Same as MR^M_{50}, our method achieved the best MR^M_{75} for both mean MR and worst MR, better than MBNet[17] by 25% and 32%, respectively.

The performances measured by MR^V_{50} and MR^T_{50} are given in Table 3. Our proposed method significantly outperforms others on MR^V_{50} when the disparities are large and has better performance overall, indicated by the lowest mean. On MR^T_{50}, while AR-CNN[16] has the best performance and we are comparable to them, AR-CNN has an inferior performance on MR^V_{50}. The reason might be because AR-CNN used thermal images as reference modality, making it more

Table 4 Comparison with state-of-the-art methods on KAIST dataset, with simulated disparity of misalignment in the horizontal direction by MR^V_{75} and MR^T_{75}, including their mean and SD over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Bold values indicate the best performance.

		Thermal images' horizontal shift distance (px)												
Methods		−10	−8	−6	−4	−2	0	2	4	6	8	10	Mean	SD
MSDS-RCNN[10]	MR^V_{75}	81.55	77.81	72.59	70.24	67.44	68.42	71.62	77.72	82.59	87.20	89.82	77.00	7.62
	MR^T_{75}	92.42	89.96	87.03	81.99	77.13	72.49	72.35	77.64	81.27	85.73	90.18	82.56	7.08
AR-CNN[16]	MR^V_{75}	96.60	94.88	93.51	88.88	81.35	62.92	71.47	78.89	88.97	94.76	97.22	86.31	11.32
	MR^T_{75}	**57.32**	56.97	**55.29**	55.33	**53.59**	61.57	56.50	56.83	57.81	59.13	60.80	57.38	2.38
MBNet[17]	MR^V_{75}	86.04	84.08	79.19	72.75	65.72	60.42	61.88	67.83	74.87	82.18	87.55	74.77	9.79
	MR^T_{75}	84.34	81.33	77.68	69.92	63.34	59.86	66.90	73.98	80.47	86.05	88.30	75.65	9.57
Ours	MR^V_{75}	**65.81**	**63.74**	**61.60**	**59.61**	**58.99**	**57.23**	**58.10**	**58.23**	**59.72**	**61.94**	**63.78**	**60.80**	**2.78**
	MR^T_{75}	59.35	**56.96**	56.40	**55.17**	54.59	**54.77**	**55.01**	**55.95**	**57.44**	**58.28**	**60.19**	**56.74**	**1.91**



(a) MSDS-RCNN[2] (b) AR-CNN[1] (c) MBNet[17] (d) Ours

Fig. 6 Qualitative comparison examples of detection results on KAIST dataset of (a) MSDS-RCNN;[10] (b) AR-CNN;[16] (c) MBNet;[17] and (d) ours. Scene N1 and D1 are original test images without simulated disparity. Scene N2 and D2 are test images with simulated disparity from shifting 10 pixels to the left and right direction, respectively. Green bounding boxes represent ground truth by Zhang et al.,[16] and red bounding boxes represent detection results. Dashed line bounding boxes denote substituted bounding boxes for methods that do not have paired bounding boxes.

biased toward thermal modality. Meanwhile, our method can maintain good performance in both modalities at the same time. Generally, we have significantly better performance in visible modality and comparable performance to AR-CNN in thermal modality. From Table 4, we have significantly better MR_{75}^{V} at all disparities and comparable performance with AR-CNN on MR_{75}^{T}. Overall, we have the best performance in both evaluation metrics and are the most robust to misalignment, indicated by the lowest SD. In summary, our bounding boxes are well placed in both modalities despite the misalignment compared to other SOTA methods.

4.4.2 *Qualitative comparison*

We provide visualization of detection results from several state-of-the-art methods on four scenes from KAIST[8] test set with a varied amount of misalignment to measure each method's quality in terms of detection precision and reliability. For methods that only have detection bounding boxes localizing objects in one modality, we replicated detection bounding boxes in the other modality with the same position and showed them as dashed line bounding boxes, i.e., we replicated thermal bounding boxes for MSDS-RCNN[10] and MBNet,[17] visible bounding boxes for AR-CNN.[16]

Scene N1 is a scene at nighttime with no misalignment between modalities and without simulated disparity. Scene D1 is a scene at daytime with significant distortion, causing each pedestrian to have different disparities, especially the leftmost pedestrian, without simulated disparity. Scene N2 is a scene at nighttime with slightly weak misalignment between modalities, and we add simulated disparity by shifting the thermal image by 10 pixels to the left direction. Scene D2 is a scene at daytime with huge misalignment from the beginning. We then add simulated disparity by shifting the thermal image by 10 pixels to the right direction for larger misalignment. We evaluate the mean $\mathrm{IoU}^{\mathrm{M}}$ of all detection bounding boxes with the highest $\mathrm{IoU}^{\mathrm{M}}$ overlap with each ground truth bounding box with at least 0.01 prediction score for each scene to measure the quality of bounding boxes of each method. As illustrated in Fig. 6, in Scene N1, our method could not demonstrate its strength since there is no misalignment. Moreover, the multi-modal regression also causes detection bounding boxes in visible modality to regress without clear information instead of staying at the same place as thermal bounding boxes, degrading the precision even further. In Scene D1, we can notice the more precise detection bounding boxes of the proposed method, especially at the leftmost pedestrian. Our method was able to adjust the detection bounding boxes in thermal modality closer to the pedestrian, despite the extreme misalignment. Still, the adjustment is not so great, which might be caused by the unusual pedestrian in dark color in thermal modality instead of bright color. In Scene N2 and D2, when misalignment is large, our method clearly shows its advantage by adjusting bounding boxes' positions for both modalities, resulting in the highest $\mathrm{mIoU}^{\mathrm{M}}$.

4.5 *Ablation Study*

We conduct ablation experiments to investigate our network's components and analyze each component's effectiveness. First, we compare the performance of a typical two-stream faster R-CNN with our proposed models composed of either only multi-modal RPN or multi-modal detector and both of them. Second, we compare the performance of our proposed models trained and tested by different NMS criteria for both RPN and detector. Lastly, we compare the performance of our proposed models trained by different mini-batch sampling criteria for both RPN and detector.

4.5.1 *Regressor comparison*

As given in Table 5, we can see that multi-modal regressor RPN does not significantly improve from a single-regressor, showing that the detector could not fully utilize RPNs output. Unquestionably, the performance improves drastically when a multi-modal regressor detector is implemented, especially with large misalignment, showing the benefit of locating objects in each modality individually. Finally, our network with both components has the best performance, indicating the effectiveness of both multi-modal RPN and multi-modal detector combined.

4.5.2 *NMS comparison*

As given in Table 6, $\mathrm{IoU}^{\mathrm{M}}$ outperforms other criteria in this experiment. $\mathrm{IoU}^{\mathrm{T}}$ also performs surprisingly well as the first runner-up. The reason might be that most of the pedestrians in KAIST dataset can be recognized by thermal modalities alone, and the misalignment is not significant in most test sets. Nevertheless, this experiment shows that $\mathrm{IoU}^{\mathrm{M}}$ can be utilized as NMS criteria for multi-modal pedestrian detection and perform the best compared to other criteria.

Table 5 Comparison of the proposed multi-modal Faster R-CNN consisting of different components on KAIST dataset, with simulated disparity of misalignment in the horizontal direction by MR$_{50}^{M}$, including their mean and SD over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Bold values indicate the best performance.

		Thermal images' horizontal shift distance (px)											
RPN's regressor	Detector's regressor	−10	−8	−6	−4	−2	0	2	4	6	8	10	Mean
Single	Single	24.41	16.42	13.08	11.03	10.19	10.07	11.03	11.66	13.69	15.81	21.15	14.41
Multi	Single	24.33	17.04	12.92	10.57	9.93	9.99	10.59	11.48	12.49	15.09	20.46	14.08
Single	Multi	18.54	13.25	**10.20**	9.17	8.57	8.96	9.21	9.90	11.44	13.19	17.18	11.78
Multi	Multi	**15.46**	**11.60**	10.21	**8.51**	**8.43**	**8.28**	**8.50**	**9.14**	**10.31**	**12.51**	**15.87**	**10.80**

Table 6 Comparison of the proposed multi-modal Faster R-CNN consisting of different NMS criteria on KAIST dataset, with simulated disparity of misalignment in the horizontal direction by MR$_{50}^{M}$, including their mean and SD over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Bold values indicate the best performance.

	Thermal images' horizontal shift distance (px)											
NMS criteria	−10	−8	−6	−4	−2	0	2	4	6	8	10	Mean
IoU$^{\mathrm{V}}$	16.17	13.18	11.67	10.71	11.10	10.82	11.42	12.10	13.38	14.56	18.76	13.08
IoU$^{\mathrm{T}}$	15.86	12.24	10.33	9.08	8.82	9.00	9.08	10.17	11.41	13.30	17.60	11.54
OR	16.95	12.64	11.37	9.92	10.24	9.85	10.06	10.72	12.09	13.97	16.53	12.21
AND	20.51	15.18	11.84	9.98	9.50	9.38	9.43	10.29	11.51	15.05	18.98	12.88
IoU$^{\mathrm{M}}$	**15.46**	**11.60**	**10.21**	**8.51**	**8.43**	**8.28**	**8.50**	**9.14**	**10.31**	**12.51**	**15.87**	**10.80**

Table 7 Comparison of the proposed multi-modal Faster R-CNN trained by different mini-batch sampling criteria on KAIST dataset, with simulated disparity of misalignment in the horizontal direction by MR$_{50}^{M}$, including their mean and SD over all shifted distances. Positive horizontal shift distances mean shifting to the right direction, and negative horizontal shift distances mean shifting to the left direction. Bold values indicate the best performance.

	Thermal images' horizontal shift distance (px)											
Sampling criteria	−10	−8	−6	−4	−2	0	2	4	6	8	10	Mean
IoU$^{\mathrm{V}}$	20.77	16.82	14.42	13.38	13.28	13.25	13.27	14.73	15.36	17.34	21.31	15.81
IoU$^{\mathrm{T}}$	14.75	11.89	10.90	9.62	9.65	9.76	9.53	10.16	10.85	13.40	16.30	11.53
OR	**14.54**	11.66	10.70	10.03	9.07	9.23	10.07	10.10	11.40	12.91	**15.26**	11.36
AND	17.03	13.07	10.92	9.97	9.39	9.46	9.62	10.76	11.74	14.38	18.52	12.26
IoU$^{\mathrm{M}}$	15.46	**11.60**	**10.21**	**8.51**	**8.43**	**8.28**	**8.50**	**9.14**	**10.31**	**12.51**	15.87	**10.80**

4.5.3 *Mini-batch sampling comparison*

From Table 7, IoU$^{\mathrm{M}}$ has the best performance when shift distance is less than or equal to 8 and also achieves the lowest mean MR$^{\mathrm{M}}$. However, IoU$^{\mathrm{M}}$ is inferior to OR when the shift distance is

10. The reason could be that OR is better at learning extreme cases, such as very large misalignment. Still, it has worse performance at no or weak misalignment as a trade-off. Overall, IoU^{M} has the best performance. It is worth considering how to make a multi-modal network perform well at any level of misalignment.

5 Conclusion

In this article, we have analyzed the current misalignment problem of existing multi-modal pedestrian detection methods. We have proposed the novel multi-modal detection method based on modal-wise regression and IoU^{M}, consisting of multi-modal NMS, multi-modal RPN, and multi-modal detector. We have also introduced new evaluation metrics for multi-modal detection, IoU^{M} and MR^{M}. The proposed method is robust to large misalignment, independently localizes pedestrians in each modality, and keeps their paired relations. Our experiments showed that when the misalignment is large or the precision requirement of bounding boxes is high, our proposed method achieves the best performance compared to state-of-the-art methods, demonstrating our robustness to misalignment and superior precision of detection bounding boxes in both modalities. However, the performance of our method when there is no misalignment is still lackluster, and even though we achieve the best performance when misalignment is significant, it is still not good enough to be reliable in crucial real-life applications such as autonomous driving cars, which has no room for any error. According to experimental results, we will improve our network's performance in more cases in our future work, not only in cases of misalignment. Moreover, no method currently independently infers objects' position for both modalities and builds their pair relations besides ours. We hope future research will consider our concern and use our proposed evaluation metric to evaluate multi-modal detection with misalignment.

References

1. E. Yurtsever et al., "A survey of autonomous driving: common practices and emerging technologies," *IEEE Access* **8**, 58443–58469 (2020).
2. U. Gawande, K. Hajari, and Y. Golhar, *Recent Trends in Computational Intelligence*, IntechOpen Publisher (2020).
3. B. Yang et al., "Convolutional channel features," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)* (2015).
4. L. Zhang et al., "Is faster R-CNN doing well for pedestrian detection?," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 443–457 (2016).
5. J. Mao et al., "What can help pedestrian detection?," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 6034–6043 (2017).
6. J. Li et al., "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia* **20**(4), 985–996 (2018).
7. S. Zhang et al., "Occlusion-aware R-CNN: detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vision (ECCV)* (2018).
8. S. Hwang et al., "Multispectral pedestrian detection: benchmark dataset and baseline," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 1037–1045 (2015).
9. A. González et al., "Pedestrian detection at day/night time with visible and fir cameras: a comparison," *Sensors* **16**(6), 820 (2016).
10. C. Li et al., "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Br. Mach. Vision Conf. (BMVC)* (2018).
11. J. Liu et al., "Multispectral deep neural networks for pedestrian detection," in *Br. Mach. Vision Conf. (BMVC)*, pp. 73.1–73.13 (2016).
12. K. Park, S. Kim, and K. Sohn, "Unified multi-spectral pedestrian detection based on probabilistic fusion networks," *Pattern Recognit.* **80**, 143–155 (2018).
13. C. Li et al., "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.* **85**, 161–171 (2019).
14. W. Treible et al., "Cats: a color and thermal stereo benchmark," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)* (2017).

15. Y. Choi et al., “Kaist multi-spectral day/night data set for autonomous and assisted driving,” *IEEE Trans. Intell. Transp. Syst.* **19**(3), 934–948 (2018).
16. L. Zhang et al., “Weakly aligned cross-modal learning for multispectral pedestrian detection,” in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)* (2019).
17. K. Zhou, L. Chen, and X. Cao, “Improving multispectral pedestrian detection by addressing modality imbalance problems,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 787–803 (2020).
18. N. Wanchaitanawong et al., “Multi-modal pedestrian detection with large misalignment based on modal-wise regression and multi-modal IoU,” in *2021 17th Int. Conf. Mach. Vision and Appl. (MVA)*, IEEE, pp. 1–6 (2021).
19. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit. (CVPR’05)*, Vol. 1, pp. 886–893 (2005).
20. P. Dollár et al., “Integral channel features,” in *Br. Mach. Vision Conf. (BMVC)* (2009).
21. P. Dollár et al., “Pedestrian detection: a benchmark,” in *Conf. Comput. Vision and Pattern Recognit. (CVPR)* (2009).
22. P. Dollár et al., “Fast feature pyramids for object detection,” *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* **36**(8), 1532–1545 (2014).
23. J. Wagner et al., “Multispectral pedestrian detection using deep fusion convolutional neural networks,” in *Eur. Symp. Artif. Neural Networks, Comput. Intell. and Mach. Learn. (ESANN)* (2016).
24. H. Choi et al., “Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks,” in *23rd Int. Conf. Pattern Recognit. (ICPR)*, pp. 621–626 (2016).
25. D. König et al., “Fully convolutional region proposal networks for multispectral person detection,” in *IEEE Conf. Comput. Vision and Pattern Recognit. Workshops (CVPRW)*, pp. 243–250 (2017).
26. D. Xu et al., “Learning cross-modal deep representations for robust pedestrian detection,” in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 4236–4244 (2017).
27. D. Guan et al., “Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection,” *Inf. Fusion* **50**, 148–157 (2019).
28. L. Zhang et al., “Cross-modality interactive attention network for multispectral pedestrian detection,” *Inf. Fusion* **50**, 20–29 (2019).
29. H. Zhang et al., “Guided attentive feature fusion for multispectral pedestrian detection,” in *Proc. IEEE/CVF Winter Conf. Appl. of Comput. Vision (WACV)*, pp. 72–80 (2021).
30. S. Li, X. Kang, and J. Hu, “Image fusion with guided filtering,” *IEEE Trans. Image Process.* **22**(7), 2864–2875 (2013).
31. T. Rukkanchanunt, M. Tanaka, and M. Okutomi, “Full thermal panorama from a long wavelength infrared and visible camera system,” *J. Electron. Imaging* **28**(3), 033028 (2019).
32. T. Shibata, M. Tanaka, and M. Okutomi, “Misalignment-robust joint filter for cross-modal image pairs,” in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, pp. 3295–3304 (2017).
33. T. Rukkanchanunt et al., “Disparity map estimation from cross-modal stereo,” in *IEEE Global Conf. Signal and Inf. Process. (GlobalSIP)*, IEEE, pp. 988–992 (2018).
34. Y. Ogino et al., “Coaxial visible and FIR camera system with accurate geometric calibration,” *Proc. SPIE* **10214**, 1021415 (2017).
35. S. Kim et al., “DASC: dense adaptive self-correlation descriptor for multi-modal and multispectral correspondence,” in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 2103–2112 (2015).
36. J. Dong et al., “Learning to align images using weak geometric supervision,” in *Int. Conf. 3D Vision (3DV)*, IEEE, pp. 700–709 (2018).
37. T. Shibata, M. Tanaka, and M. Okutomi, “Accurate joint geometric camera calibration of visible and far-infrared cameras,” *Electron. Imaging* **29**(11), 7–13 (2017).
38. M. Everingham et al., “The pascal visual object classes (VOC) challenge,” *Int. J. Comput. Vision* **88**, 303–338 (2010).
39. T.-Y. Lin et al., “Microsoft COCO: common objects in context,” in *Eur. Conf. Comput. Vision*, pp. 740–755 (2014).

40. P. Dollar et al., "Pedestrian detection: an evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(4), 743–761 (2012).
41. L. Zhang et al., "Weakly aligned feature fusion for multimodal object detection," *IEEE Trans. Neural Networks Learn. Syst.* 1–15 (2021).
42. S. Ren et al., "Faster R-CNN: towards real-time object detection with region proposal networks," in *Adv. in Neural Inf. Process. Syst.*, pp. 91–99 (2015).
43. K. He et al., "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, pp. 2980–2988 (2017).
44. R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 580–587 (2014).
45. R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, pp. 1440–1448 (2015).
46. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent. (ICLR), Conf. Track Proc.*, 7–9 May, San Diego, Califronia (2015).
47. J. Deng et al., "Imagenet: a large-scale hierarchical image database," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 248–255 (2009).
48. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. in Neural Inf. Process. Syst.*, pp. 1097–1105 (2012).

Napat Wanchaitanawong received his bachelor of engineering from the Department of Computer Engineering, Chulalongkorn University, in 2017, and received his master of engineering from the Department of Systems and Control Engineering, Tokyo Institute of Technology in 2021. He is currently a PhD student at the Department of Systems and Control Engineering, Tokyo Institute of Technology. His research interests include computer vision and image processing.

Masayuki Tanaka received his bachelor's and master's in control engineering, and a PhD from Tokyo Institute of Technology in 1998, 2000, and 2003. He was a software engineer at Agilent Technology from 2003 to 2004. He was a research scientist at Tokyo Institute of Technology from 2004 to 2008. He was an associate professor at the Graduate School of Science and Engineering, Tokyo Institute of Technology, from 2008 to 2016. He was a visiting scholar at Department of Psychology, Stanford University, from 2013 to 2014. He was an associate professor at the School of Engineering, Tokyo Institute of Technology, from 2016 to 2017. He was a senior researcher at National Institute of Advanced Industrial Science and Technology from 2017 to 2020. Since 2020, he has been an associate professor at the School of Engineering, Tokyo Institute of Technology.

Takashi Shibata received his BS and MS degrees form the Department of Physics, Tohoku University, in 2005 and 2007, respectively. He received his PhD from Tokyo Institute of Technology in 2017. He joined NEC Corporation in 2008. Since 2020, he has been a principal researcher at NTT Corporation. His research interests include image processing and pattern recognition.

Masatoshi Okutomi received his BEng from the Department of Mathematical Engineering and Information Physics, the University of Tokyo, Japan, in 1981 and an MEng from the Department of Control Engineering, Tokyo Institute of Technology, Japan, in 1983. He joined Canon Research Center, Canon Inc., Tokyo, Japan, in 1983. From 1987 to 1990, he was a visiting research scientist in the School of Computer Science at Carnegie Mellon University. In 1993, he received a DEng for his research on stereo vision from Tokyo Institute of Technology. Since 1994, he has been with Tokyo Institute of Technology, where he is currently a professor in the Department of Systems and Control Engineering, the School of Engineering.