Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection
Supplementary Material

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

1. Detection results on the KAIST dataset

Miss Rate We use MR, MR\textsuperscript{C}, and MR\textsuperscript{T} to evaluate the detection results, and compare the proposed AR-CNN method with other state-of-the-art methods (i.e. [2, 6, 3, 5, 1, 7, 4]) in Figure 1. The miss rate curves are corresponding to Table 2 in the main paper.

Visualization In Figure 2, we show some visualizations of detection results of the proposed AR-CNN. For the pedestrians with position shift problem, proposals of the sensed (color) modality are adjusted to aligned position. This phenomenon demonstrates that the Region Feature Alignment module can predict the region-wise position shift of two modalities and adaptively adjust the sensed proposals, thus enabling modality-aligned feature fusion process for better classification and localization.

2. Experiments on the color reference

In this section, we fix the color image as the reference modality. Table 1 shows that our AR-CNN still achieves the best performance and the smallest standard deviation, further validating the effectiveness of the proposed approach. Additionally, 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.

3. KAIST-Paired annotation examples

In Figure 3, we show some examples of our KAIST-Paired annotation. The bounding boxes are localized in both modalities, and a unique index is assigned to indicate the pairing relationship.

References


\textsuperscript{*}Corresponding author
Table 1. Quantitative results of the robustness to thermal position shift (i.e., we fix the color image while shifting the thermal image) on the KAIST dataset. MR$^T$ is used to evaluate the detection performance.
Figure 2. Qualitative results of the proposed method. The first row shows the reference proposals whose confidence score (in range [0, 1.0]) is greater than 0.6, while the second row illustrates the corresponding sensed proposals. In the third row, we select some proposal instances to demonstrate the effectiveness of the Region Feature Alignment module: orange dotted boxes refer to the reference proposals, which are good ones in the reference image but suffer the shift problem in the sensed modality; red bounding boxes denote the adjusted sensed proposals after the region feature alignment process. Green bounding boxes in the last two rows are the final predicted pedestrians whose confidence score is greater than 0.6.
Figure 3. Examples of our KAIST-Paired annotation. Bounding boxes in green, yellow and red indicate no-occlusion, partial occlusion, and heavy occlusion respectively. The red characters above the boxes denote the pairing information.