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

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

## 1. Detection results on the KAIST dataset

**Miss Rate** We use MR, $\text{MR}^C$, and $\text{MR}^T$ 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.

*Corresponding author

## References

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

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

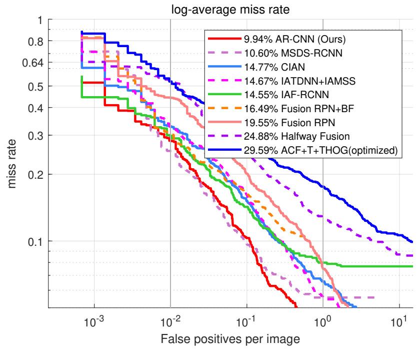
[3] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 243–250, 2017.

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

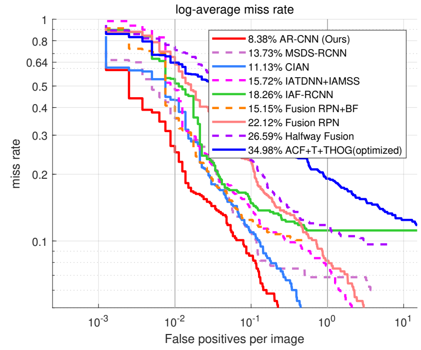
[5] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.

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

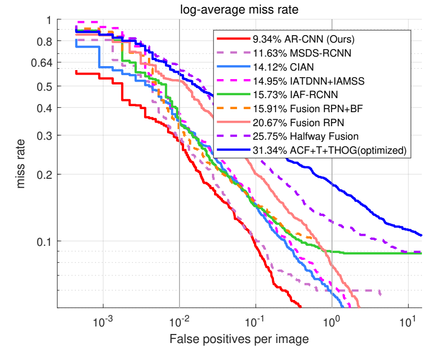
[7] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019.



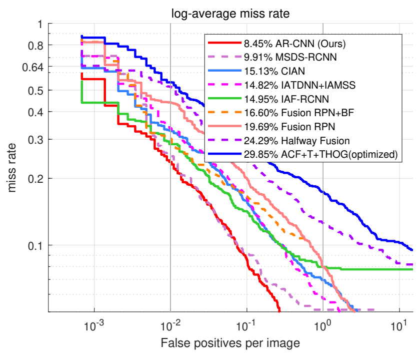

(a) Day, MR



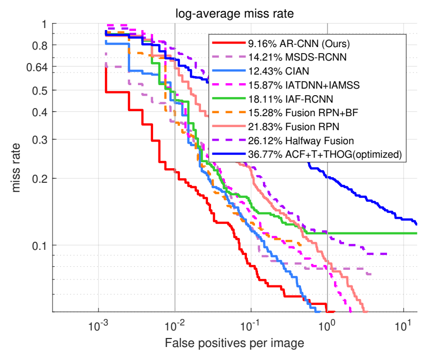

(b) Night, MR



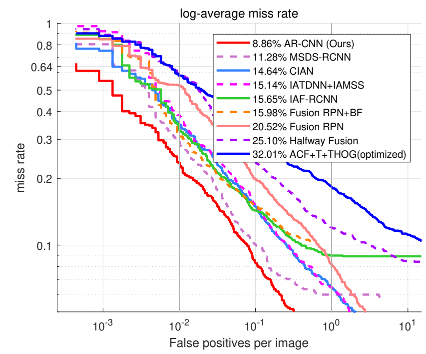

(c) All-day, MR



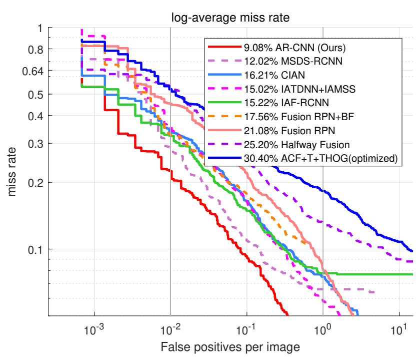

(d) Day, MR$^C$



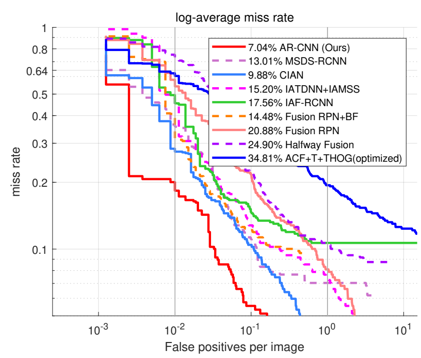

(e) Night, MR$^C$



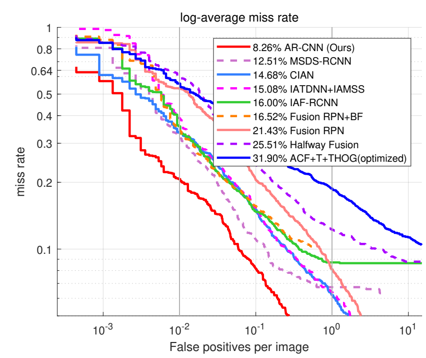

(f) All-day, MR$^C$




(g) Day, MR$^T$




(h) Night, MR$^T$




(i) All-day, MR$^T$

Figure 1. Comparisons with the state-of-the-art methods on the KAIST dataset. The scores in the legend are the log-average miss rate scores of the corresponding methods.

| Method | $S^{0^\circ}$ | | | $S^{45^\circ}$ | | $S^{90^\circ}$ | | $S^{135^\circ}$ | |
|---|---|---|---|---|---|---|---|---|---|
| | $O$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| Halfway Fusion [6] | 25.10 | 31.65 | 5.26 | 34.66 | 7.85 | 28.71 | 2.46 | 33.74 | 8.16 |
| Fusion RPN [3] | 20.52 | 29.12 | 7.10 | 29.06 | 9.20 | 22.14 | 1.59 | 30.02 | 9.95 |
| Adapted Halfway Fusion | 15.06 | 22.24 | 6.98 | 25.96 | 11.33 | 17.29 | 2.30 | 26.49 | 11.53 |
| CIAN [7] | 14.64 | 22.06 | 7.91 | 24.82 | 11.06 | 15.82 | 1.26 | 25.43 | 11.07 |
| MSDS-RCNN [4] | 11.28 | 18.21 | 6.74 | 21.00 | 9.66 | 12.93 | 1.28 | 21.71 | 10.12 |
| AR-CNN (Ours) | **8.86** | **10.86** | **1.58** | **11.66** | **2.59** | **9.51** | **0.62** | **11.47** | **2.42** |

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$^C$ 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.