https://doi.org/10.2352/EI.2024.36.17.AVM-111

Multi-Modal Pedestrian Detection via Dual-Regressor and Object-Based Training for One-Stage Object Detection Network

Napat Wanchaitanawong [a]***, Masayuki Tanaka*** [a]***, Takashi Shibata*** [b]***, Masatoshi Okutomi*** [a]***;***
[a] ***Tokyo Institute of Technology, Tokyo, Japan,*** [b] ***NTT Corporation, Kanagawa, Japan***

Abstract

Multi-modal pedestrian detection has been developed actively in the research field for the past few years. Multi-modal pedestrian detection with visible and thermal modalities outperforms visible-modal pedestrian detection by improving robustness to lighting effects and cluttered backgrounds because it can simultaneously use complementary information from visible and thermal frames. However, many existing multi-modal pedestrian detection algorithms assume that image pairs are perfectly aligned across those modalities. The existing methods often degrade the detection performance due to misalignment. This paper proposes a multi-modal pedestrian detection network for a one-stage detector enhanced by a dual-regressor and a new algorithm for learning multi-modal data, so-called object-based training. This study focuses on Single Shot MultiBox Detector (SSD), one of the most common one-stage detectors. Experiments demonstrate that the proposed method outperforms current state-of-the-art methods on artificial data with large misalignment and is comparable or superior to existing methods on existing aligned datasets.

Introduction

Pedestrian detection is one of the important research topics in computer vision [1,2]. The one-stage detection network [3,4], including SSD [5], is one of the standard architectures for many of these vision applications. However, using only visible modality has limitations, such as adverse lighting conditions or cluttered backgrounds. Multiple modalities, such as visible and thermal modalities, have been used together to surpass these limitations for pedestrian detection. Recent studies present various approaches that can combine different modalities' information.

One of the main challenges of multi-modal pedestrian detection is the misalignment problem. Most existing methods are under the assumption that the alignment of those image pairs is close to perfect. Those methods suffer performance degradation when there is misalignment. More recent methods [6] directly addressed this issue. They proposed an alignment module to adaptively align features between two modalities, which improved robustness against misalignment. However, their performance is still lackluster when the degree of misalignment is large. Furthermore, their method is only applicable to two-stage detection networks. This paper aims to design a single-stage multi-modal pedestrian detection network that is robust against large misalignment while keeping its performance in no misalignment cases.

Our work introduces several contributions, including: i) The incorporation of a dual-regressor within a single-stage multi-modal pedestrian detection network. ii) A specialized training protocol and data augmentation strategy tailored for the dual-regressor. iii) An evaluation metric specifically designed for multi-modal data with misalignment, accompanied by experiments demonstrating the effectiveness of our proposed method across data with varying degrees of misalignment.

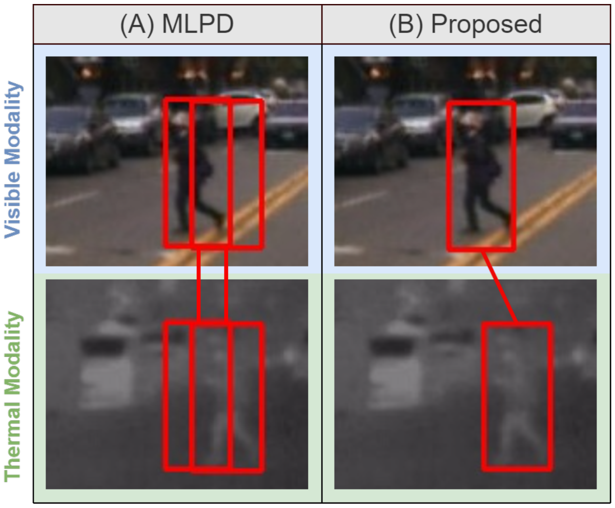


Figure 1: Visualization examples of detection results on KAIST dataset by (A) MLPD [7] and (B) our proposed method. Red boxes represent predicted bounding boxes and the lines between them indicate their pair relation.

Related Works

Multi-modal pedestrian detection

Researchers transitioned from exclusively utilizing color images in pedestrian detection task to incorporating color-thermal images, driven by their advantageous ability to leverage information from both modalities to compensate for each other's weaknesses. KAIST Multispectral Pedestrian Detection dataset [8] has been extensively employed in the realm of multi-modal pedestrian detection research, contributing to its ongoing advancements. In recent times, The field has clearly shifted towards favoring CNN-based methods [9–20] due to their exceptional performance relative to conventional methods. Nonetheless, a primary challenge in the initial stages revolved around effectively combining and leveraging information from both modalities [21–24].

Misalignment handling

The majority of existing methods operate under the assumption that visible-thermal image pairs are geometrically aligned, as shown in Fig 1(A), the bounding boxes do not precisely locate the objects in both modalities when significant misalignment is present. These approaches directly fuse features from both modal-

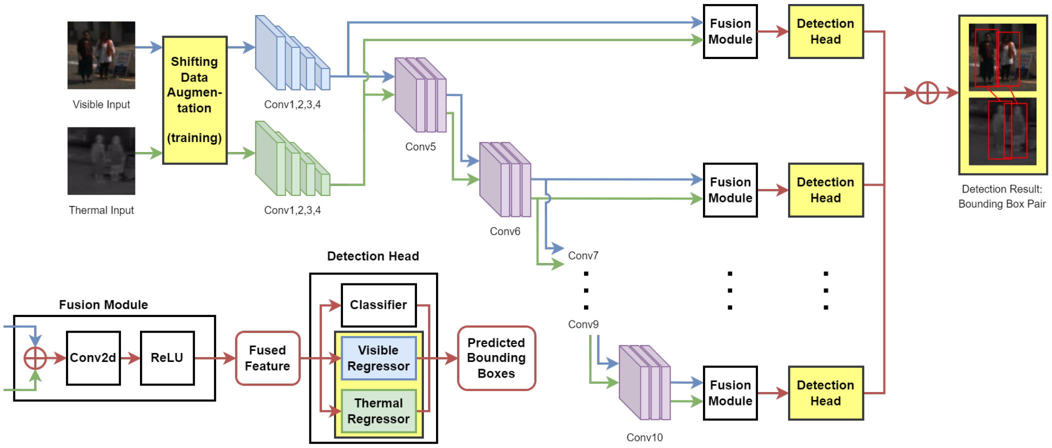


Figure 2: The overall architecture of our network. The framework is based on SSD [5] customized by MLPD [7]. Yellow blocks represent notable changes introduced in our method: shifting data augmentation in the training phase, detection heads with visible regressors and thermal regressors, and detection outputs consisting of pairs of bounding boxes. Blue, green, and red blocks/paths represent properties of visible modality, thermal modality, and fused modalities, respectively. $\oplus$ denotes channel-wise concatenation.

ities at corresponding pixel positions. Despite numerous proposals for geometric calibration and image alignment in the realm of multi-modal cameras [25–28], achieving precise and dense alignment for every pixel remains an ongoing challenge. Consequently, detectors utilizing these methods experience significantly degraded performance in regions where alignment is suboptimal. Recent advancements in methods, such as AR-CNN [6], have directly addressed this issue. They introduced an alternative annotation for the KAIST dataset, KAIST-paired annotation, pinpointing the object positions for each modality separately. This is especially invaluable for training models in multi-modal detection scenarios where misalignment is prevalent. They also further innovated with an alignment module designed to dynamically harmonize features between two modalities, elevating robustness against misalignment. MBNet [29] proposed Modality Alignment (MA) module which predicts offsets for every pixel in each modality to achieve effective alignment. Notably, existing methods fall short of harnessing the full potential of the KAIST-paired annotation. Instead, they often utilize the paired annotation to detect pedestrians in only one modality, neglecting the wealth of information the KAIST-paired annotation could offer. We propose a novel method capable of producing bounding box pairs that explicitly account for misalignment, thus maximizing the utility of the KAIST-paired annotation.

Method

Overview of the proposed method

We present one-stage multi-modal pedestrian detection framework inspired by SSD [5] and MLPD [7], as shown in Fig 2. The visible and thermal inputs initially follow distinct branches, proceeding through shared convolutional layers. Note that shifting data augmentation is only applied in the training phase, which is an addition to semi-unpaired augmentation of MLPD [7]. They are then unified within a fusion module before being input into the detection head, where we implement the proposed dual-regressor. The final output of the network is a set of bounding box pairs locating objects in both modalities.

Dual-regressor for single-stage network

We propose a refinement to the conventional single-stage detection head, introducing a dual-regressor approach for multi-modal detection, as illustrated in Fig 2. Each detection head comprises a classifier, visible regressor, and thermal regressor. The dual-regressor outputs are parameterized coordinates that represent the predicted object in both visible and thermal modalities. In contrast to SSD [5], our modified loss function accounts for the dual-regressor setup, incorporating distinct regression losses for each modality-specific regressor. The overall loss function is expressed as follows:

$$\begin{aligned} L &= \sum_i L_{cls}\left(BB_i, \widehat{BB}_i\right) \\ &+ \sum_i \left[w_i^v L_{reg}\left(BB_i^v, \widehat{BB}_i^v\right) + w_i^t L_{reg}\left(BB_i^t, \widehat{BB}_i^t\right)\right], \end{aligned} \tag{1}$$

where i denotes the index of the anchor box, a predefined bounding box positioned at various points throughout the images. These anchor boxes serve the purpose of identifying objects within specific, designated regions. L_{cls} denotes classification loss, which is binary cross entropy with sigmoid activation function. L_{reg} denotes regression loss, employing L1 loss, BB_i^v, BB_i^t denote the visible and thermal ground truth bounding boxes of anchor box i, respectively. $\widehat{BB}_i^v$, $\widehat{BB}_i^t$ denote the predicted visible and thermal bounding boxes of anchor box i, respectively. w_i^v, w_i^t denote the visible and thermal mask, determined by multi-label of the object, adopted from MLPD [7]. In essence, w_i^v, w_i^t are set to 1 when the corresponding object is perceivable in the visible or thermal modality, respectively; otherwise, they are set to 0.

Cross-modal bounding boxes' overlapping and evaluation metric for multi-modal detection

Given the potential misalignment between modalities, the coordinates of each object in each modality may differ. To quantify the overlap between two pairs of bounding boxes, we employ the "Multi-Modal Intersection over Union (IoU^M) [20]" metric. In essence, IoU^M serves as the criterion for categorizing bounding boxes into positive or negative and is applied in non-maximum

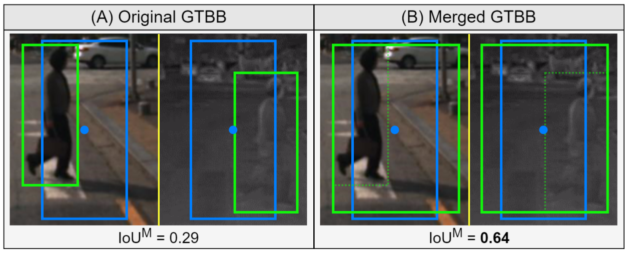


Figure 3: Visualization examples of (A) a ground truth bounding box (GTBB) pair from KAIST-paired annotation by [6] and (B) a merged ground truth bounding box. Green boxes represent ground truth bounding boxes. Blue boxes represent anchor boxes. Images are artificially shifted for better understanding.

suppression (NMS) instead of the traditional IoU for multi-modal data. The formula for IoU^M is defined as

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

where GT^V and GT^T denote paired ground truth bounding boxes of the same object in visible and thermal modalities, respectively. DT^V and DT^T denote paired detection bounding boxes of the same object in visible and thermal modalities, respectively. IoU^M serves as a valuable metric for assessing the accuracy of detection bounding boxes in both modalities, evaluating the system's capability to handle misalignment, and ensuring correct object matching between modalities. To quantify the performance, we employ "Multi-Modal Log-Average Miss Rate (MR^M)," which is derived from the Log-Average Miss Rate [30], using IoU^M as the criterion for bounding boxes' positive-negative categorization.

Object-based training and shifting data augmentation

As discussed in the context of the dual-regressor, visible and thermal ground truth bounding boxes is crucial for training the dual-regressor. We adopted the KAIST-paired annotation developed by Zhang et al. [6]. While various methods have employed this annotation in different ways, they often did not leverage its full potential. For instance, MBNet [29] merges visible and thermal bounding boxes of each pedestrian into a unified bounding box by averaging. MLPD [7] considers the same pedestrian in both visible and thermal modalities as two distinct objects, a methodology we will now label as 'bounding box-based (BB-based) training'. In contrast, our approach considers each pedestrian as a single object with two individual coordinates for visible and thermal modalities. For unpaired objects visible exclusively in one modality, whether solely in the visible or thermal domain, they are categorized as either visible-only or thermal-only objects, respectively. Subsequently, these objects are utilized to exclusively train either the visible or thermal regressor. This training approach is referred to as 'object-based training.' This distinction allows our method to precisely locate pedestrians in both modalities, even in the presence of significant misalignment.

In the sampling process, positive samples are chosen based on the overlap between each anchor box and the ground truth bounding box. To account for potential misalignment, We unify visible and thermal bounding boxes into a single bounding box by utilizing the farthest points in both horizontal and vertical directions from the vertices of the original bounding boxes. This consolidation can improve overlap computation with the anchor box, thereby reducing the likelihood of overlooking potential samples with significant misalignment. However, during regressor training, we maintain the original ground truth bounding boxes as targets for the proposed dual-regressors. The visualized example of ground truth bounding boxes is depicted in Fig 3, where the original ground truth bounding box (Fig 3(A)) serves as the target for our regressors' training: the visible regressor is trained with the visible ground truth bounding box, and the thermal regressor is trained with the thermal ground truth bounding box. In the sampling process, we utilize the merged ground truth bounding box (Fig 3(B)) to calculate the overlap with the anchor box. This approach enhances the overlap measurement, especially when misalignment is significant. Here, IoU^M increases from 0.29 to 0.64. This increase could be pivotal, potentially changing the sample's classification from negative to positive.

Our Non-Maximum Suppression (NMS) utilizes IoU^M as a suppression criterion to preserve pair relations between bounding boxes in the visible and thermal modalities. The process begins by categorizing each bounding box pair into three groups: visible-thermal, visible-only, or thermal-only objects, based on the prediction scores of both modalities. Pairs with prediction scores of both modalities below the specified threshold are classified as background and discarded. In the case where only one modality's prediction score surpasses the threshold, the pair is designated as a modality-specific object. Otherwise, it is identified as a visible-thermal object. Bounding box pairs are sorted by the average prediction scores between the visible and thermal modalities in a descending manner. The overlap calculation between pairs considers IoU^M, IoU of the visible modality (IoU^V), and IoU of the thermal modality (IoU^T). However, for visible-only or thermal-only objects, only the bounding box in the corresponding modality is considered, treating the other modality as non-existent. When either of the overlap thresholds is surpassed, the bounding box pair with the lower score is suppressed.

Furthermore, we incorporate shifting data augmentation to expose our network to misalignment scenarios, addressing a gap in the original dataset. This augmentation method involves randomly translating training images horizontally in one of the two modalities, with pixel shifts ranging from -10 to 10. This process is facilitated by a multinomial distribution, with probabilities derived from a normal distribution, to randomly determine the shift distance. Initially, the entire network is trained without shifting data augmentation. Subsequently, upon achieving a well-performing model on the validation dataset, we proceed to freeze all layers of the network except the regressors and retrain the previous checkpoint with shifting data augmentation. This step further enhances localization performance, particularly when dealing with misalignment data. This augmentation strategy contributes to the robustness of our model in handling misalignment challenges. Subsequently, other semi-unpaired augmentations adopted from MLPD [7] are still applied.

Experiments

Dataset

The KAIST dataset [8] stands out as one of the extensively utilized multi-modal pedestrian datasets, featuring over 90,000 deviation across shifted distances. Bold values indicate the best performance.

Methods	Thermal images' horizontal shift distance (px)											Mean	SD
	-10	-8	-6	-4	-2	0	2	4	6	8	10		
MSDS-RCNN [15]	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 [6]	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 [29]	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
MLPD [7]	21.33	13.07	9.57	**7.10**	7.07	6.97	7.89	9.49	10.59	15.27	21.86	11.84	5.48
Our past work [31]	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**
Ours	**14.82**	**10.21**	**8.20**	7.27	**6.76**	**6.84**	**7.21**	**8.34**	**9.35**	**11.93**	**15.22**	**9.65**	3.08

Table 2: Performance Evaluation of Varied Components and Training Strategies in the Proposed Network on the KAIST dataset with simulated disparity of misalignment by MR^M_{50}, mean, and standard deviation across shifted distances.

Type of regressor	Training strategy	Shifting data augmentation	Thermal images' horizontal shift distance (px)											Mean	SD
			-10	-8	-6	-4	-2	0	2	4	6	8	10		
Single	BB-based	-	22.91	15.01	9.74	7.87	7.04	7.17	8.43	9.66	11.04	15.19	21.12	12.29	5.57
Single	Object-based	-	20.36	13.27	8.78	7.72	6.93	7.21	7.69	9.18	10.32	13.61	20.41	11.41	4.98
Single	Object-based	✓	20.21	13.65	9.42	7.94	7.28	7.20	7.85	8.61	11.07	13.36	19.60	11.47	4.74
Dual	BB-based	-	19.74	12.92	9.84	8.24	6.90	6.94	7.95	9.50	10.51	14.04	20.34	11.54	4.77
Dual	Object-based	-	17.10	11.59	8.67	7.99	7.35	6.99	7.37	**8.34**	9.45	12.53	16.03	10.31	3.56
Dual	Object-based	✓	**14.82**	**10.21**	**8.20**	**7.27**	**6.76**	**6.84**	**7.21**	**8.34**	**9.35**	**11.93**	**15.22**	**9.65**	**3.08**

frames recorded during both day and night to account for varying light conditions. Initially presumed to be geometrically aligned, the dataset's annotations revealed numerous errors, including imprecise localization, misclassification, and misaligned regions, as reported by prior studies [6, 15]. To address these issues, several researchers [9] have created improved versions of annotations as alternatives to the original dataset.

Implementation and evaluation details

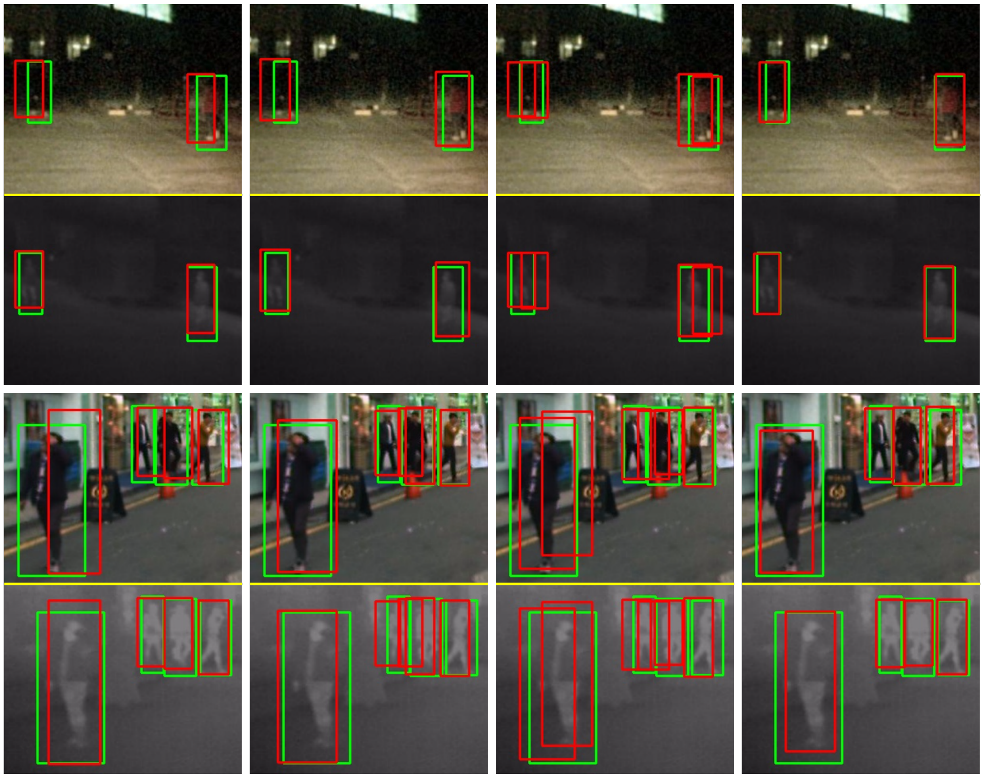
We adopted an SSD modified into MLPD [7]. The architecture utilized VGG16 pre-trained on ImageNet with batch normalization for Conv1 to Conv5, and the remaining convolutional layers (Conv6 onwards) were initialized with values drawn from a normal distribution (std=0.01). The model underwent training with Stochastic Gradient Descent (SGD), using an initial learning rate, momentum, and weight decay of 0.0001, 0.9, and 0.0005, respectively. The mini-batch size was set to 6, and the input image size was resized to 512 (H) x 640 (W) We integrated MLPD's semi-unpaired data augmentation, maintaining the same parameters, and introduced our shifting data augmentation to bolster the training process against misalignment. The standard deviation of the normal distribution for the shifting data augmentation was set to 4. The prediction score threshold of NMS is set to 0.1. The overlap threshold IoU^M, IoU^V, and IoU^T of NMS are set to 0.425, 0.75, and 0.75, respectively. First, we train the whole network without shifting data augmentation for 30 epochs. Then, we continue the training from last checkpoint only on dual-regressor with shifting data augmentation for another 30 epochs.

We conducted our experiments using KAIST Dataset [8]. Given our specific focus on addressing the misalignment problem, we adopted the annotations provided by Zhang et al. [6] for both training and testing. Recognizing that the test data did not include sufficient scenes with significant misalignment, we conducted a "simulated disparity experiment" to replicate misalignment at various degrees. We horizontally shift thermal images of the test data by 2, 4, 6, 8, and 10 pixels in both directions, while the visible images remained unchanged. This process will net us 11 subsets of test data with different degrees of misalignment. We evaluated the performance of our methods against all available state-of-the-art methods with accessible source code. For methods producing a single bounding box for each object, we substituted visible and thermal bounding boxes with that single bounding box. The detection performance was quantified using the Multi-Modal Log-Average Miss Rate (MR^M) over the range of $[10^{-2}, 10^0]$ False Positive Per Image (FPPI) with an IoU^M threshold of 0.5 (MR^M_{50}).

Comparison with existing methods

Performance comparison. Table 1 shows the performance comparison between various state-of-the-art methods, including our past work [20]. The proposed method emerges as the top-performing solution, consistently outperforming state-of-the-art approaches across various simulated disparity distances on the KAIST dataset. Specifically, at smaller misalignment distances, our model showcases performance comparable to the MLPD baseline, indicating that the introduced modifications maintain competitive accuracy under standard conditions. However, the strength of the proposed method emerges at larger misalignment distances (e.g., -10 pixels), where it significantly surpasses MLPD and other methods, which is particularly evident at larger misalignment distances, emphasizing the effectiveness of the proposed model in addressing challenges associated with substantial misalignment. Additionally, the proposed method consistently outperforms our previous work across all shift distances, demonstrating enhanced robustness and performance in handling misalignment challenges. The mean and standard deviation values further support the reliability and stability of the proposed method across diverse misalignment scenarios.

Qualitative comparison. Fig 4 illustrates comparison examples of detection results on the KAIST dataset, showcasing the performance of our method against other state-of-the-art approaches. i) First Scene: In a scene where pedestrians are separate but challenging to recognize due to dark lighting and substantial misalignment, our method stands out, producing precise bounding boxes for all pedestrians, whereas alternative methods either struggle to locate pedestrians accurately or generate multiple bounding boxes for a single pedestrian, leading to false positives. ii) Second Scene: In a more crowded scene where pedestrians are numerous and clearly distinct, but serious misalignment is present,



(a) AR-CNN [6] (b) MBNet [29] (c) MLPD [7] (d) ours

Figure 4: Qualitative comparison examples of detection results on KAIST dataset of (a) AR-CNN [6], (b) MBNet [29], (c) MLPD [7], and (d) ours. Green boxes represent ground truth bounding boxes. Red boxes represent predicted bounding boxes. Image pairs are cropped in the same position to make the contrast between methods more apparent. Prediction score threshold is set to 0.1. Thermal images of scene 1 and 2 are shifted to the left and right direction by 10 pixels, respectively.

our method showcases its ability to generate accurate bounding boxes for all pedestrians. In contrast, other methods encounter challenges in precise localization. For instance, MLPD resorts to creating two bounding box pairs for a single pedestrian, attempting to cover them in both modalities.

Impact of components and training strategies

Table 2 provides an insightful ablation study, exploring the impact of different components and training strategies on our proposed network's performance. We examined variations in the type of regressor, training strategy (BB-based or Object-based), and the inclusion of shifting data augmentation. The results indicate that the performances of single-regressor networks are almost the same. They could not utilize from the object-based training and misalignment data. Furthermore, the integration of dual-regressors, combined with a BB-based training strategy, does not lead to any performance improvement. This is because BB-based training does not consider the relationship between objects in different modalities. In contrast, combining dual-regressors with an object-based training strategy ensures precise pedestrian localization, facilitating accurate pedestrian matching even under varying degrees of misalignment, ultimately leading to improved performance. Additionally, the incorporation of shifting data augmentation allows the model to learn from data exhibiting diverse misalignment, providing valuable insights not present in the original training data and contributing to the best performance.

Conclusion

This paper proposed the one-stage multi-modal pedestrian detection network, leveraging a dual-regressor and object-based training to address the misalignment challenges prevalent in existing methods. The proposed Multi-Modal Log-Average Miss Rate (MR^M) metric provides a comprehensive evaluation criterion for multi-modal detection, accounting for misalignment. The simulated disparity experiments on the KAIST dataset demonstrated the superiority of our proposed method.

References

[1] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE TPAMI*, 2011.

[2] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: A benchmark," in *Proc. CVPR*. IEEE, 2009.

[3] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. ECCV,*. Springer, 2016.

[6] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen

Lei, and Zhiyong Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. ICCV*, 2019.

[7] Jiwon Kim, Hyeongjun Kim, Taejoo Kim, Namil Kim, and Yukyung Choi, "Mlpd: Multi-label pedestrian detector in multispectral domain," in *IEEE RA-L*, 2021.

[8] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. CVPR*, 2015.

[9] Shu Wang Jingjing Liu, Shaoting Zhang and Dimitris Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proc. BMVC*, 2016.

[10] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *Proc. ESANN*, 2016.

[11] Hangil Choi, Seungryong Kim, Kihong Park, and Kwanghoon Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," in *Proc. ICPR*, 2016.

[12] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. CVPRW*, 2017.

[13] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. CVPR*, 2017.

[14] Kihong Park, Seungryong Kim, and Kwanghoon Sohn, "Unified multi-spectral pedestrian detection based on probabilistic fusion networks," *Pattern Recognition*, 2018.

[15] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proc. BMVC*, 2018.

[16] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, 2019.

[17] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang, "Illumination-aware faster r-cnn for robust multispectral pedestrian detection," *Pattern Recognition*, 2019.

[18] 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*, 2019.

[19] Heng Zhang, Elisa Fromont, Sebastien Lefevre, and Bruno Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in *Proc. WACV*, 2021.

[20] Napat Wanchaitanawong, Masayuki Tanaka, Takashi Shibata, and Masatoshi Okutomi, "Multi-modal pedestrian detection with large misalignment based on modal-wise regression and multi-modal iou," in *Proc. MVA*. IEEE, 2021.

[21] Shutao Li, Xudong Kang, and Jianwen Hu, "Image fusion with guided filtering," *IEEE TIP*, 2013.

[22] Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi, "Misalignment-robust joint filter for cross-modal image pairs," in *Proc. ICCV*, 2017.

[23] Thapanapong Rukkanchanunt, Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi, "Disparity map estimation from cross-modal stereo," in *Proc. GlobalSIP*. IEEE, 2018.

[24] Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi, "Unified image fusion framework with learning-based application-adaptive importance measure," *IEEE TCI*, 2018.

[25] Yuka Ogino, Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi, "Coaxial visible and fir camera system with accurate geometric calibration," in *Thermosense: Thermal Infrared Applications XXXIX*. Int. Society for Optics and Photonics, 2017.

[26] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn, "Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence," in *Proc. CVPR*, 2015.

[27] Jing Dong, Byron Boots, Frank Dellaert, Ranveer Chandra, and Sudipta Sinha, "Learning to align images using weak geometric supervision," in *Proc. 3DV*. IEEE, 2018.

[28] Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi, "Accurate joint geometric camera calibration of visible and far-infrared cameras," *Proc. EI*, 2017.

[29] Kailai Zhou, Linsen Chen, and Xun Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Proc. ECCV*, 2020.

[30] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE TPAMI*, 2012.

[31] Napat Wanchaitanawong, Masayuki Tanaka, Takashi Shibata, and Masatoshi Okutomi, "Multi-modal pedestrian detection with misalignment based on modal-wise regression and multi-modal IoU," *Journal of Electronic Imaging*, 2023.

Author Biography

Napat Wanchaitanawong *received his B.Eng. from Department of Computer Engineering, Chulalongkorn University in 2017 and received his M.Eng. from Department of Systems and Control Engineering, Tokyo Institute of Technology in 2021. He is currently a Ph.D. student at Department of Systems and Control Engineering, Tokyo Institute of Technology.*

Masayuki Tanaka *received his Ph.D. degree from Tokyo Institute of Technology in 2003. 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 Stanford University, from 2013 to 2014. He was an Associate Professor at 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. He was an Associate Professor at Tokyo Institute of Technology, from 2020 to 2023. Since 2023, he has been a Professor at Tokyo Institute of Technology.*

Takashi Shibata *received the M.S. degree form the Department of Physics, Tohoku university in 2007. He received the Ph.D. degree from Tokyo Institute of Technology in 2017. He joined NEC Corporation in 2008. From 2020 to 2022, he was a Principal Researcher at NTT Corporation. His research interests include image processing and pattern recognition.*

Masatoshi Okutomi *received a B.Eng. degree from the Department of Mathematical Engineering and Information Physics, the University of Tokyo, Japan, in 1981 and an M.Eng. degree 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, USA. In 1993, he received a D.Eng. degree 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.*