Dear Editorial Board of IEEE Robotics and Automation Letters

We revise and resubmit the manuscript entitled **"MLPD: Multi-Label Pedestrian Detector in Multispectral Domain"** for consideration by **"IEEE Robotics and Automation Letters"**. We confirm that this work is original and has not been published elsewhere in any format, nor is it currently under consideration for publication elsewhere.

Most multi-modality fusion methods have been developed by using the fully overlapped image pair. In multispectral pedestrian detection, the KAIST dataset is the most widely used dataset, which is taken with a special equipment called beam-splitter. Even though this equipment allows a zero baseline between two sensors, this kind of sensor system has a difficulty to apply to practical applications. From a practical view, the stereo setting is used as an alternative way. Unlike the beam-splitter system, this stereo system allows a certain distance between two sensors, so that two issues arise, which affect the fusion method and detection performance. The first issue is that there are non-overlapped areas in the image where only information from one sensor appears. The other issue is that there is a pixel-level alignment problem known as misalignment due to parallax. This problem is proportional to the baseline distance between two sensors, and it can occur in image pairs due to the asynchronization.

In the manuscript, we tackle the multispectral pedestrian detection by using a fully overlapped image pair, and we introduce a generalized multispectral pedestrian detection framework that detects the pedestrian in both paired (fully overlapped) and un-paired (partially overlapped) conditions. The main contents are summarized as follows: 1) We address constraints of previous fusion methods, which make the methods hard to be applied to real world applications and introduce a new perspective of the multispectral pedestrian detection in unpaired conditions; 2) We propose a generalized multispectral pedestrian framework for ideal and practical conditions, which is built upon multi-label learning with a semi-unpaired augmentation strategy; 3) We test the proposed method, considering various unpaired cases, and it achieves the comparable or better result in comparison to the state-of-the-art algorithm.

To demonstrate the validity of the proposed method, we conduct various experiments in multispectral pedestrian datasets, as well as synthetically generated datasets. That is, we conducted extensive experiments regarding various conditions: 1) General paired images; 2) Sensor failure 3) Stereo and EO/IR setting simulation. The experimental results indicate the superiority of our proposed method in paired and unpaired conditions in comparison to the state-of-the-art pedestrian detection algorithm.

We would like to thank you, in advance, for your consideration of the re-submitted work and are looking forward to receiving your decision on the manuscript. We hope you will find that our results would be of interest to readers of **IEEE *Robotics and Automation Letters***.

Sincerely,
Jiwon Kim, Hyeongjun Kim, Taejoo Kim, Namil Kim and Yukyung Choi

Revision Report of Manuscript RA-L 21-0716

Title: MLPD: Multi-Label Pedestrian Detector in Multispectral Domain

Authors: Jiwon Kim*1, Hyeongjun Kim*1, Taejoo Kim*1, Namil Kim2 and Yukyung Choi†1

We would like to thank all the reviewers for spending their time and making efforts to read our manuscript and provide useful suggestions to improve the work. We have read the comments carefully, and tried to follow the suggestions as closely as possible. Specific changes in this revision will be summarized below.

Editor:

The paper presents a pedestrian detection method that fuses RGB and thermal images using a modified SSD object detector. A challenge addressed in this work is handling imperfect alignment and synchronization of the multi-modal data stream. While the reviewers found the submission generally interesting there are several aspects that require improvement before it can be considered for publication.

Response: Again, thank you for allowing us to strengthen our manuscript with your valuable comments and queries. We have worked hard to incorporate your feedback and hope that these revisions persuade you to accept our submission.

[C1] There is some concern regarding the handling of alignment and synchronization of the images as the datasets used do not suffer significantly from these issues.

Response: Please, refer to [C5], [C6].

[C2] There is also some confusion with regards to the apparent lack of fusion being used in the CVC-14 dataset experiments.

Response: Please, refer to [C10], [C28].

[C3] Regarding the proposed architecture two concerns are raised: i) whether or not the architecture results in fusion being performed and ii) the existence of newer and better performing baseline architectures. A revision will need to address these concerns.

Response: Please, refer to [C8], [C14].

[C4]Finally, the entire paper should be checked for completeness of equations as well as updating the reference section to use the standard IEEE style.

Response: Thank you for raising this issue. We have thoroughly checked all the references to follow the IEEE style template.

Reviewer No.2:

The paper presents a single-stage detection framework for detecting pedestrians using RGB + thermal cameras. This paper's main motivation is that these two cameras may not be perfectly paired (aligned and/or synchronised) in practice, and its solution addresses this limitation. To this end, an SDD object detector (with a VGG backbone) has been manipulated to include a thermal input in addition to an RGB image. Then a multi-level fusion strategy is proposed to fuse the features from these two sensor modalities. In the output, the detected bounding boxes are labelled if they are either detected in RGB, thermal or both cameras. A data augmentation strategy suggested making the model robust to the unpaired cases. The method has been evaluated on KAIST and CVC-14 datasets, as well as a synthesised dataset based on KAIST.

Response: We would like to thank you for spending your time and for making efforts to read our manuscript and provide useful suggestions to improve the work.

Although the paper targets addressing an important problem in robotic perception, i.e. detection using multi-sensor modalities, the motivation, solution and experiments are not very convincing:

Response: Thank you for raising this point. We have added some explanations and experiments to follow the comments provided by the reviewers to further improve the quality of our manuscript.

[C5] the paper motivates the problem by arguing that the assumption of paired (aligned and/or synchronized) sensors is not practical (I agree with this), but it has been tested on the real datasets, which barely suffer from this problem.

Response:

In earlier ways, special equipment such as a beam splitter was used to mechanically match the principle point of RGB and thermal sensors. In this way, we can create a fully-overlapped image pair without non-overlapped areas in the image. A popular example of this dataset is the KAIST multispectral dataset. However, this kind of special equipment is difficult to manufacture, and the sensor system including such equipment becomes large in size as shown in Fig.1-(a). Our terminology “paired image” means such a fully overlapped image without non-overlapped areas in the image.

For practical and general approaches, there are efforts to avoid the special equipment, which allows the distance between RGB and thermal sensors as shown in Fig.1-(b) and (c) like stereo-setting.

In these approaches, we can obtain a partially overlapped image pair including both an overlapped area (all the sensor images) and a non-overlapped area (only a single sensor image). According to the sensor resolution, areas that do not overlap can appear on the sides or at the boundary. Our terminology "unpaired image" means such image containing both overlapped area and non-overlapped area.

In both cases, there is a difference in the overlapped area. In the KAIST pedestrian dataset, the beam splitter (Fig.1-(a)) allows us to capture the image pair in a zero baseline between two sensors. However, in stereo-setting (Fig.1-(b) and Fig.1-(c)), there is a pixel-level alignment problem (a.k.a misalignment) when the two images are registered by corresponding pixels with subpixel accuracy due to parallax. This alignment problem is proportional to the baseline distance between two sensors. Also, this pixel-wise alignment problem can occur in the "unsynchronized" case where two images are not captured, simultaneously.

In this perspective, our goal is to develop a multispectral pedestrian detector which can robustly recognize pedestrians regardless of whether a pair of images is perfectly overlapped (paired) or only partially overlapped (unpaired). It is not easy to obtain all the realistic unpaired datasets, we only use the fully-overlapped KAIST dataset to train the model, and to simulate the partially overlapped image pair to prove the effectiveness of the practical scenario. Given a fully overlapped image pair, and certain parameters such as the distance of baseline, it can geometrically generate almost realistic unpaired image pairs.

We apologize that our description and terminology are not clear to understand the motivation of the paper. In the revised manuscript, we added a more descriptive explanation of this issue and re-stated the definition of our terminology.

[C6] Also, I don't see that **any solution is suggested for unsynchronised sensors.**

Response:

As mentioned in [C5], the unsynchronized sensors can cause misalignment in the overlapped area. CVC-14 dataset was captured by the stereo-setting as shown in Fig.1-(b), so this dataset originally should be classified into unpaired datasets. However, the contributor of the dataset cropped the non-overlapped area in each image, so all the provided images are fully overlapped (paired dataset) as shown ***Review-Fig 1***. Moreover, the contributor mentioned that this dataset has some troubles in misaligned ground truth bounding boxes, incorrect extrinsic parameters in some sequences, and incorrect/missing ground truth in thermal images due to the unsynchronized capturing way and mistakes by providers.

Therefore, our intention for a solution to handle unsynchronized sensors was that the proposed method worked well in the CVC-14 dataset. In other words, we concluded that our model can robustly estimate the pedestrians in the overlapped area where some misalignment problem occurs from the unsynchronized condition.

We apologize that our description is not clear to convey our intention and to understand the result. In the revised manuscript, we entirely revised section-IV-A-3 for more descriptive explanations of this issue.

[P.4, IV.A.3).] Edited version: "The CVC-14 [17] dataset is a multispectral pedestrian dataset taken with a stereo camera configuration. The dataset is composed of Grey-Thermal pairs consisting of 7085 and 1433 frames for training and test sets, and provides individual annota-tions in each modality. Unlike the KAIST dataset in which two sensors are mechanically aligned, this dataset originally provides multispectral image pairs with non-overlapped areas and overlapped areas containing some misalignment issues. However, the author of the dataset released the cropped image pairs without the non-overlapped areas. Therefore, we treated this dataset as the fully-overlapped (paired) dataset for our purpose,but it still suffers from the pixel-level misalignment problem. Moreover, there are some issues, such as inaccu-rate ground truth boxes, incorrect extrinsic parameters, and unsynchronized capture systems. Nevertheless, this dataset has been used by many works [9], [12], [21], [22]because it is one of the few practical datasets captured in a stereo setup."

Review-Fig 1. *Camera setup for the CVC-14 dataset and registered sample frames showing different field of view. This figure shows the difference between paired images and unpaired images. The CVC-14 dataset provides only 'Paired' image pairs.*

[C7] the contribution is incremental (perhaps only fusion strategy can be considered as a main contribution of the paper).

Response:

Thank you for giving us an insightful comment on our paper. As you mentioned, one of the main contributions of our paper is the fusion strategy. Another proposed contribution is a general and scalable multispectral pedestrian detection framework. Most of the previous works [4,5,6,7,9,10,11,12,13,19,21,22], using the KAIST dataset can work well in a fully-overlapped image pair condition, not in a partially overlapped image pair condition. However, with fully-overlapped image pairs, the proposed multi-class labeling and augmentation strategy can train the model to work well in both paired and unpaired image pairs. Lastly, we firstly provide a new benchmark of various multispectral pedestrian detection models in realistic unpaired conditions (containing non-overlapped areas in the image). We expect that our research will be a milestone

to step forward to the next level in future work. To improve the delivery of the contribution, we described this issue in the revised manuscript.

[C8] The model used (SSD with VGG backbone) is not state-of-the-art detection models. We have more efficient ` like YOLO v3, v4& v5, FCOS (ICCV 2019), DETR (ECCV 2020) with much better performance compared to SSD and many better backbones like (ResNet).

Response:

Thank you for pointing out this issue. We agree that applying more efficient detection models can show better performance than the current model. A major reason for using an SSD-style model is to make a fair comparison with other models. In most multispectral pedestrian detection works, the main research topic is how to adaptively fuse the multispectral image pairs. After using SSD and Faster-RCNN as baseline detection models in earlier works, these models as baseline detection models were used in the following works, because it is not possible to know whether the reason for the improvement is the proposed fusion method or the performance of the detector itself. However, we strongly agree on the need of applying the proposed method to the latest model. For this reason, we will update the published code and maintain it continuously. According to your comment, we added an experimental result to compare the backbone model in the revised version. [***Table I.***]

TABLE I. Experiment results on KAIST dataset.

Methods	Backbone	Miss Rate(IoU = 0.5)		
		ALL	DAY	NIGHT
ACF [4]	-	47.32	42.57	56.17
Halfway Fusion [5]	VGG-16	25.75	24.88	26.59
Fusion RPN+BF [18]	VGG-16	18.29	19.57	16.27
IAF R-CNN [10]	VGG-16	15.73	14.55	18.26
IATDNN + IASS [11]	VGG-16	14.95	14.67	15.72
CIAN [19]	VGG-16	14.12	14.77	11.13
MSDS-RCNN [6]	VGG-16	11.34	10.53	12.94
AR-CNN [9]	VGG-16	9.34	9.94	8.38
MBNet [12]	ResNet-50	8.13	8.28	7.86
MLPD(Ours)	VGG-16	**7.58**	**7.95**	6.95
MLPD(Ours)	ResNet-50	7.61	8.36	**6.35**
MLPD(Ours)	ResNet-101	9.10	10.13	7.60

(Revised) Table I. *Experiment results on KAIST dataset*

[C9] some implementation details are not explained. To label the same box if it appears in each sensor modalities, we need the boxes to be paired in both sensor modalities (the same ID should be given to the same person's bounding box that appears in each sensor). I am not sure how this information is available in these datasets.

Response:

We apologize that the description of the multi-label assignment is not clear to understand. We used only the KAIST dataset which contains the paired (fully overlapped) RGB and thermal image pair with the aligned ground truth. The multi-label is determined depending on whether the ground truth is in the area where the multispectral images are overlapped, or the area where the multispectral images are not overlapped. To be specific, we defined the multi-label as a 2-dimensional vector; If the ground truth is in overlapped areas, the multi-label is defined as [1, 1], and if the ground truth is in non-overlapped area and it is only in RGB image, the multi-label is defined as [1, 0]. ([0, 1] for only in thermal image). We rewrote the explanation to avoid any confusion in the revised manuscript as follows:

[P.3, III.B. 14-30] Edited version:

"Let $\mathcal{Y} = \{ \overrightarrow{y_1}, \overrightarrow{y_2}, \overrightarrow{y_3} \}$ denote the label vector space of the RGB label vector $\overrightarrow{y_R}$ and thermal label vector $\overrightarrow{y_T}$. The label vectors are determined depending on whether the ground truth is in the area, where the multispectral images are, is overlapped or non-overlapped. More specifically, to assign the multi-label vector representing the state of the input pair, three cases of the label vector are defined as follows: 1) $\overrightarrow{y_1}$= [1, 0] 2) $\overrightarrow{y_2}$ = [0, 1] vice versa; 3) $\overrightarrow{y_3}$ = [1, 1]. Basically, the label vector is assigned as $\overrightarrow{y_1}$ or $\overrightarrow{y_2}$ when the corresponding images are unpaired after the applying semi-unpaired augmentation. Similarly, it is labeled as $\overrightarrow{y_3}$ when inputs keep the paired condition. Note that those label vectors $\overrightarrow{y_R}, \overrightarrow{y_T} \in \{ \overrightarrow{y_1}, \overrightarrow{y_2}, \overrightarrow{y_3} \}$ are used as an input state when training the proposed model. With the proposed strategy, the model can adaptively generate the feature map according to the state of the input pair, so that the model can robustly detect objects in both paired and unpaired cases."

[C10] Also in CVC-14, it is mentioned that the only RBG camera's bounding box annotations have been used. Then what is the use of one of the claimed contributions, i.e. multi-label loss, in this dataset and how, is multi-label loss trained in this case, since all the boxes would always be [1,0] label according to Eq. 1?

Response:

Thank you for raising this important issue. It is true that we did not use multi-label loss in CVC-14 experimental results. CVC-14 dataset is a popular benchmark in many multispectral pedestrian

works. AR-CNN [9] and MBNet [12] followed the protocol of Park et al.[], and used only RGB annotations for evaluation on the CVC-14 dataset. Except for multi-label loss, we have two contributions as a shared multi-fusion layer and a semi-unpaired augmentation. The purpose of this comparison is to verify whether these two methods can improve the detection performance in the paired condition (a fully-aligned image pair) compared to other baseline methods as shown in ***Table II.***

TABLE II. Experiment results on the CVC-14 dataset.

Methods		Miss rate(IoU = 0.5)		
		ALL	DAY	NIGHT
Grey + Thermal	MACF [21]	69.71	72.63	65.43
	Choi *et al.* [22]	63.34	63.39	63.99
	Halfway Fusion [21]	31.99	36.29	26.29
	Park *et al.* [21]	26.29	28.67	23.48
	AR-CNN [9]	22.1	24.7	18.1
	MBNet [12]	**21.1**	24.7	**13.5**
	MLPD† (Ours)	21.33	**24.18**	17.97

MLPD† : MLPD trained without the multi-label learning.

(Revised) Table II. *Experimental results on the CVC-14 dataset.*

[C11] Experiments may not be convincing enough to show the superiority of the suggested framework on these real datasets against the potentially stronger baselines. What about trying SOTA RGB based object detections on these datasets and ignoring the other sensor modality ?

Response:

Thank you for providing a great comment. We conducted an additional experiment in some single-modality detection models. We tested CSP[*] which is the state-of-the-art model in RGB-based pedestrian detection benchmark, YoloV5 [**] which is the latest single-stage model in object detection, and other models (e.g. YOLO v3, YOLO v4 and YOLO-ACN) in the KAIST dataset. As shown in ***Review-Table. I***, the fusion-based proposed method (MLPD) achieves better performance than all the latest models. Since the KAIST dataset has both day and night condition images, the single-modality detection model can suffer from environmental changes such as illumination changes.

Methods	Train modality	AP
MLPD	RGB+Thermal	**85.43**
CSP[*]	RGB	65.14
CSP[*]	Thermal	70.77
YOLOv5[**]	RGB	70.18
YOLOv5[**]	Thermal	75.35
YOLOv3[***]	Thermal	73.5
YOLOv4[***]	Thermal	76.9
YOLO-ACN[***]	Thermal	76.2

Review-Table. I *The result of other studies (e.g. CSP, YOLO v3, YOLO v4, and YOLO-ACN) depending on train modality with respect to AP.*

[*] W. Liu, S. Liao, W. Ren, W. Hu and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019.
[**] [Online]. Available: https://github.com/ultralytics/yolov5
[***] Y. Li, S. Li, H. Du, L. Chen, D. Zhang and Y. Li, "YOLO-ACN: Focusing on Small Target and Occluded Object Detection," in IEEE Access, vol. 8, pp. 227288-227303, 2020.

[C12] The Ablation studies are not explained well. How did the authors implement each baseline (table V)? for example, how a model without shared multi-fusion is tested? What about a baseline that naively adds a thermal image as an additional channel before inputting to the framework?

Response:

We apologize for such parts that are not clearly explained in the manuscript. To solve the problem, we have added a detailed explanation in the revised manuscript as follows:

[P.6, IV.E.] Edited version: "Although the proposed method shows significant im-provement, we would like to further understand the role of each component and how their combination works.We perform a series of ablation experiments and present the results in Table V. Baseline network is the SSD-likeHalfway fusion model and achieves 11.77% of miss-rate.The performance improves to 9.51% after only applying semi-unpaired augmentation. Then it further reaches 8.49% after adopting a multi-fusion method as modality-specific information can be preserved until the last layer.Lastly, the multi-label learning strategy is applied, and it improves performance by a large margin. From this fact,we conclude that the proposed method can encourage the model to learn more generalized and discriminative features to detect pedestrians."

In consideration of your comment, we conducted additional experiments for early fusion baseline which naively concatenates the RGB and thermal image along the channel axis. For all the experiments, we used the same settings and input size, except for specified fusion strategy. Note that we do not apply the shared fusion model to the early fusion baseline because there is no need to explicitly fuse the intermediate feature.

Review-Table. II shows that our halfway fusion-based model is more accurate than the final early fusion baseline by 0.19 miss rate. Interestingly, the proposed methods, such as semi-unpaired augmentation and multi-label learning, help to improve the performance dramatically in early fusion models. We can clearly see that the proposed methods perform really well on various fusion models and can help improve the detection accuracy by a large margin.

Fusion Method	SUA	MLL	SMF	Miss Rate(IoU = 0.5)		
				ALL	DAY	NIGHT
Early	-	-		11.21	13.41	6.54
	✓	-		8.69	9.78	6.42
	✓	✓		7.77	8.95	**5.47**
Halfway	✓	✓	✓	**7.58**	**7.95**	6.95

SUA : Semi-Unpaired Augmentation
MLL : Multi-Label Learning
SMF : Shared Multi-Fusion

Review-Table. II. *Ablation experiments of proposed methods.*

[C13]AP is the most popular metric for object detection; why it is not evaluated using AP, and the miss rate.

Response:

As you mentioned, AP is the most popular evaluation metric in object detection. Therefore, we also evaluated our model with respect to the AP score, and the result is shown in ***Review-Table III.*** However, for some reasons, the miss-rate metric suggested by Dollar et al. [15] has been more commonly used in recent RGB-based pedestrian detection studies [*, **,***].

	Backbone	Miss rate	AP
MLPD(Ours)	VGG-16	**7.58**	85.43
MLPD(Ours)	Resnet-50	7.61	**85.45**
MLPD(Ours)	Resnet-101	9.10	84.11

Review-Table III. *Qualitative result of the proposed method depending on the backbone network.*

[*] W. Liu, S. Liao, W. Ren, W. Hu and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection", in IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Jun. 2019.
[**] I. Hasan, S. Liao, J. Li, S. U. Akram and L. Shao, "Generalizable pedestrian detection: The elephant in the room", 2020. [Online]. Available: arXiv preprint arXiv:2003.08799 1.2
[***] Y. Tan, H. Yao, H. Li, X. Lu and H. Xie, "PRF-Ped: Multi-scale Pedestrian Detector with Prior-based Receptive Field," in IEEE International Conference on Pattern Recognition (ICPR), May 2021

Reviewer No.3:

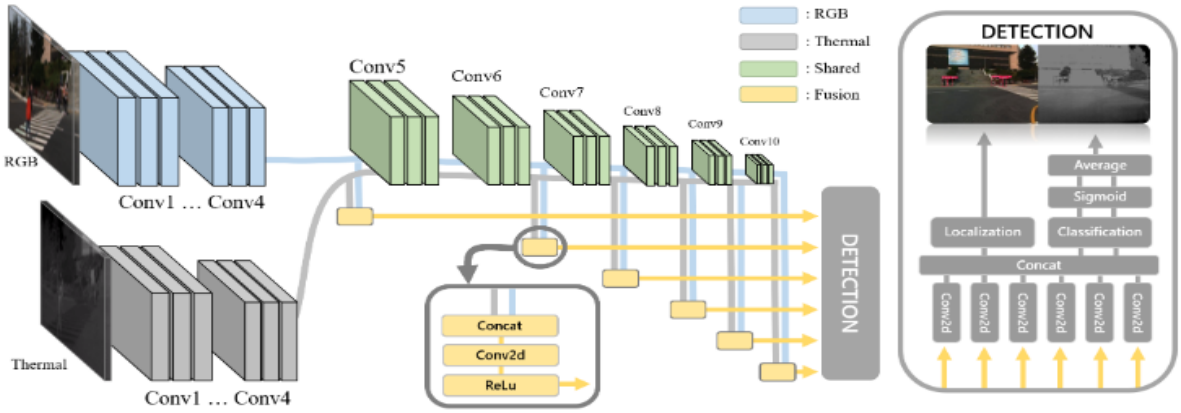
The paper presents a fusion method for RGB and thermal images to detect pedestrians. An SSD detector is adapted for fusing the two input modalities. By dividing the prediction output into only RGB, only thermal, or both, the method learns to output modality-aware predictions which leads to more robustness. It furthermore enables the network to predict on so-called unpaired modality inputs where only one of the two modalities is given. This leads to more complete coverage of detections over a scene representation (combination of RGB and thermal image).

Response: Thank you for the nice summary. We really appreciate you for taking your time to read our paper.

[C14] The fusion idea, as well as the splitting of labels, are nice ideas. I only have doubts that training splitting the labels does not really lead to a fusion network but rather, that there are two networks within one network. In other words, the network does not really fuse the inputs. Can the authors comment on that?

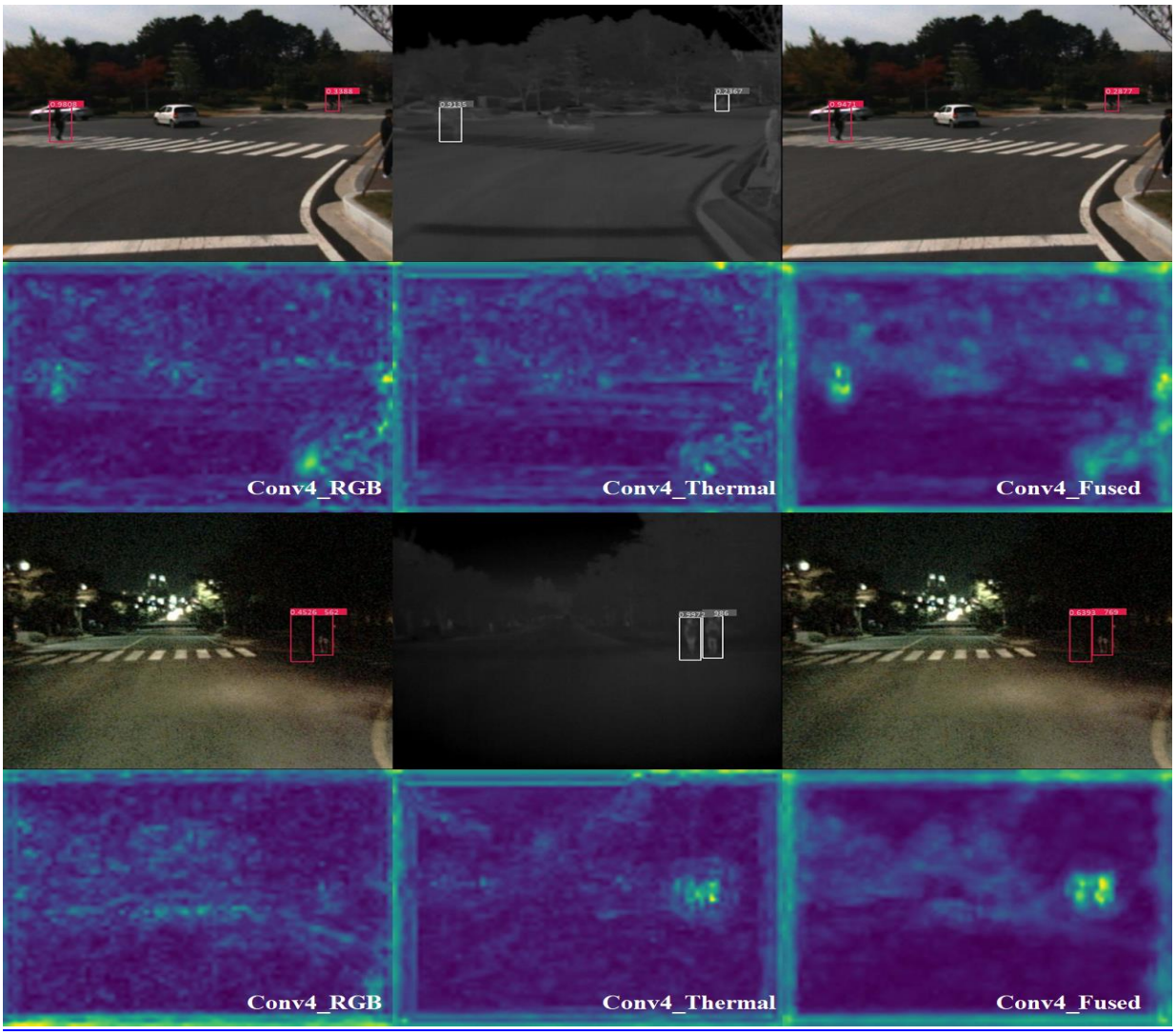
Response:

Thank you for raising this point. Instead of fusing the inputs, we fused two feature maps after passing through a shared model as shown in Fig.2 (yellow box). In the prediction stage, we averaged the multi-label confidence score to a final confidence score.



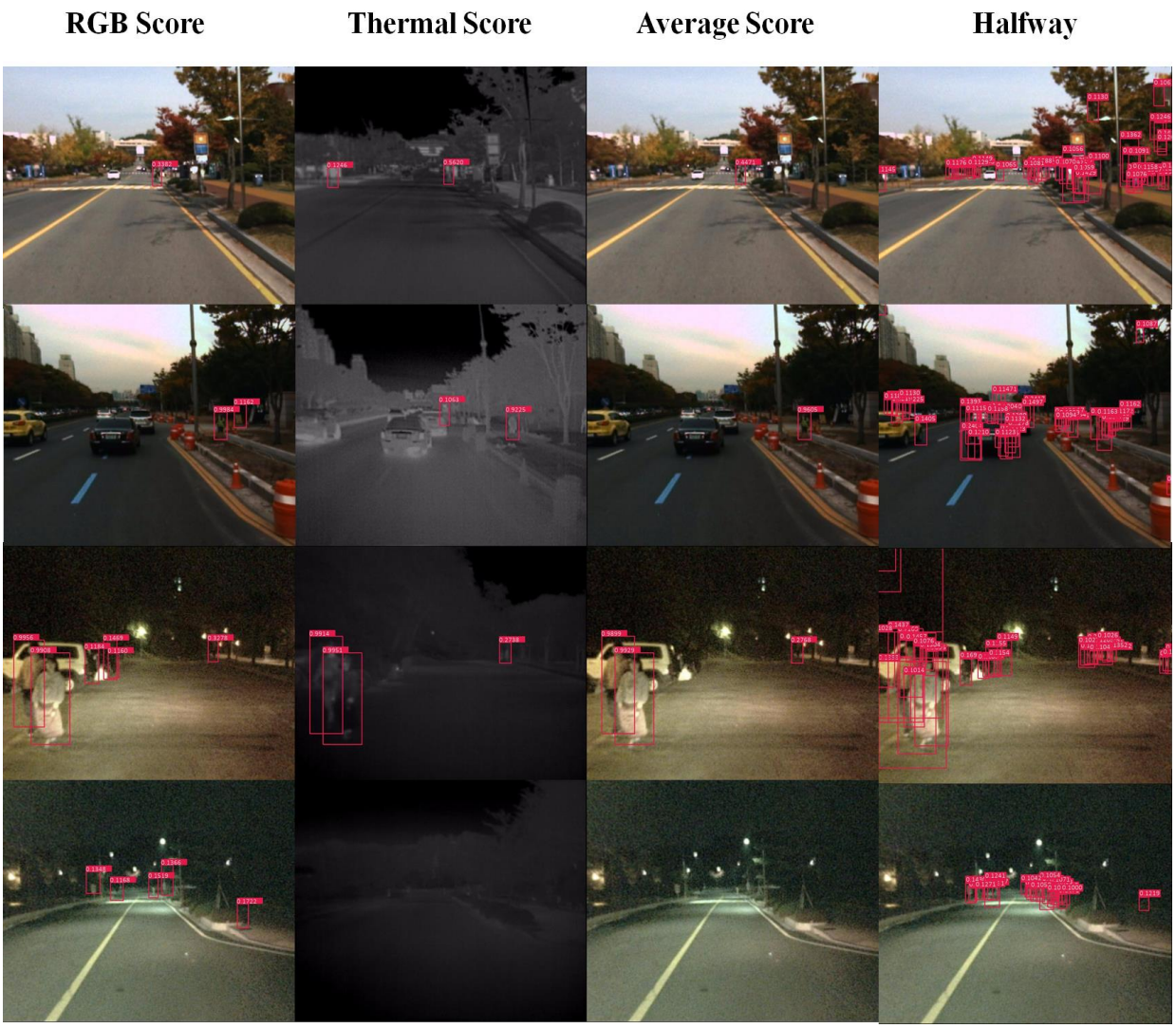
(Revised) Fig.2 Proposed architecture. *Our method is a SSD-like network which consists of two independent branches(i.e. RGB and thermal). They use independent convolutional layers before Conv5. Then they share the remaining group of convolutional layers until the end. In the Multi Fusion Module, features of each modality are concatenated. Next, other convolutional layers are used to decrease the number of channels. The outputs are fed into the detection head afterward. This architecture resembles the framework of SSD [14] but there are three main differences. 1) We adopt this architecture for multi-modality fusion; 2) We leverage multi-label learning for training; 3) We use a score function method for the final prediction.*

We argue that the proposed multi-label loss encourages the half-way fusion model to learn a more distinguishable feature map. In conventional half-way fusion models, the fused feature map is learned to estimate single label confidence. In this case, even if the discrimination of the feature map from each modality model is insufficient, it is taught in a way that makes the fused feature map become discriminative. However, since the multi-label loss enforces it to distinguish which modality is more affected in the fused feature map, the discrimination of the feature map from each modality should be discriminative, in order for the fused feature map to naturally become more discriminative. Moreover, though averaging multi-label confidence, the false positive samples can be reduced more than single label confidence. (***Review-Fig 2)***



Review-Fig 2. *This figure shows two cases of the visualized feature maps before and after the conv4 layer where feature maps from two modalities are fused, and corresponding detection results. The first and third columns represent a RGB and thermal modality, respectively. Also, the last column shows the result when two modalities are fused. The first row is the visualized detection result and the following row is the corresponding visualized feature maps. Likewise, the third and fourth rows represent the other case.*

As shown in Table.V in the paper, the multi-label loss boosts the performance. To show the validity of the argument, we first visualize the activation map before and after fusion layer in ***Review-Fig 2***. After passing through the fusion layer, the areas near pedestrians are more active and the background areas are less active. Moreover, we visualize the detection result of the baseline, SSD-like Halfway, and the proposed method. Here, we only visualize bounding boxes that have prediction scores that are greater or equal to 0.1. As shown in ***Review-Fig 3***, the proposed method estimates results with fewer false positive samples than other baseline models.



Review-Fig 3*. The first and second columns show the predicted results of the proposed method by using confidence scores in both modalities, independently. In contrast, the third column shows the predicted results of the proposed method by using an average score. Also, the last column is the result of the baseline (Halfway). The result explicitly shows that false positives tend to disappear after the score fusion is applied since it decreases the scores for false positives.*

[C15] The intro is well written and shows the motivation of the presented approach. Personally, I would already try to mention Fig. 1 already in the intro to connect the cover figure to the text (the later reference can still be there). Going over the figure I had problems at the beginning understanding what the difference between "conventional" and "commercialized" multispectral sensors is. I think these two titles do not really represent what the figure should tell the reader since it is more about paired vs. unpaired modalities. Perhaps one could think about better titles here?

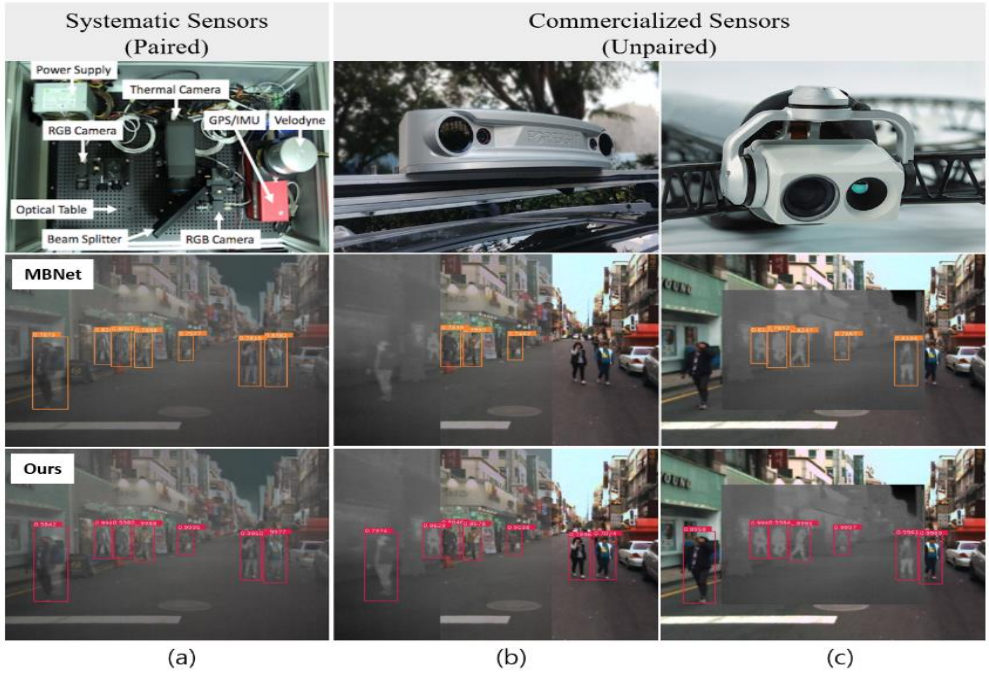
Response:

Thank you for your valuable comments. We apologize that the terminology we used may confuse readers to understand the meaning of “paired” and “unpaired”. We first restate the terminology of “paired” and “unpaired”. Our terminology “paired image” means a fully overlapped (aligned) image taken by the specially designed device such as beam splitter as shown in Fig.1-(a). Inversely, if the image pair is taken by a multispectral stereo-setting or EO/IR setting as shown in Fig.1-(b,c), there is an overlapped area (both sensor images) and a non-overlapped area (a single sensor image) in the image. Our terminology “unpaired image” means such image containing both overlapped area and non-overlapped area.

Our intention to use “conventional” means the earlier work of multispectral pedestrian dataset which provided a fully-overlapped (aligned) RGB and thermal image pair from special equipment. After introducing this multispectral dataset, there are several efforts to make the dataset practical and concise for mass production and reproducibility. Therefore, we categorized these sensor systems such as a multispectral stereo setting and EO/IR as “Commercialized”.

In the revised version, we changed “conventional” to “systematic” to reduce ambiguity for understanding our intention. Moreover, we added more explanations to make readers clearly understand what the titles stand for. We included the edited version of ***Fig.1*** as below.

Edited version:



(Revised) Fig 1. Examples of multispectral image pair according to sensor configurations *When the proposed method is compared with the state of the art, our results show the best performance for both*

paired and unpaired multispectral inputs. (a) Paired RGB-Thermal with beam splitter configuration (b) Unpaired RGB-Thermal with stereo configuration (c) Unpaired RGB-Thermal EO/IR configuration.

[C16] "This has the assumption that the input pair is almost overlapped and synchronized." - This sentence implies, that the datasets are not used because of synchronized images (meaning that the various modalities represent a scene for the same timestamp). I argue, that this is not the case since synchronization is a valid/correct assumption. Also, the presented method relies on synchronized images.?

Response:

As mentioned in [C5] and [C15], we apologize that the terminology we used may confuse readers to understand the meaning of "paired" and "unpaired". We first restate the terminology of "paired" and "unpaired". Our terminology "paired image" means a fully overlapped (aligned) image taken by the specially designed device such as beam splitter as shown in Fig.1-(a). Inversely, if the image pair is taken by a multispectral stereo-setting or EO/IR setting as shown in Fig.1-(b, c), there is an overlapped area (both sensor images) and a non-overlapped area (a single sensor image) in the image. Our terminology "unpaired image" means such image containing both overlapped area and non-overlapped area.

In both cases, there is a difference in the overlapped area. In the KAIST pedestrian dataset, the beam splitter (Fig.1-(a)) allows us to capture the image pair in a zero baseline between two sensors. However, in stereo-setting (Fig.1-(b) and Fig.1-(c)), there is a pixel-level alignment problem (a.k.a misalignment) when the two images are registered by corresponding pixels with subpixel accuracy due to parallax. This alignment problem is proportional to the baseline distance between two sensors. Also, this pixel-wise alignment problem can occur in the "unsynchronized" case where two images are not captured simultaneously.

What we wanted to convey is that most fusion approaches used a fully-overlapped (aligned) at the same timestamp (synchronized) as input images. In this paper, we tested a fully-overlapped image benchmark in both synchronized (KAIST pedestrian dataset) and unsynchronized sensors (CVC-14 dataset) respectively. Note that the CVC-14 dataset provides the overlapped image pair, but this dataset has some trouble with synchronized problems.

We apologize that our description is not clear and detailed to convey the motivation of the paper. According to your valuable comment, we revised an ambiguous description of this issue and restated the terminology, contribution, and motivation of the proposed method.

[C17] "A multi-label learning method is defined as assigning more specific labels in every object so that this strategy can encourage models to learn more sophisticated and finer features." Based on this definition and in general, I am not sure if the approach can be called multi-label learning, since there are not added more SPECIFIC labels (in the sense of more detailed parts, etc.) during training. The only rule is " that each pedestrian can be seen one of the images at least" Could the authors comment on that.

Response:

We apologize that the description of multi-label learning is not clear to understand. Generally, pedestrian detection estimates a single variable for the confidence of the pedestrian. In this paper, the proposed model estimates two variables for the confidence of the pedestrian. This is because we handle both a fully-overlapped image pair and a partially overlapped image pair (stereo-setting). In a partially overlapped image pair, some ground truths only exist in one of the multispectral image pairs.

Therefore, we defined the multi-label confidence to represent the presence in the modality. To be specific, each variable means whether the target exists in a certain image. For example, if the ground truth box exists in both RGB and thermal images, we assigned the label as [1, 1]. Inversely, if the ground truth box exists in only the RGB image, we assigned the label as [1, 0]. Since the model predicts more than a single label (confidence within [0,1]), we defined the proposed method as multi-label learning.

For reference, it is surveyed that the terminology of multi-label learning has been used in a similar concept to ours in the following paper[18]. According to your valuable comment, we revised an ambiguous description about this issue as follow:

[P.3, III.B. 14-30] Edited version:

" Let $\mathcal{Y} = \{ \overrightarrow{y_1}, \overrightarrow{y_2}, \overrightarrow{y_3} \}$ denote the label vector space of the RGB label vector $\overrightarrow{y_R}$ and thermal label vector $\overrightarrow{y_T}$. The label vectors are determined depending on whether the ground truth is in the area, where the multispectral images are, is overlapped or non-overlapped. More specifically, to assign the multi-label vector representing the state of the input pair, three cases of the label vector are defined as follows: 1) $\overrightarrow{y_1}$= [1, 0] 2) $\overrightarrow{y_2}$ = [0, 1] vice versa; 3) $\overrightarrow{y_3}$ = [1, 1]. Basically, the label vector is assigned as $\overrightarrow{y_1}$ or $\overrightarrow{y_2}$ when the corresponding images are unpaired after the applying semi-unpaired augmentation. Similarly, it is labeled as $\overrightarrow{y_3}$ when inputs keep the paired condition. Note that those label vectors $\overrightarrow{y_R}, \overrightarrow{y_T} \in \{ \overrightarrow{y_1}, \overrightarrow{y_2}, \overrightarrow{y_3} \}$ are used as an input state when training the proposed model. With the proposed strategy, the model can adaptively generate the feature map according

to the state of the input pair, so that the model can robustly detect objects in both paired and unpaired cases."

[C18] The methods part is difficult to understand at the beginning since all the variables for eq. 1 are introduced at the end of the section.

Response:

Thank you for raising this point. In the edited version of the manuscript, we changed as follow:

[P.3, III.A. 8] Original version:

$$\phi_{Fused} = f^{shr}([f_R^{spc}(I_R) \oplus f_T^{spc}(I_T)]) \quad (1)$$

"where ϕ_{Fused} refers to fused a feature map. However, we observe that input features of the detection header usually loss modality-specific information."

[P.3, III.A. 15-21] Edited version:

$$\phi_{Fused} = f^{shr}([f_R^{spc}(I_R) \oplus f_T^{spc}(I_T)]) \quad (1)$$

"where ϕ_{Fused} refers to a fused feature map. $f_R{}^{spc}$, $f_T{}^{spc}$, and f^{shr} denote the modality-specific part given RGB, thermal input images and the modality-shared part, respectively. I_R and I_T refer to corresponding images in RGB and thermal domains and $\oplus$ indicates a concatenation. We observe that input features of the detection head usually lose modality-specific information."

[C19] It is mentioned, that the authors "observe that input features of the detection header usually loss modality-specific information" following fusion of eq.1. Could this be explained, how this is observed?

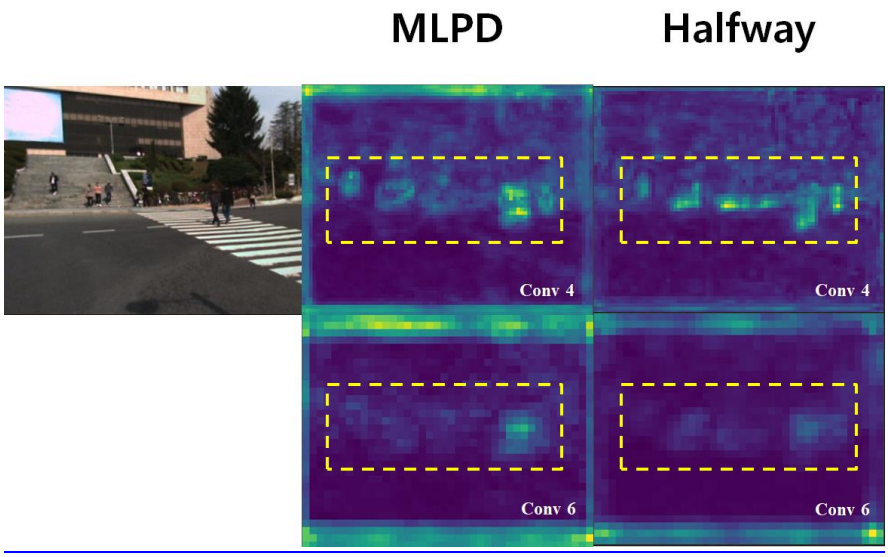
Response:

As mentioned above in [C14], the proposed multi-label loss encourages the half-way fusion model to learn a more distinguishable feature map. In the conventional half-way fusion model, the fused feature map is learned to estimate single label confidence. In this case, even if the discrimination of the feature map from each modality model is insufficient, it is taught in a way that makes the fused feature map become discriminative. However, since the multi-label loss enforces it to

distinguish which modality is more affected in the fused feature map, the discrimination of the feature map from each modality should be discriminative, in order for the fused feature map to naturally become more discriminative.

Moreover, compared to the conventional half-way fusion model, we carefully designed the shared model architecture. In the conventional half-way fusion model, after two independent models, two feature maps from each model immediately are fused and the fused feature map passes through the shared model. Accordingly, domain-specific information gradually disappears. However, in the proposed model, two feature maps from each model are not immediately fused. These features pass through the shared model independently, and then they are fused before feeding to the detection head. Therefore, the domain-specific information can be maintained.

As shown in ***Review-Fig 4***, we visualize the last feature map of the independent model and the intermediate feature map of the shared model. The fused feature map of the proposed method is more active in the area near the pedestrians because the above two approaches encourage models to learn a more discriminative feature map in independent models and to retain the domain-specific information in the fused feature map.



Review-Fig 4. *This figure shows the visualization result of the proposed method in comparison with the baseline model. It is observed that the feature map of the proposed method is more distinct and active after passing through the shared-multi fusion layer.*

[C20] "As mentioned above, this environment is not practical and it is not easy to reproduce the pixel-level alignment between multispectral images in the real-world applications." Although it is clear to me what is meant here, I would change this formulation (there are also others where the term pixel-level alignment is used). Pixel-level alignment for my understanding is the mapping of pixels from one to the other image which can be easily done by a stereo-calibration. You also need this here, since eq. 3 can only be obtained if there is a mapping/transformation between RGB and thermal images. Also the examples in Fig 1 b) and c) are aligned (to my understanding) since the RGB pixels are mapped to their corresponding thermal pixels. It is just, that the intersection of both images does not cover the full image.

Response:

The KAIST multispectral pedestrian dataset [4] was first introduced. This dataset provides fully overlapped RGB and thermal image pairs, which means that all the areas of the image are covered by both multispectral images taken at the same time. Even though most fusion methods have preferred to use such fully overlapped datasets, this kind of dataset has difficulty being used in real-world applications, because of the need for special equipment to capture both images at the zero-baseline distance as shown in Fig.1-(a).

From a practical view, the stereo setting is used as an alternative way as shown in Fig.1-(b) and Fig.1-(c). Unlike the sensor system in Fig.1-(a), this system allows a certain distance between two sensors. Thus, two issues arise, which affect the fusion method and detection performance. The first issue is that there are non-overlapped areas in the image where only information from one sensor appears. The other issue is that there is a pixel-level alignment problem known as misalignment due to parallax. This misalignment problem is proportional to the baseline distance between two sensors, and it can occur in image pairs due to the asynchronization.

From these perspectives, our intention to “As mentioned above, this environment is not practical and it is not easy to reproduce the pixel-level alignment between multispectral images in the real-world applications." is that multispectral stereo-setting as shown in Fig.1-(b) and Fig.1-(c) is a more practical approach to obtain the multispectral image pair, but this setting suffers from a misalignment problem to some extent. Also, the definition of pixel-level alignment you described in the review is correct. We apologize that our description and terminology are not clear to understand the content of the paper. In the revised manuscript, we added a more descriptive explanation of this issue and re-stated the definition of our terminology.

[C21] The semi-unpaired augmentation is a simple combination of two existing, commonly used techniques, which is why I would not claim to introduce a NEW augmentation.

Response:

Thank you for raising this important issue. As you point out, the augmentation is not new but the combination of existing augmentation techniques. However, this augmentation strategy is firstly adopted in multispectral pedestrian detection to simulate a partially overlapped image pair. Even though the component of the combination is not new, this has the novelty to do simulation via the proposed augmentation strategy.

We are sorry for making the misunderstanding, therefore we have thoroughly checked the manuscript to take out such parts that may confuse readers. We revised the paper to convey the novelty of the augmentation strategy which can efficiently simulate a partially overlapped image pair given a fully-overlapped image pair to learn the model to handle both conditions.

[P.4, III-C. 06-09] Original version:

"Therefore, we introduce a simple, yet effective data augmentation technique, called semi-unpaired augmentation."

[P.4, III-C. 06-09] Edited version:

"Therefore, we ***present*** a simple, yet effective ***way to break the pair by applying a simple data augmentation strategy***, called semi-unpaired augmentation."

[C22] In eq. 4 \phi_4 is missing.

Response:

Thank you for raising this point. In our opinion, it is probably ϕ_5, right? We excluded the ϕ_5 on purpose, and it is also shown in Fig. 2.

[C23] "We calculate the prediction score by taking the average of RGB and thermal confidence scores." Doesn't this mean, that your overall confidence of a pedestrian being visible only in RGB or thermal drops to 0.5? How is this treated in the NMS since there might be really low confidence values right? Why is the y_{cls}^{BG} not taken into account?

Response:

As you mention, the prediction score drops to 0.5 when pedestrians are only visible in RGB or thermal. However, the evaluation is conducted on a total of 9 reference points, and the lowest

reference point is 10^{-2}. Thus, the threshold is set to 0.01 in the NMS and the score drop is not critical since 0.5 is high enough. For the details of the evaluation, please refer to [C27]. Also, when reference points are higher than $10^{-0.5}(= 0.316)$, we observe that there are more positive effects than negative effects since they drop the scores of false positives as shown in ***Review-Fig 5***.

Review-Fig 5. Qualitatively result. *This figure shows the result of the proposed method with respect to bounding box scores. The first two columns show bounding boxes with scores for each domain. Then the last column shows the average score, which allows the model to remove false positives.*

[C24] Experiments: As mentioned above, I have doubts about the fusion characteristic of the network. Perhaps the authors could show that their network outperforms a baseline where two separate networks (one with RGB input and one with thermal input) generate two separate predictions which are combined at the end.

Response:

Thank you for your valuable comment. We conducted such experiments to compare the early fusion model which naively concatenates the RGB and thermal image along the channel axis, and the late fusion model which combines two separate predictions at the end. For all the experiments, we used the same settings and input size, but we do not apply all the proposed methods, such as semi-unpaired augmentation, multi-label learning, and shared fusion to the late fusion model due to the separate models.

Review-Table. IV shows that our halfway model fusion model outperforms the other two fusion methods. In a naïve setting, the performance of the early fusion model is a little bit higher than that of the halfway fusion model, but the performance of the final model is the opposite of the previous result. In the late fusion model, the performance drop occurs severely, and it seems that it is because the multispectral information cannot be reflected in the feature map of the network like two other methods. One interesting thing is that the proposed methods, such as semi-unpaired augmentation and multi-label learning, help to improve the performance dramatically in early fusion models. We can clearly see that the halfway fusion model is the best option for the current application, and the proposed methods can significantly improve the accuracy of various fusion models.

Fusion Method	SUA	MLL	SMF	Miss Rate(IoU = 0.5)		
				ALL	DAY	NIGHT
Early	-	-		11.21	13.41	6.54
	✓	✓		7.77	8.95	**5.47**
Halfway	-	-	-	11.77	13.50	8.37
	✓	✓	✓	**7.58**	**7.95**	6.95
Late				17.14	17.62	16.15

SUA : Semi-Unpaired Augmentation
MLL : Multi-Label Learning
SMF : Shared Multi-Fusion

Review-Table. IV. *Ablation experiments of proposed methods.*

[C25] "As shown in Fig 1 (c), due to the sensor configuration, this dataset has a misalignment problem that cannot guarantee the pixel-level alignment between multispectral pairs, unlike the fully-aligned multispectral pair in the KAIST dataset." Again, to my understanding, the images are aligned and it is more about the coverage of the intersection.

Response:

As you mention above, in stereo-setting (Fig.1-(b) and Fig.1-(c)), there is a pixel-level alignment problem (a.k.a misalignment) when the two images are registered by corresponding pixels with subpixel accuracy due to parallax. This alignment problem is proportional to the baseline distance between two sensors. Also, this pixel-wise alignment problem can occur in the "unsynchronized" case where two images are not captured simultaneously.

Even though the CVC-14 dataset is one of the popular benchmarks in multispectral pedestrian works, there are several issues in the CVC-14 dataset. To be specific, the CVC-14 dataset was captured by the stereo-setting as shown in Fig.1-(b), so this dataset originally should be classified as an unpaired dataset. However, the contributor of the dataset cropped the non-overlapped area in each image, so all the provided images are fully overlapped. Moreover, the contributor mentioned that this dataset has some misalignment problems due to the unsynchronized capturing way.

Therefore, our intention was that the CVC-14 dataset suffers from the above problems. We apologize that our description is not clear to convey our intention and to understand the content. In the revised version, we rewrote this sentence to remove the ambiguity for understanding the content.

[P.4, IV.A.3).] Edited version:

"The CVC-14 [17] dataset is a multispectral pedestrian dataset taken with a stereo camera configuration. The dataset is composed of Grey-Thermal pairs consisting of 7085 and 1433 frames for training and test sets, and provides individual annota-tions in each modality. Unlike the KAIST dataset in which two sensors are mechanically aligned, this dataset originally provides multispectral image pairs with non-overlapped areas and overlapped areas containing some misalignment issues. However, the author of the dataset released the cropped image pairs without the non-overlapped areas. Therefore, we treated this dataset as the fully-overlapped (paired) dataset for our purpose,but it still suffers from the pixel-level misalignment problem. Moreover, there are some issues, such as inaccu-rate ground truth boxes, incorrect extrinsic parameters, and unsynchronized capture systems. Nevertheless, this dataset has been used by many works [9], [12], [21], [22] because it is one of the few practical datasets captured in a stereo setup."

[C26] "We mimicked EO/IR (narrow-baseline) and stereo (wide- baseline) sensor configuration in (c) and (d) respectively." I think it is the other way round, right?

Response:

Thank you very much for your careful comment on our mistake. In the revised version, we restated the sentence and entirely changed the figure with a more concise caption to correct it.

[P.5, IV.A.4). 19] Edited version:

"We **mimic stereo setup and EO/IR** configuration in (c) and (d), respectively."

[C27] Perhaps one could think to show the equation for the FPPI metric

Response:

Thank you for raising this point. References seem to be missing from the section describing the evaluation metric. We will add a reference paper ***[22]*** for readers who want to know the equation and a detailed explanation of the FPPI.

[P.5, IV.A.5). 1-4] Edited version:

"We use the standard log-average miss rate (LAMR) sampled against a false positive per image (FPPI) in the range of $[10^{-2}, 10^{0}]$ **as suggested by Dollar et al[22]**."

[22] *P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 743-761, Apr 2012.*

[C28] Concerning the Evaluation on CVC-14 Dataset, if I understand that correct, then also the MLPD is trained with the bounding boxes only in RGB. It might be interesting to also show the results of the MLPD trained on both bounding box annotations (RGB and thermal) as an additional row in Table II, just for completeness.

Response:

Thank you for providing the insightful suggestion for our paper. We first explain why we provided the result of the model trained by only RGB annotations. Even though the CVC-14 dataset is one of the popular benchmarks in multispectral pedestrian works, there are several issues such as misaligned ground truth bounding boxes, incorrect extrinsic parameters in some sequences, and incorrect/missing ground truth (See the ***Revision-Fig 6***). These issues are already known to the

authors, so previous multispectral pedestrian detection works followed the protocol of Park et al.[22], and used only RGB annotations for training and evaluation.

Revision-Fig 6. Several issues in the CVC-14 dataset. *The first row images show (a) misaligned ground truth due to the parallax problem and unsynchronized images. The following row images represent (b) incorrect extrinsic parameters in some sequences. Last row images include (c) incorrect or missing ground truth (yellow dotted boxes) cases caused by mistakes of the provider.*

The purpose of this comparison is to verify whether these two methods (a multi-fusion layer and a semi-unpaired augmentation) can improve the detection performance in the paired condition (a fully-aligned image pair) compared to other baseline methods as shown in ***Table II***.

TABLE II. Experiment results on the CVC-14 dataset.

Methods		Miss rate(IoU = 0.5)		
		ALL	DAY	NIGHT
Grey + Thermal	MACF [21]	69.71	72.63	65.43
	Choi *et al.* [22]	63.34	63.39	63.99
	Halfway Fusion [21]	31.99	36.29	26.29
	Park *et al.* [21]	26.29	28.67	23.48
	AR-CNN [9]	22.1	24.7	18.1
	MBNet [12]	**21.1**	24.7	**13.5**
	MLPD† (Ours)	21.33	**24.18**	17.97

MLPD† : MLPD trained without the multi-label learning.

(Revised) Table II. *Experimental results on the CVC-14 dataset.*

Nevertheless, according to your valuable comment, we trained MLPD with bounding box annotations in both modalities. The result of this experiment is as follows [***Review-Table V***].

	Test annotation	**Miss rate**
MLPD-RGB	RGB	21.33
MLPD-Thermal	Thermal	16.47
MLPD-RGB+Thermal	**RGB**	**45.20**
MLPD-RGB+Thermal	**Thermal**	**33.84**

Review-Table V. The results of the MLPD trained on both bounding box annotations (RGB and thermal)

As shown in ***Review-Table V***, we can see that the performance of MLPD-RGB+Thermal is worse than that of MLPD-RGB and MLPD-Thermal, respectively. These results show the incorrectness of the ground truth to some extent as mentioned above, and these results can be justification for other baseline models using only RGB annotation.

[C29] In Evaluation on Unpaired Datasets the authors state: “This experiment has the meaning that most of the previous fusion methods do not handle the case where both input images are not perfectly aligned and synchronized."Correct me if I am wrong, but I thought the CVC-14 dataset does also contain image pairs where the intersection does not cover the whole image?

Response:

Thank you for the valuable feedback. As mentioned above, the CVC-14 dataset provided all the overlapped image pairs, but there are some misalignment problems due to the sensor system and unsynchronized capturing way. According to the definition of the terminology as “paired”, CVC-14 should be classified as a paired dataset as shown in ***Review-Fig 1***, but this dataset contains the pixel-level alignment problem, unlike the KAIST pedestrian dataset.

Review-Fig 1. *Camera setup for the CVC-14 dataset and registered sample frames showing a different field of view. This figure shows the difference between paired images and unpaired images. The CVC-14 dataset provides only ‘Paired’ image pairs.*

In the revised manuscript, we have revised that sentence which may confuse readers.

[P.6, IV.D. 6-7] Edited version:

“We demonstrate the robustness and generality of the proposed method in synthesized datasets. This experiment has the meaning that most of the previous fusion methods ***could not handle the case where both input images are unpaired cases (containing both overlapped and non-overlapped areas).”***

[C30] Additionally, concerning Table IV the "simulated" images with "Sides blackout" do not represent the case of a stereo set-up, since here the change of perspective between the two images is not considered.

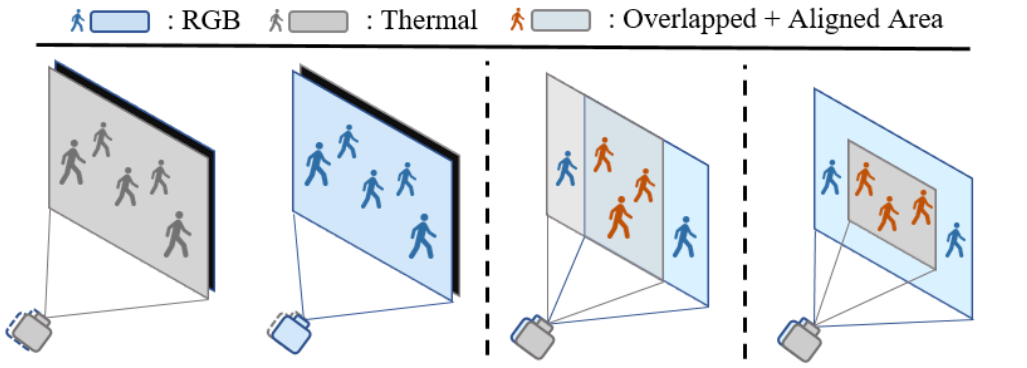
Response:

We agree with your comment. The purpose of the black simulation is to verify that the proposed model can robustly estimate the target given only a single modality image. As mentioned above in [C19], the proposed multi-loss and fusion layer has an advantage for encouraging the model to learn more deterministic domain-specific features. Moreover, we argue that this blackout image is a special case of unpaired conditions, having only non-overlapped images.

[C31] This is also wrong in Fig. 3 c). Please make sure to either make this clear, that the stereo set-up cannot be simulated like this or leave it out.

Response:

Thank you for pointing out this issue. As you mentioned above, it is not easy to simulate the real stereo image pair from the paired image in terms of the reproducibility of the parallax. Nevertheless, we tried to do our best to mimic the situation in stereo-setting with non-overlapped areas on the sides and at the boundary. Moreover, we argued that it is meaningful to conduct such unpaired simulations and experiments in terms of practical use of multispectral images in real-world applications. To clarify this, we modified Fig 3 (c) as below.



(Revised) Fig 3. Cases of non-overlapped areas *which correspond to the region where either information of sensors is missing. (a) RGB blackout; (b) Thermal blackout; (c) Sides blackout; (d) Surrounding blackout.*

[C32] In general "detection header" is normally called "detection head"

Response: Thank you for giving us detailed feedback to make it sound more professional. We took your advice and replaced “detection header” with “detection head”.

[C33] there are several sentences that are difficult to understand due to typos or wrong sentence structure.

Response: We agree with you on that and have found some typos, grammatically wrong sentences as well as wrongly structured phrases. Therefore, we have thoroughly read the manuscript multiple times to make sure all sentences are well organized and there is no mistake left.

[C34] papers/scientific reports are written in present

Response: Thank you for providing the feedback. In the edited version of the manuscript, we changed all past tense verbs to follow the rule.

[C35] incorrect usage of the term "aligned"

Response: Thank you for providing the valuable comment. We agree with you and have incorporated this feedback throughout our paper. This started from our misunderstanding of the term. To avoid any confusion, we have thoroughly checked our manuscript and either taken out or replaced the term to make it more clear to understand.

[C36] references: authors should be written instead of et al.

Response: Thank you for raising this point. We changed the entire reference to follow the IEEE style reference format.

Reviewer No.5:

This paper presents a novel approach for multispectral pedestrian detection, that is, pedestrian bounding box regression from an RGB + thermal image pair. While the presented target domain is pedestrians in traffic scenes, the proposed framework has the potential to be applied to other scenarios, as well.

In contrast to the majority of existing methods for multispectral pedestrian detection, this work lifts the assumption of a paired input image pair. Namely, it does not require pixel-level alignment nor full overlap between the input color and thermal images. To this end, the authors present three main contributions: (i) a single-stage detection framework with a novel multi-modality fusion parametrization, (ii) the introduction of multi-label learning to increase robustness in the unpaired input scenario, and (iii) a novel data augmentation technique to synthesize unpaired training data from paired images.

The proposed framework has been validated on popular pedestrian detection benchmarks, as well as on a novel synthesized dataset of unpaired color and thermal image pairs. The presented experimental evidence demonstrates the effectiveness of the approach for pedestrian detection, both in the paired and unpaired input scenarios.

Response: Thank you for the nice summary of our paper.

[C37] The article offers a comprehensive list of related works and clearly highlights the significance of the proposed contributions. At the same time, the writing of the entire manuscript could be improved, to fix both spelling and wording and increase readability.

Response: Thank you for the appreciation and the suggestion to further improve our paper. We will thoroughly review the manuscript keeping in mind your advice.

[C38] Further, a major issue of the manuscript is that many of its technical components are not clearly described or are not well formalized. It is crucial for the following comments to be addressed in order to allow accurate understanding of the proposed contributions and improve the overall presentation quality:

Response: Thank you for raising this point. We have read our paper multiple times to increase the completeness of the manuscript and tried our best to follow all valuable comments from reviewers.

[C39] The multi-label learning in Section III-B should be formalized better, particularly Equation 3. Are the labels computed in Equation 3 applied to each pedestrian bounding box? If so, this is not stated anywhere.

Response: Thank you for the suggestion and for providing such valuable feedback. In our revised manuscript, we have deleted some parts that may confuse readers and replaced them with formalized explanations including whether the equation is applied to each pedestrian bounding box. The restated sentences are as follows:

[P.3, III.B. 14-30] Edited version:

"Let $\mathcal{Y} = \{ \overrightarrow{y_1}, \overrightarrow{y_2}, \overrightarrow{y_3} \}$ denote the label vector space of the RGB label vector $\overrightarrow{y_R}$ and thermal label vector $\overrightarrow{y_T}$. The label vectors are determined depending on whether the ground truth is in the area, where the multispectral images are, is overlapped or non-overlapped. More specifically, to assign the multi-label vector representing the state of the input pair, three cases of the label vector are defined as follows: 1) $\overrightarrow{y_1}$= [1, 0] 2) $\overrightarrow{y_2}$ = [0, 1] vice versa; 3) $\overrightarrow{y_3}$ = [1, 1]. Basically, the label vector is assigned as $\overrightarrow{y_1}$ or $\overrightarrow{y_2}$ when the corresponding images are unpaired after the applying semi-unpaired augmentation. Similarly, it is labeled as $\overrightarrow{y_3}$ when inputs keep the paired condition. Note that those label vectors $\overrightarrow{y_R}, \overrightarrow{y_T} \in \{ \overrightarrow{y_1}, \overrightarrow{y_2}, \overrightarrow{y_3} \}$ are used as an input state when training the proposed model. With the proposed strategy, the model can adaptively generate the feature map according

to the state of the input pair, so that the model can robustly detect objects in both paired and unpaired cases."

[C40] The data augmentation step in Section III-C is not well formulated, namely the notation "[RGB[X], T[X]], RGB[O], T[X]]" has not been defined in relation to what the authors refer to as "modality".

Response: Thank you for raising this point. We double-checked the mentioned part and found some notations are unnecessarily used, which only deteriorates the readability. Therefore, we simply took out the notations.

[C41] Further, it is not clear how a horizontal flip of one of the two modalities can result in a sensible training image pair, as it would yield two mirrored images across the two domains.

Response: We apologize for such parts that are not clearly explained in the manuscript. It is assumed that the KAIST benchmark dataset only contains paired images, and we believe this assumption is valid as corresponding images in both modalities are automatically paired by a beam-splitter when taking them. In a paired situation, boxes are labeled as [1, 1] which we call multi-label. In the training process, the pair can be broken depending on whether the semi-unpaired augmentation has been applied. Basically, it is changed to [1, 0] and [0, 1] when semi-unpaired augmentation is applied to either of the domains, respectively. We included a detailed explanation of the concept to avoid any confusion in the revised manuscript as previously mentioned in [C39].

[C42] Finally, it is said that the data augmentation is applied to both images and bounding boxes, which misleadingly sounds like each bounding box can be flipped and cropped independently from all other bounding boxes.

Response:

Thank you for the insightful comment. We agree that the sentence would be interpreted in that way. Therefore, we rephrased the sentence as follows:

[P.4, III-C, 22-24] Original version:

"Note that we apply these techniques to both images and bounding boxes, so all boxes that are augmented are used as ground truth."

[P.4, III-C, 22-24] Edited version:

"Note that we apply the technique to both modalities independently, so all boxes that are augmented by geometric transformations are used as ground truth with the previously defined multi-label."

[C43] In Section III-D, the classification function f_cls is not well formalized, namely the text fails to describe what are the confidence scores corresponding to the background, color, and thermal and how are they computed. Further, it is not clear what is the prediction score, and how is it used. Lastly, it is not clear how the three-dimensional vector in Equation 5 relates to the two-dimensional labels in Equation 3, as described in Equation 6.

Response: Thank you for raising this point. We changed the equation in Section III-D as follows:

[P.4, III.D.6-11] Original version:

"Then we define the classification function(f_{cls}) as follows:

$$f_{cls}(\phi^*) = [\hat{y}_{cls}^{BG}, \hat{y}_{cls}^{R}, \hat{y}_{cls}^{T}] \quad (5)$$

where $\hat{y}_{cls}$refers to the confidence score of the predicted bounding box and BG, R and T refer to a background, RGB and thermal, respectively. We calculate the prediction score by taking the average of RGB and thermal confidence scores. For a multi-label classification, our network is optimized by minimizing the binary cross entropy (BCE) loss function in an end-to-end manner. It is formulated as:

$$L_{cls} = L_{BCE}(GT_{multi-label}, f_{cls}(\phi^*)) \quad (6)$$ "

[P.4, III.D.6-14] Edited version:

"Then we define $\hat{y}_R$ and $\hat{y}_T$ which refer to the confidence score vector corresponding to the predicted bounding box, and the same with thermal bounding boxes respec-tively as follows:

$$[\hat{y}_R, \hat{y}_T] = \sigma(f^{cls}(\phi^*)) \quad (4)$$

where $f^{cls}$and σ refer to the classification layer and sigmoid function, respectively. The prediction score is calculated by taking an average of RGB and thermal confidence scores corresponding to the same bounding box. For a multi-label classification, our network is optimized by minimizing the binary cross entropy (BCE) loss function in an end-to-end manner. It is formulated as:

$$\boldsymbol{L_{cls} = L_{BCE}([y_R, y_T], [\hat{y}_R, \hat{y}_T])} \quad (5)$$ "

[C44] In Section I, the first time the KAIST dataset is mentioned in the second paragraph, it should be cited.

Response: Thank you for the valuable feedback. We added a citation after first mentioning the dataset as follows:

[P.1, I. 13-14] Edited version:

" Initially, the KAIST multispectral pedestrian dataset [4] was firstly introduced."

[C45] In Section I, in the second-to-last paragraph when the SSD-like Halfway baseline is mentioned, the corresponding works should be cited.

Response: Thank you for thoroughly checking our paper to give us valuable feedback. We have added a corresponding citation as follows:

[P.2, I. 50-54] Edited version:

"By applying the proposed method to the Halfway baseline based on SSD[8], we show a significant improvement in both paired and unpaired conditions with a fast inference time."

[C46] In Fig 2, the last sentence of the caption mentions two main differences to the SSD framework, but then actually lists three.

Response: Thank you for pointing out the error. We have changed the word "two" to "three" in the caption for Fig 2 as follows:

[P.2, Fig 2. 4-5] Edited version:

"This architecture resembles the framework of SSD [8] but there are three main differences."

[C47] Equation 1 has virtually all of its terms and operators undefined

Response: Thank you for giving us the feedback. We checked the part and added some definitions of undefined terms and operators in the updated manuscript as follows:

[P.3, III.A. 15-20] Edited version:

"Where ϕ_{Fused} refers to a fused feature map. $f_R{}^{spc}$, $f_T{}^{spc}$, and f^{shr} denote the modality-specific part given RGB, thermal input images, and the modality-shared part, respectively. $I_R$and $I_T$refer to corresponding images in RGB and thermal domains and $\oplus$ indicates a concatenation. "

[C48] Equation 7 contains the undefined term L_loc.

Response: Thank you for letting us know about the error. We added the definition of each notation following your advice.

[C49] In Table II, Choi et al. is cited as [21], [22], while it only refers to [22].

Response: Thank you for pointing out the error. We took out the first reference.

[C50] The ablation experiments presented in Section V should rather be presented together with the other experimental results in Section IV.

Response: Thank you for raising the point. We created the Ablation Study subsection in section IV-E following your feedback.

[C51] Fig 5 should be referenced in text, in Section IV. The caption should state for which dataset these qualitative results are.

Response: Thank you for giving us detailed feedback to increase the completeness of our manuscript. We additionally explained which dataset the experiment was conducted with. Also, we referenced Fig 5 in Section IV to make readers understand our work better.

[C52] The entire References section must be revised to fix the formatting of all the citations to comply with the submission template.

Response: Thank you for raising this issue. We have thoroughly checked all the references to follow the IEEE style template.

MLPD: Multi-Label Pedestrian Detector in Multispectral Domain

Jiwon Kim*[1], Hyeongjun Kim*[1], Taejoo Kim*[1], Namil Kim[2] and Yukyung Choi†[1]

***Abstract*— Multispectral pedestrian detection has been actively studied as a promising multi-modality solution to handle illumination and weather changes. Most multi-modality approaches have the assumption that all inputs are fully-overlapped. However, these kinds of data pairs are not common in practical applications due to the complexity of the existing sensor configuration. In this paper, we tackle multispectral pedestrian detection, where all input data is not paired. To do that, we propose a novel single-stage detection framework which leverages multi-label learning to learn input state-aware features by assigning a separate label according to the given state of the input image pair. We also present a novel augmentation strategy by applying geometric transformations to synthesize the unpaired multispectral images. In extensive experiments, we demonstrate the efficacy of the proposed method on various real-world conditions, such as the fully-overlapped images and partially-overlapped images, in stereo-vision. Code and a demonstration video are available at https://github.com/sejong-rcv/MLPD-Multi-Label-Pedestrian-Detection.**

***Index Terms*— Multispectral pedestrian detection, unpaired images, synthetic datasets, and multi-label learning**

I. INTRODUCTION

Pedestrian detection is one of the important topics that are actively discussed in robotics and computer vision. It is a highly desired research area as it can improve the level of intelligence in various applications, such as robots, vehicles, drones, etc. Conventional RGB-based pedestrian detection is often challenged by illumination and weather changes. A substantial number of fusion methods have been developed in order to improve detection accuracy in these conditions. Among them, multispectral solutions [1], [2], [3] including thermal images are of great interest to both academic research and industry, owing to their robustness in all-day conditions.

The KAIST multispectral pedestrian dataset [4] was first introduced. This dataset provides fully-overlapped RGB and thermal image pairs, which means that all the areas of the image are covered by both multispectral images taken at the same time. Even though most fusion methods have preferred to use such fully-overlapped datasets, this kind of dataset has difficulty being used in real-world applications, because of the need for special equipment to capture both images at a zero-baseline distance as shown in Fig.1-(a).

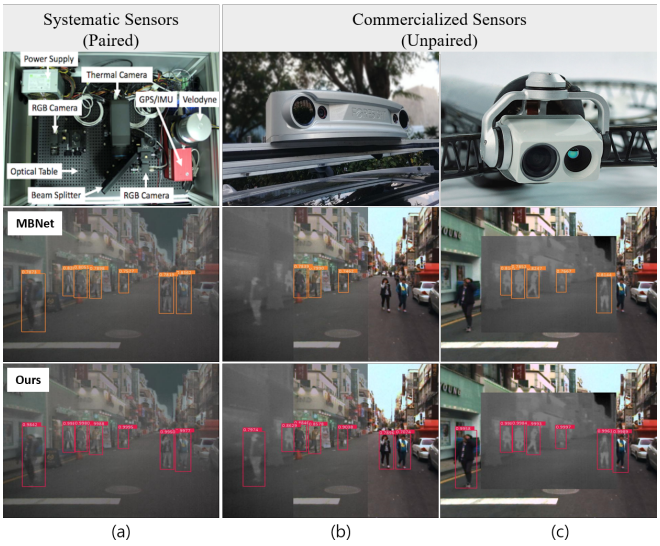


Fig. 1. **Examples of multispectral image pair according to sensor configurations** When the proposed method is compared with the state of the art, our results show the best performance for both paired and unpaired multispectral inputs. (a) Paired RGB-Thermal with beam splitter configuration (b) Unpaired RGB-Thermal with stereo configuration (c) Unpaired RGB-Thermal with EO/IR configuration.

From a practical view, the stereo setting is used as an alternative way as shown in Fig.1-(b) and Fig.1-(c). Unlike the sensor system in Fig.1-(a), this system allows a certain distance between two sensors. Thus, two issues arise, which affect the fusion method and detection performance. The first issue is that there are non-overlapped areas in the image where only information from one sensor appears. The other issue is that there is a pixel-level alignment problem known as misalignment due to parallax. This misalignment problem is proportional to the baseline distance between two sensors, and it can occur in image pairs due to the asynchronization.

In this perspective, we tackle most existing multispectral pedestrian detection methods that are mainly studied only when the multispectral images are fully-overlapped. For clarity, we first define the terminology "paired image" for a fully-overlapped image pair, and "unpaired image" for a partially-overlapped image pair including both overlapped and non-overlapped areas.

†Corresponding author

[1]Jiwon Kim, Hyeongjun Kim, Taeju Kim and Yukyung Choi are with School of Intelligent Mechatronic Engineering, Sejong University, South Korea {jwkim, hjkim, tjkim, ykchoi}@rcv.sejong.ac.kr

[2]Namil Kim is with NAVER LABS, South Korea namil.kim@naverlabs.com

Since it is not easy to obtain all the realistic partially-overlapped datasets, we focus on the general and scalable detection method to handle both overlapped and non-overlapped areas in the image by using only a fully-overlapped multispectral image pair. To do that, we introduce some novel methods and a training strategy, called multi-label learning, to learn more discriminative features, and present a semi-unpaired augmentation to stochastically generate unpaired inputs. By applying the proposed methods to the Halfway baseline based on SSD [8], we show a significant improvement in both paired and unpaired conditions with a fast inference time.

We summarize our contributions as follows: 1) We address constraints of previous fusion methods, which makes the methods hard to be applied to real-world applications and introduce a new perspective of multispectral pedestrian detection in unpaired conditions; 2) We propose a generalized multispectral pedestrian framework in ideal and practical image conditions, which is built upon multi-label learning with a novel augmentation strategy; 3) We test the proposed method on various unpaired cases, and it achieves competitive and better results compared to the state-of-the-art method.

II. RELATED WORKS

A. Multispectral Pedestrian Detection

Hwang *et al.* [4] proposed a baseline hand-crafted method and released the KAIST multispectral pedestrian benchmark with original annotations. Since the release of the KAIST dataset, a lot of multispectral pedestrian methods have been proposed for all-day vision. J. Li *et al.* [5] firstly proposed a deep learning-based fusion model with useful analyses about comparisons of various fusion architectures. Li *et al.* [6] introduced an auxiliary task which applies a semantic segmentation to the detection model and showed a better result than the detection-only model. To improve detection performance, the following works handled various topics of the detection model and the dataset itself. Zheng *et al.* [7] proposed Gated Fusion Units (GFU) to effectively aggregate multi-scale feature maps from SSD-based models and Zhang *et al.* [9] proposed a Region Feature Alignment (RFA) module which adaptively compensates a misalignment of feature maps in both modalities. In terms of imbalance problems in multispectral datasets, Li *et al.* [10] and Guan *et al.* [11] used illumination information to overcome a modality imbalance between RGB and thermal images. Zhou *et al.* [12] introduced Modality Balance Network (MBNet) that can solve modality imbalance problems in both illumination and features, simultaneously. Most previous studies have a big assumption that multispectral image pairs are fully-overlapped without non-overlapped areas. However, this condition is only possible with a systematic device, such as a beam-splitter, so we believe that this issue must be addressed in order for multispectral solutions to be applied in real-world conditions. In this paper, we tackle the unpaired multispectral pedestrian detection issue and the solution to handle unpaired inputs which commonly occur in real-world applications, such as stereo-vision.

B. Unpaired Cases for Multispectral Pedestrian Detection

To make a fully overlapped multispectral image pair is cost-expensive, so there have been some studies to address this issue. Kim *et al.* [13] assumed an extremely unpaired condition where only one single image is available and introduced an adversarial feature learning method for multispectral pedestrian detection. Even though the final model can detect the pedestrian in the single modality input, there is no further study for generality and robustness in paired and unpaired conditions. In this paper, we handle more realistic and various unpaired cases from such extreme cases like RGB and Thermal blackout as shown in Fig.3-(a) and (b) to common cases as shown in Fig.3-(c) and (d) that can come from a multispectral stereo-setting. Moreover, we demonstrate the generality of the proposed method through the fact that performance gains occur in various paired cases.

C. Multi-label Learning

A multi-label learning method is defined as assigning more specific labels in every object so that this strategy can encourage models to learn more sophisticated and finer features. Therefore, the key issue of multi-label learning is how to assign meaningful labels. With this advantage, multi-label learning has been used in detection tasks during the past few years. Zhou *et al.* [14] applied multi-label learning approach to part detectors to capture partial occlusion patterns to alleviate the heavily occluded cases and Tao Gong *et al.* [15] also reported the improvement of performance when it is applied. Most of previous works mainly focused on assigning finer labels. Unlike previous methods, we define the multi-label according to the state of multispectral pairs and we expect that a model can recognize the input state and generate the state-aware feature for pedestrian detection. This is the first attempt in multispectral fusion methods because this kind of method is not important to previous fusion methods where the fully-overlapped image pair without non-overlapped areas is given.

III. METHODS

We propose a generalized multispectral pedestrian detection framework which comprises three novel contributions, such as shared multi-fusion layers, multi-label learning, and semi-unpaired augmentation scheme. In this section, we explain the detail of each contribution.

A. Architecture

In most multispectral pedestrian detection works, the main research topic is how to adaptively fuse the multispectral image pairs. In the early years, researchers have

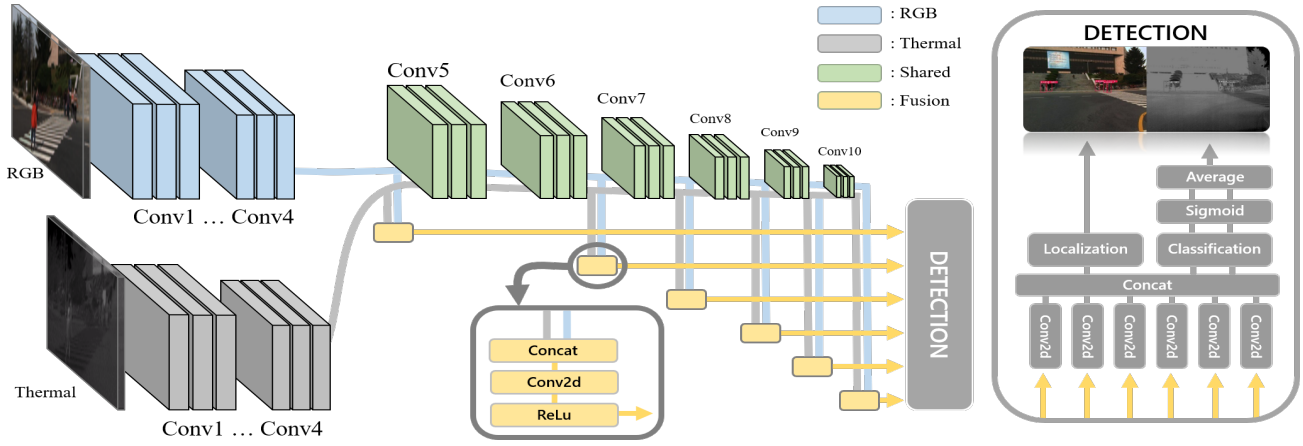


Fig. 2. **Proposed architecture.** Our method is an SSD-like network which consists of two independent branches(i.e. RGB and thermal). They use independent convolutional layers before Conv5. Then they share the remaining group of convolutional layers until the end. In the Multi Fusion Module, features of each modality are concatenated. Next, other convolutional layers are used to decrease the number of channels. The outputs are fed into the detection head afterward. This architecture resembles the framework of SSD [8] but there are three main differences. 1) We adopt this architecture for multi-modality fusion; 2) We leverage multi-label learning for training; 3) We use a score function method for the final prediction.

adopted SSD [8] and Faster-RCNN [16] as a baseline network because it is easier for them to figure out whether the reason for an improvement is the proposed fusion method or the detector itself. Thus, our detection model is also based on the SSD-like Halfway fusion model [5]. As shown in Fig 2, the model consists of the modality-specific part, modality-shared part, and detection head. In general Halfway fusion model, the feature maps, from each modality-specific part, are firstly merged and then are fed into the modality-shared part to generate input features of the detection head as follows:

$$\phi_{Fused} = f^{shr}([f_R^{spc}(I_R) \oplus f_T^{spc}(I_T)]) \quad (1)$$

where ϕ_{Fused} refers to a fused feature map. f_R^{spc}, f_T^{spc} and f^{shr} denote the modality-specific part given RGB, thermal input images, and the modality-shared part, respectively. I_R and I_T refer to corresponding images in RGB and thermal domains and ($\oplus$) indicates a concatenation.

We observe that input features of the detection head usually lose modality-specific information. We argue that the modality-shared part does not preserve information of each modality given the merged feature input. Therefore, we introduce a re-parametrization technique with the fusion layer. Instead of feeding concatenated feature map to the shared part, we feed the feature map of each modality tower, separately, and merge them before feeding them into the detection head. An interesting point is that input features of the detection head can keep modality-specific information by adding only a few fusion layers. The re-parametrization is conducted as eq (2), compared to the original formulation in eq (1).

$$\phi_{Fused} = \mathbb{F}([f^{shr}(f_R^{spc}(I_R)) \oplus f^{shr}(f_T^{spc}(I_T))]) \quad (2)$$

where $\mathbb{F}$ denotes the proposed fusion layer. We design the fusion layer as light as possible, for real-time applications, before feeding fused features into the detection head. As shown in Fig 2, the fusion layer is based on a single convolutional layer with an activation function.

B. Multi-label as the Input State

A multi-label learning method assigns more detailed class labels to encourage a model to learn more discriminative features. In most previous fusion methods, a fully-overlapped image pair is used as input images so that all objects are located and visible in both RGB and thermal domains. Moreover, these methods would fail to detect objects in situations where one of the input data has some problems, such as sensor-fault, blackout, and saturation. Therefore, we acknowledge partially-overlapped cases to handle realistic problems in the detection framework. To do that, we introduce the multi-label learning strategy in the multispectral pedestrian detection framework.

Let $\mathcal{Y} = \{\vec{y_1}, \vec{y_2}, \vec{y_3}\}$ denote the label vector space of the RGB label vector $\vec{y_R}$ and thermal label vector $\vec{y_T}$. The label vectors are determined depending on whether the ground truth is in the areas, where the multispectral images are, is overlapped or non-overlapped. More specifically, to assign the multi-label vector representing the state of the input pair, three cases of the label vector are defined as follows: 1) $\vec{y_1} = [1, 0]$, 2) $\vec{y_2} = [0, 1]$ vice versa; 3) $\vec{y_3} = [1, 1]$. Basically, the label vector is assigned as $\vec{y_1}$ or $\vec{y_2}$ when the corresponding images are unpaired after applying semi-unpaired augmentation. Similarly, it is labeled as y_3 when inputs keep the paired condition. Note that those label vectors $\vec{y_R}, \vec{y_T} \in \{\vec{y_1}, \vec{y_2}, \vec{y_3}\}$ are used as an input state when training the proposed model. With the proposed strategy, the model can adaptively generate the feature map according to the state of the input pair, so that the model can robustly detect objects in both paired and unpaired cases.

C. Semi-unpaired Augmentation

Even though unpaired cases should be handled in pedestrian detection, the problem is how to obtain realistic unpaired pairs. A naïve approach is to collect a multispectral dataset and annotate all objects in every scene. However, it is not easy to collect images from all kinds of sensor configurations. Therefore, we present a simple, yet effective way to break the pair by applying a simple data augmentation strategy, called *semi-unpaired augmentation*. As mentioned above, the major goal of the proposed method is the generality of the detection framework in both paired and unpaired conditions. That is, the model can distinguish which modality pedestrian has been affected by. To do that, we generate unpaired images from paired multispectral images. To prevent distortions in the augmented images, we only use geometric transformations, such as horizontal flip and random resized crop. More specifically, the horizontal flip is independently applied to each modality with a probability of 0.5. Similarly, the random resized crop is applied with a probability of 0.5 afterward. In other words, the augmentation technique breaks the pair with a probability of 0.75. Note that we apply the technique to both modalities independently, so all boxes that are augmented by geometric transformations are used as ground truth with the previously defined multi-label.

D. Optimization

As we mentioned, ϕ_i refers to a fused feature map that is fed into the detection head. The detection head takes multiple fused features with different resolution maps as inputs to detect pedestrians of various sizes. The concatenated feature map($\phi^\star$) is defined as follows:

$$\phi^\star = \phi_4 \oplus \phi_6 \oplus \phi_7 \oplus \phi_8 \oplus \phi_9 \oplus \phi_{10} \tag{3}$$

Then we define $\hat{y}_R$ and $\hat{y}_T$ which refer to the confidence score vector corresponding to the predicted bounding box, and the same with thermal bounding boxes respectively as follows:

$$[\hat{y}_R, \hat{y}_T] = \sigma(f^{cls}(\phi^\star)) \tag{4}$$

where f^{cls} and σ refer to the classification layer and sigmoid function, respectively. The prediction score is calculated by taking an average of RGB and thermal confidence scores corresponding to the same bounding box. For a multi-label classification, our network is optimized by minimizing the *binary cross entropy (BCE)* loss function in an end-to-end manner. It is formulated as:

$$L_{cls} = L_{BCE}([y_R, y_T], [\hat{y}_R, \hat{y}_T]) \tag{5}$$

Our loss term for localization (i.e. box regression) is the same as SSD [8]. Finally, the final loss term(L) is the weighted sum of the two loss terms as follows:

$$L = L_{loc} + \lambda L_{cls} \tag{6}$$

where λ is a weight factor to balance two loss terms, and L_{loc} and L_{cls} indicate the loss terms for localization and classification, respectively. We set λ to 1 in our experiments. Hence, this does not affect the result.

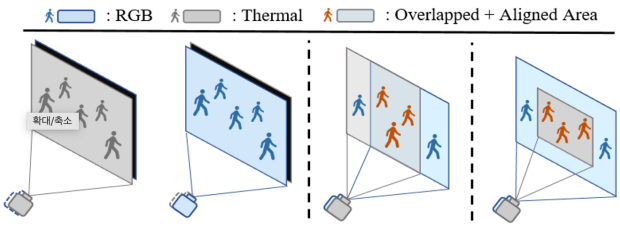
IV. EXPERIMENTS

A. Experimental Setup

1) Implementation: The baseline model is based on a modified version of SSD in PyTorch. We redesign the feature aggregation module through the proposed re-parametrization strategy and the fusion layer. Based on the knowledge that most pedestrians can be expressed by the lengthwise bounding box, we set the parameter of the anchor box as 1/1 and 1/2 for aspect ratios, $[2^0, 2^{1/3}, 2^{2/3}]$ for fine scales and 40, 80, 160, 200, 280, 360 for scales levels. We use VGG16 pre-trained on ImageNet with batch normalization, from *Conv1* to *Covn5*, and remaining convolution kernels are initialized with values drawn from the normal distribution (std=0.01). The model is trained by *Stochastic Gradient Descent(SGD)* with the initial learning rate, momentum, weight decay, as 0.0001, 0.9, and 0.0005. The mini-batch size is set to 6 and the input image size is resized to 512(H) x 640(W). We provide other hyper-parameters in the implementation.

2) KAIST Dataset: The KAIST Multispectral Pedestrian Dataset [4] consists of 95,328 *fully-overlapped* RGB-Thermal pairs in an urban environment. The provided ground truth consists of 103,128 pedestrian bounding boxes in 1,182 instances. In the experiment, we follow the standard criterion as *train02*, which samples one frame out of every 2 frames so that total 25,076 frames are used for training. For evaluation, we also follow the standard evaluation criterion as *test20*, which is sampled one out of every 20 frames, so all results are evaluated on 2,252 frames consisting of 1,455 frames in day-time and 797 frames in night-time. Note that we use the paired annotations for training [9] and the sanitized annotations for evaluation [6]. This is the standard criterion for a fair comparison with recent related works.

3) CVC-14 Dataset: The CVC-14 [17] dataset is a multispectral pedestrian dataset taken with a stereo camera configuration. The dataset is composed of Grey-Thermal pairs consisting of 7085 and 1433 frames for training and test sets, and provides individual annotations in each modality. Unlike the KAIST dataset in which two sensors are mechanically aligned, this dataset originally provides multispectral image pairs with non-overlapped areas and overlapped areas containing some misalignment issues. However, the author of the dataset released the cropped image pairs without the non-overlapped areas. Therefore, we treated this dataset as the fully-overlapped (paired) dataset for our purpose, but it still suffers from the pixel-level misalignment



(a) RGB Blackout (b) Thermal Blackout (c) Sides Blackout (d) Surrounding Blackout

Fig. 3. **Cases of non-overlapped areas** which correspond to the region where either information of sensors is missing. (a) RGB blackout; (b) Thermal blackout; (c) Sides blackout; (d) Surrounding blackout.

problem. Moreover, there are some issues, such as inaccurate ground truth boxes, incorrect extrinsic parameters, and unsynchronized capture systems. Nevertheless, this dataset has been used by many works [9], [12], [21], [22] because it is one of the few practical datasets captured in a stereo setup.

4) Synthetic Datasets for Unpaired Images: We introduce realistic synthetic datasets to demonstrate the robustness in the unpaired input (containing both overlapped and non-overlapped areas). As shown in Fig 3, non-overlapped areas are defined as the location where only a single modality can be visible. This region is natural and variable according to the relative location of each RGB and thermal sensor.

Among them, we define the most common cases of the non-overlapped areas as shown in Fig 3. Given KAIST multispectral image pairs, we generate four types of unpaired cases, such as (a) RGB blackout; (b) Thermal blackout; (c) Sides blackout, and (d) Surrounding blackout. The first two cases, (a) and (b), represent a sensor failure situation, called sensor fault where one sensor does not work at all. For example, RGB sensors have bad visibility at night and thermal sensors sometimes suffer from the crossover. To generate such cases, we fill all zero values in either RGB or thermal images stochastically. We mimic stereo setup and EO/IR configuration in (c) and (d), respectively. The (c) case can be generated by vertically dividing the original image into 3 equal-sized subsections. The subsections are applied RGB only, thermal only, and both images stochastically. Lastly, to generate (d) case, we select one of fully-aligned RGB and thermal images, crop to the smaller size, and inset the cropped image into the original counterpart. The cropped range is 96 pixels on top and bottom, and 120 pixels on the left and right side, accordingly. We believe that this synthesized image is useful to validate the robustness of the multispectral fusion model in unpaired condition and this synthesized image has little discrepancy against the real-world unpaired image because we carefully select all parameters to generate the synthetic case according to the real-world sensor configuration.

5) Evaluation Metric: We use the standard *log-average miss rate* (LAMR) sampled against a *false positive per image* (FPPI) in the range of $[10^{-2}, 10^{0}]$ as suggested by Dollar *et al.* [20] for a representative score which is the most popular metric for a pedestrian detection task. This metric only focuses on a high-precision region rather than a low-precision region so it is more appropriate for commercial solutions.

TABLE I. Experiment results on KAIST dataset.

Methods	Backbone	Miss Rate(IoU = 0.5)		
		ALL	DAY	NIGHT
ACF [4]	-	47.32	42.57	56.17
Halfway Fusion [5]	VGG-16	25.75	24.88	26.59
Fusion RPN+BF [18]	VGG-16	18.29	19.57	16.27
IAF R-CNN [10]	VGG-16	15.73	14.55	18.26
IATDNN + IASS [11]	VGG-16	14.95	14.67	15.72
CIAN [19]	VGG-16	14.12	14.77	11.13
MSDS-RCNN [6]	VGG-16	11.34	10.53	12.94
AR-CNN [9]	VGG-16	9.34	9.94	8.38
MBNet [12]	ResNet-50	8.13	8.28	7.86
MLPD(Ours)	VGG-16	**7.58**	**7.95**	6.95
MLPD(Ours)	ResNet-50	7.61	8.36	**6.35**
MLPD(Ours)	ResNet-101	9.10	10.13	7.60

TABLE II. Experiment results on the CVC-14 dataset.

Methods		Miss rate(IoU = 0.5)		
		ALL	DAY	NIGHT
Grey + Thermal	MACF [21]	69.71	72.63	65.43
	Choi *et al.* [22]	63.34	63.39	63.99
	Halfway Fusion [21]	31.99	36.29	26.29
	Park *et al.* [21]	26.29	28.67	23.48
	AR-CNN [9]	22.1	24.7	18.1
	MBNet [12]	**21.1**	24.7	**13.5**
	MLPD† (Ours)	21.33	**24.18**	17.97

MLPD† : MLPD trained without multi-label learning.

B. Evaluation on KAIST Dataset

We show the performance of pedestrian detection in a fully-overlapped and aligned conventional multispectral dataset. Since the goal of the proposed method is to improve the generality of pedestrian detection in both paired and unpaired cases, it is important to prove the superiority in paired cases. The detection performance can be found in Table I. The table clearly shows that our proposed method surpasses previous methods by a large margin. Compared to the performances of the previous state-of-the-art methods, such as MBNet [12], AR-CNN [9], we achieve not only better accuracy but also lower computational load, as shown in Fig 4. From this result, we argue that the proposed fusion model can preserve model-specific information and multi-label learning itself can help to learn more discriminative features in paired conditions.

C. Evaluation on CVC-14 Dataset

To evaluate the robustness to those conditions, we adopt this dataset and compare the result with other methods [9], [12], [21], [22]. To make a fair comparison,

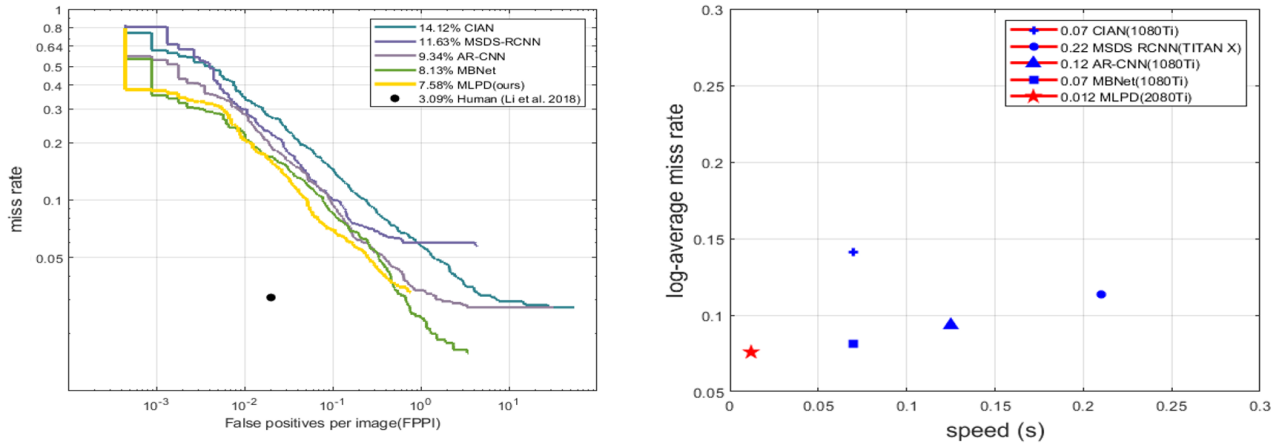


Fig. 4. **Performance** Our method outperforms the state-of-the-art method and shows an approximately 400% speed increase. Note that we excluded the time for post-processing such as Non-maximum Suppression.

we also follow the protocol introduced in [21] as other studies also adopt this. However, this protocol which only uses *bounding boxes* in grey images is contrary to the key idea of multi-label learning. Despite the fact that multi-label learning cannot be applied in the experiment, the proposed method still achieves a competitive result compared to the state-of-the-art method as shown in Table II. The result verifies that the proposed model is robust to realistic problems, which normally deteriorates the performance of models.

D. Evaluation on Unpaired Datasets

We demonstrate the robustness and generality of the proposed method in synthesized datasets. This experiment has the meaning that most of the previous fusion methods could not handle the case where both input images are unpaired cases (containing both overlapped and non-overlapped areas). However, this scenario is natural in real-world sensor applications due to the stereo-vision system and sensor resolution as shown in Fig 1. For a fair comparison with recent works, we apply the same protocol to all methods. In Table III, when the proposed method is applied, our model demonstrates a significant improvement over previous works in every case. More specifically, the proposed approach achieves 25.02% miss rate, which implies it still maintains a lower miss rate than SSD whereas other studies show 80.06% of the average miss rate on the thermal blackout dataset. Likewise, our approach outperforms all the other methods by a large margin as shown in Table IV. We also show that the proposed model can detect pedestrians in the non-overlapped areas (Fig 5), unlike other comparisons. Note that we can see that the proposed dataset is not easy and trivial from the fact that methods that perform well in a paired dataset show a large performance drop. We can conclude that the proposed method effectively generalizes a model in both paired(overlapped) and unpaired(non-overlapped) cases with stronger detection performance in all-day conditions.

TABLE III. Experiment results on the KAIST dataset regarding sensor failure

Methods	RGB	Thermal	Miss Rate(IoU = 0.5)		
			ALL	DAY	NIGHT
SSD-RGB [8]	-	Black	34.63	25.38	53.86
MSDS-RCNN [6]	-	Black	82.97	76.04	97.68
AR-CNN [9]	-	Black	77.03	67.54	97.85
MBNet [12]	-	Black	80.20	71.88	100
MLPD(Ours)	**-**	**Black**	**25.02**	**17.34**	**41.30**
SSD-thermal [8]	Black	-	21.12	25.63	12.58
MSDS-RCNN [6]	Black	-	36.36	39.53	28.67
AR-CNN [9]	Black	-	17.70	21.95	8.64
MBNet [12]	Black	-	55.56	57.49	46.81
MLPD(Ours)	**Black**	**-**	**16.88**	**20.92**	**8.25**

TABLE IV. Experiment results on the KAIST dataset when two cameras are unpaired.

Methods	Miss Rate(IoU = 0.5)		
	Sides Blackout (Thermal-RGB)	Sides Blackout (RGB-Thermal)	Surrounding Blackout
SSD-RGB [8]	51.08	63.75	34.63
SSD-Thermal [8]	59.98	38.73	55.06
MSDS-RCNN [6]	59.42	43.00	47.22
ARCNN [9]	54.59	32.18	57.58
MBNet [12]	63.81	56.65	46.99
MLPD(Ours)	**21.77**	**15.40**	**16.42**

TABLE V. Ablation experiments of proposed methods

semi-unpaired augmentation	shared multi-fusion	multi-label learning	Miss Rate(IoU = 0.5)		
			ALL	DAY	NIGHT
-	-	-	11.77	13.50	8.37
✓	-	-	9.51	9.85	8.56
✓	✓	-	8.49	9.13	7.38
✓	✓	✓	**7.58**	**7.95**	**6.95**

E. Ablation Study

Although the proposed method shows significant improvement, we would like to further understand the role of each component and how their combination works. We perform a series of ablation experiments and present the results in Table V. Baseline network is the SSD-like Halfway fusion model and achieves 11.77% of miss-rate.

The performance improves to 9.51% after only applying semi-unpaired augmentation. Then it further reaches 8.49% after adopting a multi-fusion method as modality-specific information can be preserved until the last layer. Lastly, the multi-label learning strategy is applied, and it improves performance by a large margin. From this fact, we conclude that the proposed method can encourage the model to learn more generalized and discriminative features to detect pedestrians.

V. CONCLUSIONS

In this paper, we address the problem that previous multispectral solutions and their algorithms have mostly used the fully-overlapped RGB and thermal image pair from the specially designed sensor system. In order for multispectral solutions to be widely applied in real-world applications, the algorithm should handle the unpaired image condition where multispectral image pairs contain both overlapped and non-overlapped areas of the images. To do that, we propose a generalized multispectral pedestrian detection framework that detects the pedestrian in both paired and unpaired conditions given a single model trained by only paired image sets. With three novel contributions, such as multi-label learning, semi-unpaired augmentation, and a novel fusion layer, we can successfully encourage the model to learn more generalized and discriminative features regardless of whether the image pair is overlapped or not. In the experimental section, we demonstrate that the proposed methods can improve the detection performance and handle misalignment problems to some extent. For the effectiveness in non-overlapping conditions, we simulate the realistic unpaired image set and provide extensive comparisons with other state-of-the-art models.

In future work, we will consider how to verify the unpaired condition in the real world. In the paper, we try to mimic the unpaired images given the paired images with a geometric prior, but it still has some limitations to simulate the real-world image. Moreover, since the pixel-level alignment problem in the overlapped areas is an important issue for multispectral fusion methods, we will develop the proposed method to work well on this issue.

REFERENCES

[1] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.

[2] I. Sa, Z. Chen, M. Popovic, R. Khanna, F. Liebisch, J. Nieto, and R. Siegwart, "WeedNet: Dense Semantic Weed Classification Using Multispectral Images and MAV for Smart Farming," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, 2018.

[3] Y. Yue, C. Yang, J. Zhang, M. Wen, Z. Wu, H. Zhang, and D. Wang, "Day and night collaborative dynamic mapping in unstructured environment based on multimodal sensors," in *Proceeding of the IEEE International Conference on Robotics and Automation*, 2020.

[4] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[5] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proceeding of the British Machine Vision Conference*, 2016.

[6] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proceeding of the British Machine Vision Conference*, 2018.

[7] Y. Zheng, I. H. Izzat, and S. Ziaee, "GFD-SSD: Gated fusion double SSD for multispectral pedestrian detection," *arXiv preprint*, arXiv:1903.06999, 2019.

[8] W. Liu et al., "SSD: Single shot multibox detector," in *Proceedings of the European conference on Computer Vision*, 2016.

[9] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[10] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognition*, vol. 85, 2019.

[11] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, 2019.

[12] K. Zhou, L. Chen, and X. Cao, "Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems," in *Proceedings of the European conference on Computer Vision*, 2020.

[13] M. Kim, S. Joung, K. Park, S. Kim, and K. Sohn, "Unpaired Cross-Spectral Pedestrian Detection Via Adversarial Feature Learning," in *Proceeding of the International Conference on Image Processing*, 2019.

[14] C. Zhou and J. Yuan, "Multi-label Learning of Part Detectors for Heavily Occluded Pedestrian Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[15] T. Gong, B. Liu, Q. Chu, and N. Yu, "Using multi-label classification to improve object detection," *Neurocomputing*, vol. 370, 2019.

[16] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks." In *Advances in Neural Information Processing Systems*, 2015.

[17] A. González, Z. Fang, Y. S. Salas, J. Serrat, D. Vázquez, J. Xu, and A. M. López., "Pedestrian detection at day/night time with visible and FIR cameras: A comparison," *Sensors*, vol. 16, no. 6, 2016.

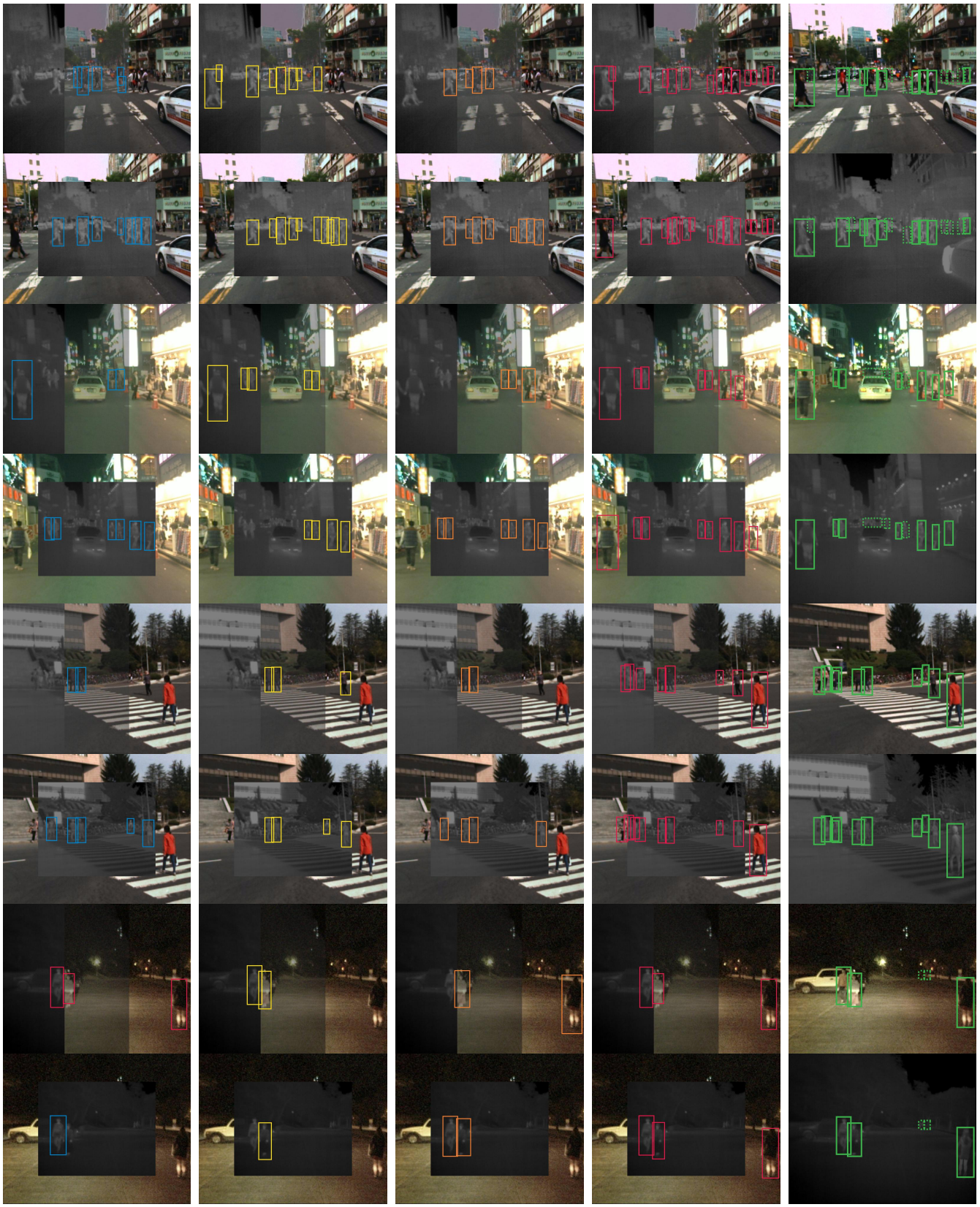
[18] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully Convolutional Region Proposal Networks for Multispectral Person Detection," in *Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[19] L. Zhang, Z. Liu, S. Zhang, X. Yang, H. Qiao, K. Huang, and A. Hussain, "Cross-modality interactive attention network for multispectral pedestrian detection," *Information Fusion*, vol. 50, 2019.

[20] P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743-761, 2012.

[21] K. Park, S. Kim, and K. Sohn, "Unified multi-spectral pedestrian detection based on probabilistic fusion networks," *Pattern Recognition*, vol. 80, pp. 143–155, 2018.

[22] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," in *Proceeding of the International Conference on Pattern Recognition*, 2016.



(a) MSDS [6] (b) ARCNN [9] (c) MBNet [12] (d) MLPD(Ours) (e) Ground Truth

Fig. 5. **Qualitative results** The qualitative results of the proposed method on the KAIST dataset. The first row shows sides blackout and the second row shows surrounding blackout and it is repeated for the remaining group of rows. The comparative results prove the robustness of our method. To follow the evaluation criterion in [4], we excluded too tiny ground truth boxes where their height is less than or equal to 55 pixel and used dotted lines to draw the excluded bounding boxes.