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 changes. Most of multi-modality approaches have the assumption that all inputs are perfectly aligned and synchronized. However, these kinds of data pairs are not common in practical applications. Especially if discrepancy between both data increases, it becomes more difficult to obtain. In this paper, we tackle multispectral pedestrian detection, where all input data is not perfectly 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. Appropriately, we introduce a novel augmentation technique, called 'semi-unpaired augmentation', by applying geometric distortion to synthesize the unpaired multispectral input images. In extensive experiments, we demonstrate efficacy of the proposed method on various real world conditions, such as the fully-overlapped images and partially overlapped images, like 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 is actively discussed in robotics and computer vision. It is a highly desired research area as it can improve the level of automation in various applications, such as robots, autonomous vehicles, drones and 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 including thermal images [1], [2], [3] are of great interest to both research and industry, owing to their robustness in all day conditions.

The KAIST multispectral pedestrian dataset is one of the popular multispectral benchmarks. This dataset provides fully aligned and synchronized RGB, as well as thermal image pairs, which corresponds to the prerequisite used in fusion studies. However, this kind of sensor configuration still has constraints to be applied in practice. It requires a specially designed hardware device such as beam-splitter, to make perfectly aligned image pairs so that it is not easy to be used in real-world applications. A practical way to collect multispectral images is to use a RGB and thermal stereo-vision system. Despite of the convenience, it can obtain only partially-overlapped image pairs according to the distance between sensors. Moreover, since the provided resolution of the thermal sensor is usually smaller than that of the RGB sensor, it is difficult to generate fully-overlapped multispectral images in this setting.

In this perspective, the successive datasets provide multispectral images in such sensor configurations. However, because of two main reasons, these datasets [4], [5] are still not widely used for fusion methods. The first reason is due to the parallax between two images, these datasets often provide ground truth of each modality separately. Another reason, is that both images from each modality are not fully-aligned, so that it is not directly used in most of fusion works. This has the assumption that the input pair is almost overlapped and synchronized.

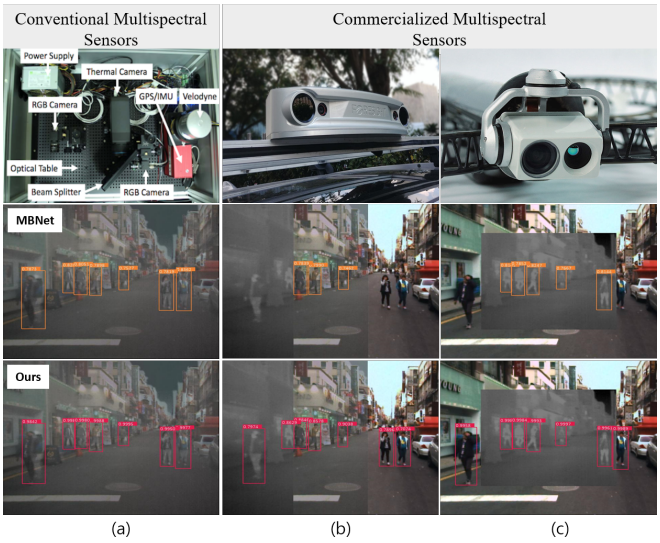

Fig. 1. **Examples of multispectral image pair according to sensor configurations** When the proposed method is compared with the state of the art, Our results show the best performance for both paired and unpaired multispectral inputs. (a) Paired RGB-Thermal with beam splitter configuration (b) Unpaired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

†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

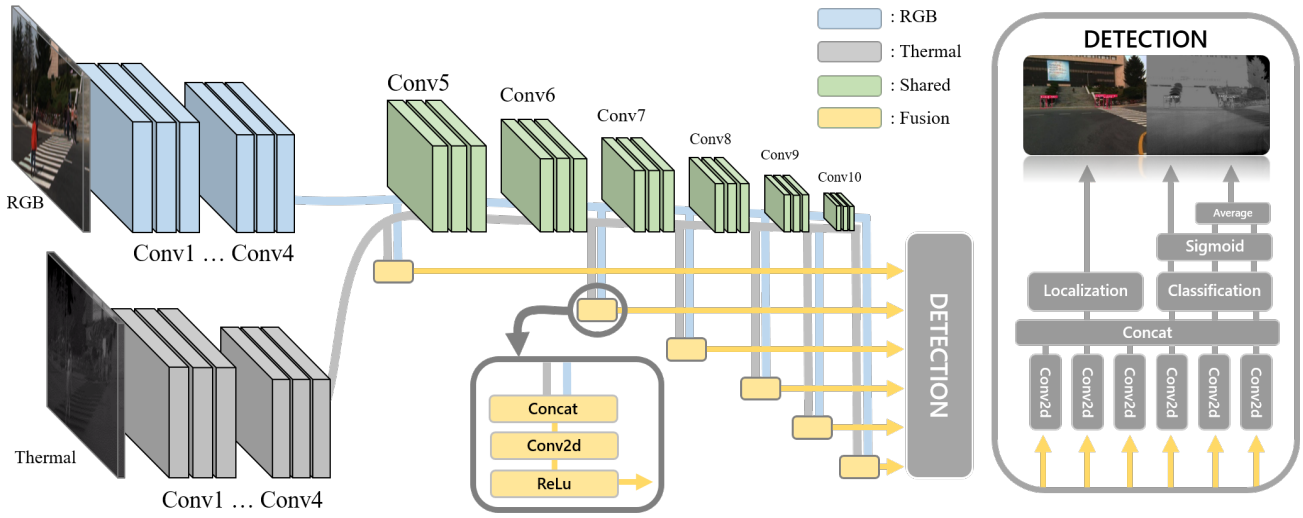

Fig. 2. **Proposed architecture.** Our method is a SSD-like network which consists of two independent branches(i.e. RGB and thermal). They use independent convolutional layers before Conv5. Then they share the remaining group of convolutional layers until the end. In Multi Fusion Module, features of each modality are concatenated. Next, other convolutional layers are used to decrease the number of channels. The outputs are fed into the detection header afterwards. This architecture resembles the framework of SSD [10] but there are two main differences. 1) We adapt 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.

In this paper, we tackle multispectral pedestrian detection, where all input data is not perfectly paired as a pixel-level alignment. Our approach is based on a single-stage object detection framework, which has an advantage in computation time. Our method employs a multi-label learning strategy to train the input state-aware model by assigning separate labels according to the given state of the input image pair. Suitably, we apply a novel augmentation technique, called semi-unpaired augmentation, to stochastically generate unpaired input pairs. By applying the proposed method to the SSD-like Halfway baseline, 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 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 technique; 3) We test the proposed method on various unpaired cases, and it achieves competitive and better results compared to state-of-the-art method.

II. RELATED WORKS

A. Multispectral Pedestrian Detection

Hwang *et al.* [6] 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. The study [7] firstly proposed a deep learning-based fusion model with useful analyses about comparisons of various fusion architectures. Li *et al.* [8] introduced an auxiliary task which applies a semantic segmentation to the detection model, and showed the 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.* [9] proposed Gated Fusion Units (GFU) to effectively aggregate multi-scale feature maps from SSD [10]-based models and Zhang *et al.* [11] 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.* [12] and Guan *et al.* [13] used illumination information to overcome a modality imbalance between RGB and thermal images. Zhou *et al.* [14] introduced Modality Balance Network (MBNet) that can solve modality imbalance problems in both illumination and features simultaneously. Most of previous studies have a big assumption that multispectral image pairs are fully-aligned or almost overlapped with little parallax. However, this condition is only possible in limited environments, such as beam-splitter, so we believe that this issue must be addressed in order for multispectral solutions to be applied in the real-world condition. In this paper, we tackle the unpaired multispectral pedestrian detection issue and the solution to handle unpaired input pairs which commonly occur in the real-world applications such as stereo-vision.

B. Unpaired Images for Multi-modal Fusion

To make fully-aligned pairs is cost-expensive and labour-intensive, and there have been many studies to alleviate this issue. Jeong *et al.* [15] presented a feature matching method to refine the coarse-level alignment

for the pixel-level accuracy. Kim *et al.* [16] pointed out the same issue in pedestrian detection tasks, where the fully-aligned image pairs are used for fusion methods. To handle this issue, this work proposed the combinatory model of embedding network, unified pedestrian detection network and adversarial network. Even though this method showed the improvement in some unpaired cases, the major drawback is the performance degradation in the paired case. In this paper, we handle more realistic unpaired cases that can come from the multispectral sensor configuration in real-world applications, and we also demonstrate the generality of the proposed method through the fact that performance gains occur in both paired and unpaired 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.* [17] applied the multi-label learning approach to part detectors to capture partial occlusion patterns to alleviate the heavily occluded cases and Tao Gong *et al.* [18] 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 method, because this kind of method is not important to previous fusion methods where the fully-aligned image pair 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 of contributions.

A. Architecture

Our detection model is based on the SSD-like halfway fusion model [7]. As shown in Figure 2, the model consists of the modality-specific part, modality-shared part and detection header. 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 header as follows:

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

where ϕ_{Fused} refers to fused a feature map. However, we observe that input features of the detection header usually loss modality-specific information. We argue that the modality-shared part does not preserve information of each modality given the merged feature input. Therefore, we introduce a re-parameterization technique with the fusion layer. Instead of feeding concatenated($\oplus$) 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 header. An interesting point is that input features of the detection header can keep modality-specific information by adding only few fusion layers. The re-parametrization is conducted as Equation (2), compared to the original formulation in Equation (1).

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

where f_R^{spc} and f_T^{spc} denote the modality-specific part given RGB and thermal input images respectively, and f^{shr} and $\mathbb{F}$ denote the modality-shared part and the proposed fusion layer, respectively. We design the fusion layer as light as possible, for real-time applications, before feeding fused features into the detection header. As shown in Figure 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 of previous fusion methods, a fully-aligned image pair was used as input images so that all the objects are located and visible in both RGB and thermal domains. 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. 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 aligned and overlapped cases to handle in the detection framework. To do that, we introduce the multi-label learning strategy in the multispectral pedestrian detection framework. We instead assign the multi-label as the state of the input pair. The basic rule is that each pedestrian can be seen one of the images at least. For example, if multispectral images are taken by a wide-baseline stereo setup, there is a non-overlapped region in each image. At the time, we expect that the model can recognize which region this pedestrian belongs to, such as the RGB-only region, thermal-only region and overlapped region. In this perspective, we define the criterion to assign the label as follows:

$$GT_{multi-label} = \begin{Bmatrix} [1,0], & \text{RGB:O, Thermal:X} \\ [0,1], & \text{RGB:X, Thermal:O} \\ [1,1], & \text{RGB:O, Thermal:O} \end{Bmatrix} \quad (3)$$

$[1,0]$ when one pedestrian is only found in a RGB domain; $[0,1]$ vise versa; $[1,1]$ when pedestrians co-exist

in both domains. 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 the unpaired case should be handled in pedestrian detection, the problem is how to obtain the realistic unpaired pair. 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 introduce a simple, yet effective data augmentation technique, 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 image pairs from fully-aligned image pairs. To prevent distortions in the augmented image, we only use geometric transformations, such as horizontal flip and random resized crop. More specifically, horizontal flip is independently applied to each modality with a probability of 0.5, such as [RGB[X],T[X]], [RGB[O],T[X]], [RGB[X],T[O]] and [RGB[O],T[O]]. Simillary, random resized crop is applied with a probability of 0.5 afterwards. In other words, the augmentation technique breaks the pair with a probability of 0.75. Note that we apply these techniques to both images and bounding boxes, so all boxes that are augmented are used as ground truth.

D. Optimization

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

$$\phi^\star = \phi_4 \oplus \phi_6 \oplus \phi_7 \oplus \phi_8 \oplus \phi_9 \oplus \phi_{10} \tag{4}$$

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

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

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:

$$L_{cls} = L_{BCE}(GT_{multi-label}, f_{cls}(\phi^\star)) \tag{6}$$

Our loss term for localization(i.e. box regression) is the same as SSD [10]. Finally, the final loss term is defined as follows:

$$L = L_{loc} + \lambda L_{cls} \tag{7}$$

where λ is a weight factor to balance two loss terms. 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-parameterization strategy and the fusion layer. Based on the knowledge that most of pedestrians can be expressed by the lengthwise bounding box, we set the parameter of the anchor box as 1/1 and 2/2 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 used 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 SGD with the initial learning rate, momentum, weight decay, as 0.01, 0.9, and 0.0005, accordingly. The mini-batch size is set to 6 and input image size is resized to 512(H) x 640(W). We provide other hyper-parameters in the implementation.

2) KAIST Benchmark Dataset: KAIST Multispectral Dataset [6] consists of total 95,328 *fully-aligned* 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 followed 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 followed 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 used the paired annotations for training [11] and the sanitized annotations for evaluation [8]. This is the standard criterion for a fair comparison with recent related works.

3) CVC-14 Dataset: The CVC-14 [4] is a multispectral pedestrian dataset taken with a wide-baseline camera configuration. The CVC-14 dataset is composed of Grey-Thermal pairs consisting of 7,085 and 1,433 frames for training and test sets. Note that the CVC-14 dataset provides individual annotations in each modality, unlike other multispectral datasets. As shown in Figure 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. The misalignment problem varies in some ways because the special device, such as beam-splitter, is needed to collect a fully-aligned dataset, and it often occurs in commercially used configurations. In this perspective,

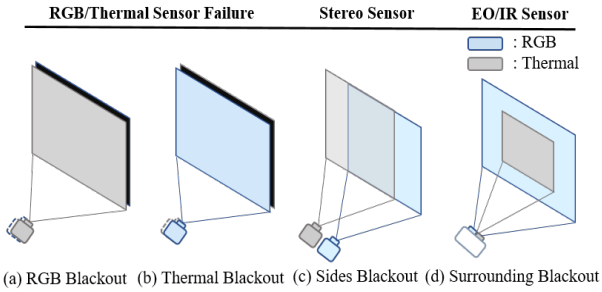

Fig. 3. **Cases of fusion dead zone (FDZ)** which is corresponding to the region where only a single modality can be visible.

we conduct the experiment in the CVC-14 dataset to prove that the proposed method can robustly detect pedestrians in real-world unpaired conditions.

4) Synthetic Datasets for Unpaired Case: To demonstrate the robustness in unpaired cases, we introduce realistic synthetic datasets given real multispectral images. We firstly define the *fusion dead zone (FDZ)* which corresponds to the region where only a single modality can be visible. This region is a natural and variable according to the relative location of each RGB and thermal sensors. Among them, we define most common cases of *fusion dead zone* as shown in Figure 3. Given KAIST multispectral image pairs, we generates 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 a bad visibility at night and suburb conditions and thermal sensors sometimes suffer from the crossover. To generate such cases, we filled all zero values in either RGB or thermal images stochastically. We mimicked EO/IR (narrow-baseline) and stereo (wide-baseline) sensor configuration in (c) and (d) respectively. CVC-14 was also taken by the sensor configuration where the resolutions of RGB and thermal cameras do not match as (d). 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 in top and bottom, and 120 pixels in 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 sampled against a *false positive per image (FPPI)* in the range of $[10^{-2}, 10^{0}]$ 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 low-precision region so it is more appropriate for commercial solutions.

TABLE I. Experiment results on KAIST dataset

Methods	Annotation	Miss Rate(IoU = 0.5)		
		ALL	DAY	NIGHT
ACF [6]	Original	47.32	42.57	56.17
Halfway Fusion [7]	Original	25.75	24.88	26.59
Fusion RPN+BF [20]	Original	18.29	19.57	16.27
IAF R-CNN [12]	Original	15.73	14.55	18.26
IATDNN + IASS [13]	Original	14.95	14.67	15.72
CIAN [19]	Original	14.12	14.77	11.13
MSDS-RCNN [8]	Original	11.34	10.53	12.94
AR-CNN [11]	Paired	9.34	9.94	8.38
MBNet [14]	Paired	8.13	8.28	7.86
MLPD(ours)	**Paired**	**7.58**	**7.95**	**6.95**

TABLE II. Experiment results on 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.* [21], [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 [11]	22.1	24.7	18.1
	MBNet [14]	**21.1**	24.7	**13.5**
	MLPD (Ours)	21.33	**24.18**	17.97

B. Evaluation on KAIST Dataset

We show the performance of pedestrian detection in conventional fully-aligned dataset. Since the goal of the proposed method is improving the generality of the 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 [14], such as MBNet, we achieve not only better accuracy but also lower computational load, as shown in Figure 4. From this result, we argue that the proposed fusion model can preserve the model-specific information and multi-label learning itself can help to learn more discriminative feature in the paired condition.

C. Evaluation on CVC-14 Dataset

Since the CVC-14 dataset has a misalignment problem, we can evaluate the robustness of the proposed method in the real-world unpaired case. For the fair comparison, we followed the protocol in [21], that uses *only bounding boxes in RGB images* as a training set. We report our results in Table II. Even with the same setting in the

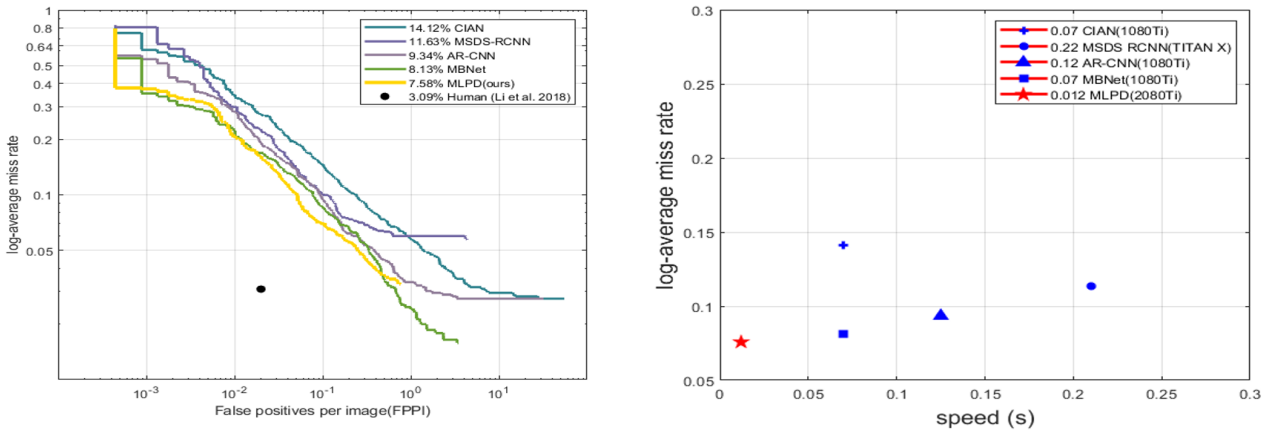

Fig. 4. **Performance** Our method outperforms the state-of-the-art method and shows an approximately 400% speed increase. Note that we excluded the time for post-processing such as Non-maximum Suppression.

KAIST dataset, the proposed method achieves competitive results compared to the state-of-the-art method. Our method outperforms other models in the day condition. At the night condition, however, our method is slightly weak. This is because other methods were developed to efficiently handle asynchronous annotations in RGB and thermal domains, but we do not tune the model for such conditions. This conclusion is that the proposed method has an efficacy to robustly detect pedestrians in the real-world misalignment condition.

D. Evaluation on Unpaired Datasets

Lastly, we demonstrate the robustness and generality of the proposed method in the synthetized dataset. This experiment has the meaning that most of previous fusion methods does not handle the case where both input images are not perfectly aligned and synchronized. However, this scenario is a natural in real-world sensor applications due to stereo-vision system, vibration, and sensor resolution as shown in Figure 1. For fair comparison with recent works, we apply the same protocol in 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 25.92% miss rate, which implies it still maintains the 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 large margin in Table IV. 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 shows a large performance drop. We can conclude that the proposed method effectively generalizes a model in both paired and unpaired cases with stronger detection performance in all day conditions.

TABLE III. Experiment results on the KAIST dataset regarding sensor failure
(i.e. RGB blackout, thermal blackout)

Methods	RGB	Thermal	Miss Rate(IoU = 0.5)		
			ALL	DAY	NIGHT
SSD-RGB [10]	-	Black	34.63	25.38	53.86
MSDS-RCNN [8]	-	Black	82.97	76.04	97.68
AR-CNN [11]	-	Black	77.03	67.54	97.85
MBNet [14]	-	Black	80.20	71.88	100
MLPD(ours)	**-**	**Black**	**25.02**	**17.34**	**41.30**
SSD-thermal [10]	Black	-	21.12	25.63	12.58
MSDS-RCNN [8]	Black	-	36.36	39.53	28.67
AR-CNN [11]	Black	-	17.70	21.95	8.64
MBNet [14]	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.
(T-R) RGB:left cutoff 30% Thermal:right cutoff 30%,
(R-T) RGB:left cutoff 30% Thermal:right cutoff 30%

Methods	Miss Rate(IoU = 0.5)		
	Sides Blackout (T-R)	Sides Blackout (R-T)	Surrounding Blackout
SSD-RGB [10]	51.08	63.75	34.63
SSD-Thermal [10]	59.98	38.73	55.06
MSDS-RCNN [8]	59.42	43.00	47.22
ARCNN [11]	54.59	32.18	57.58
MBNet [14]	63.81	56.65	46.99
MLPD(Ours)	**21.77**	**15.40**	**16.42**

TABLE V. **Ablation experiments of proposed methods**
(i.e. semi-unpaired augmentation, shared multi-fusion and multi-label learning on the KAIST Benchmark dataset.)

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

V. DISCUSSION

Although the proposed method shows significant improvement on multiple tasks, we would like to understand the role of each component and how their combination operates in practice. We perform a series of ablation experiments and present in Table V. All ablations are conducted on the KAIST dataset. Overall, all three modules improve the performance over a baseline model. We observe that the three components contribute to the accuracy in differently. One interesting point is that the augmentation only model shows the better result in the day condition, but the worse result in the night condition, while the model shows the better result in both conditions with the proposed fusion model. These results suggest that it is beneficial to use the proposed fusion model. More specifically, the augmentation can encourage the model to learn the modality-specific feature, but the baseline model cannot preserve such information to generate the fused feature map. However, the proposed fusion model can preserve such information and convey to the fused feature. Lastly, since multi-label learning strategy explicitly gives the feedback to discern the state of modality to the model, it can improve the accuracy by a large margin. From this fact, we conclude that the proposed method can encourage the model to learn more generalized and discriminative feature to detect pedestrians.

VI. CONCLUSIONS

In this paper, we addressed the problem that previous multispectral solutions and their algorithms have mostly used the fully-aligned RGB and thermal image pair from special devices great human efforts. In order for multispectral solutions to be widely applied in the real world, the algorithm should handle the unpaired condition where multispectral image pairs are partially aligned or overlapped from general sensor configurations such as stereo-vision setting, EO/IR sensors and etc. To do that, we proposed a generalized multispectral pedestrian detection framework that detects the pedestrian in both paired and unpaired conditions given the same model environment. With three novel contributions such as multi-label learning, semi-unpaired augmentation and fusion layers, we can successfully encourage the model to learn more generalized and discriminative features according to the state of the input pair. Through extensive experiments on various paired and unpaired datasets, we demonstrated the effectiveness of the proposed method and in most cases we achieved significant improvement as compared to the baseline model and state-of-the-art results.

References

[1] Ha, Qishen, *et al.*"MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes." 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017.

[2] Sa, Inkyu, *et al.*"weednet: Dense semantic weed classification using multispectral images and mav for smart farming." IEEE Robotics and Automation Letters 3.1 (2017): 588-595.

[3] Yue, Yufeng, *et al.*"Day and night collaborative dynamic mapping in unstructured environment based on multimodal sensors." 2020 IEEE international conference on robotics and automation (ICRA). IEEE, 2020

[4] González, Alejandro, *et al.*"Pedestrian detection at day/night time with visible and FIR cameras: A comparison." Sensors 16.6 (2016): 820.

[5] https://www.flir.in/oem/adas/adas-dataset-form/

[6] Hwang, Soonmin, *et al.*"Multispectral pedestrian detection: Benchmark dataset and baseline." Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). (2015).

[7] Liu, Jingjing, *et al.*"Multispectral deep neural networks for pedestrian detection." The British Machine Vision Conference (BMVC). (2016).

[8] Li, Chengyang, *et al.*"Multispectral pedestrian detection via simultaneous detection and segmentation." The British Machine Vision Conference (BMVC). (2018).

[9] Zheng, Yang, Izzat H. Izzat, and Shahrzad Ziaee. "GFD-SSD: gated fusion double SSD for multispectral pedestrian detection." arXiv preprint arXiv:1903.06999 (2019).

[10] Liu, Wei, *et al.*"Ssd: Single shot multibox detector." European Conference on Computer Vision (ECCV). (2016).

[11] Zhang, Lu, *et al.*"Weakly aligned cross-modal learning for multispectral pedestrian detection." Proceedings of the IEEE/CVF International Conference on Computer Vision. (CVPR) (2019).

[12] Li, Chengyang, *et al.*"Illumination-aware faster R-CNN for robust multispectral pedestrian detection." Pattern Recognition 85 (2019): 161-171.

[13] Guan, Dayan, *et al.*"Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection." Information Fusion 50 (2019): 148-157.

[14] Zhou, Kailai, Linsen Chen, and Xun Cao. "Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems." Proceedings of the European Conference on Computer Vision (ECCV). (2020).

[15] Jeong, Somi, *et al.*"Learning to find unpaired cross-spectral correspondences." IEEE Transactions on Image Processing 28.11 (2019): 5394-5406.

[16] Kim, Minsu, *et al.*"Unpaired Cross-Spectral Pedestrian Detection Via Adversarial Feature Learning." 2019 IEEE International Conference on Image Processing (ICIP). IEEE, (2019).

[17] Zhou, Chunluan, and Junsong Yuan. "Multi-label learning of part detectors for heavily occluded pedestrian detection." Proceedings of the IEEE International Conference on Computer Vision. (2017).

[18] Gong, Tao, *et al.*"Using multi-label classification to improve object detection." neurocomputing 370 (2019): 174-185.

[19] Zhang, Lu, *et al.*"Cross-modality interactive attention network for multispectral pedestrian detection." Information Fusion 50 (2019): 20-29.

[20] Konig, Daniel, *et al.*"Fully convolutional region proposal networks for multispectral person detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2017).

[21] Park, Kihong, Seungryong Kim, and Kwanghoon Sohn. "Unified multi-spectral pedestrian detection based on probabilistic fusion networks." Pattern Recognition 80 (2018): 143-155.

[22] Choi, Hangil, *et al.*"Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks." 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, (2016).

(a) MSDS [8] (b) ARCNN [11] (c) MBNet [14] (d) MLPD(Ours) (e) Ground Truth

Fig. 5. **Qualitative results** The qualitative results of the proposed method. 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 [6], 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.