Hybrid Attention for Robust RGB-T Pedestrian Detection in Real-World Conditions

Arunkumar Rathinam, *Member, IEEE*, Leo Pauly, Abd El Rahman Shabayek, Wassim Rharbaoui, Anis Kacem, Vincent Gaudillière, and Djamila Aouada

***Abstract*—Multispectral pedestrian detection has gained significant attention in recent years, particularly in autonomous driving applications. To address the challenges posed by adversarial illumination conditions, the combination of thermal and visible images has demonstrated its advantages. However, existing fusion methods rely on the critical assumption that the RGB-Thermal (RGB-T) image pairs are fully overlapping. These assumptions often do not hold in real-world applications, where only partial overlap between images can occur due to sensors configuration. Moreover, sensor failure can cause loss of information in one modality. In this letter, we propose a novel module called the Hybrid Attention (HA) mechanism as our main contribution to mitigate performance degradation caused by partial overlap and sensor failure, i.e. when at least part of the scene is acquired by only one sensor. We propose an improved RGB-T fusion algorithm, robust against partial overlap and sensor failure encountered during inference in real-world applications. We also leverage a mobile-friendly backbone to cope with resource constraints in embedded systems. We conducted experiments by simulating various partial overlap and sensor failure scenarios to evaluate the performance of our proposed method. The results demonstrate that our approach outperforms state-of-the-art methods, showcasing its superiority in handling real-world challenges.**

***Index Terms*—Deep learning for visual perception, multi-modal perception for HRI, sensor fusion, human detection and tracking, computer vision for transportation.**

I. Introduction

PEDESTRIAN detection is one of the important domains within computer vision for robotics, playing a significant role in applications such as self-driving vehicles, surveillance automation, and mobile robot navigation [1]. RGB cameras are commonly preferred sensors for such applications. However, they tend to suffer from overexposure in daylight, low illumination in night scenarios, and high-contrast lighting. To address these shortcomings, a number of sensors and fusion solutions were investigated. In particular, thermal cameras seem to provide several advantages in terms of costs, algorithms, and data [2]. Among them, RGB images provide texture and color information, while thermal images focus on the infrared heat emitted by the objects and are therefore invariant to lighting conditions [3]. RGB and thermal images are thefore complementary with each other by nature. This led the community to collect multispectral datasets such as KAIST [1], CVC [4] or FLIR [5], providing thermal data in addition to RGB data.

KAIST dataset provides fully-overlapping RGB-T image pairs, i.e. both images are acquired at the same time and cover the same field of view. However, acquiring such image pairs requires specialised sensor setup over conventional stereo setup which is widely used in real-world applications. In stereo setups, partial overlap will occur inherently due to a different camera Field of View (FoV) and pixel-level misalignment will occur due to parallax [6].

Information discrepancy between one image and the other can cause features to be out of their corresponding positions, resulting in decreased algorithm performance and less accurate predictions during the inference process [7]. Even in the KAIST dataset that has fully overlapping image pairs, the authors attempted to reduce the pixel-level misalignment problem. This was achieved by further improving the original data labels to *"sanitised"* cross-modal annotations [8] and *"paired"* modality-specific annotations [7] (in this letter, different *modalities* correspond to different spectral images: RGB or thermal).

Recent methods, such as multi-label learning [6], aimed to learn more discriminative features while using semi-unpaired augmentation to generate unpaired inputs between two modalities where considering a single bounding-box label is irrelevant. Even with different learning approaches, the robustness of an algorithm is questioned when one of the modalities is unavailable or partially available; for example, when a malfunction in one of the camera sensor arrays leads to a partial or even complete loss of one modality. This situation is investigated in very few existing literature [6], and the performance drop in such scenarios is quite high even for best performing models on fully overlapping images.

To improve performance and enhance algorithm robustness, our research examines the issue of partial overlap caused by various factors such as stereo configurations, sensor malfunctions, and others. This can result in partial or complete invisibility of

Received 27 May 2024; accepted 23 October 2024. Date of publication 21 November 2024; date of current version 4 December 2024. This article was recommended for publication by Associate Editor Shrey Pareek and Editor Gentiane Venture upon evaluation of the reviewers' comments. This work was supported by Luxembourg National Research Fund (FNR) through the Project BRIDGES2020/IS/14755859/MEET-A/Aouada. *(Arunkumar Rathinam and Leo Pauly contributed equally to this work.) (Corresponding author: Arunkumar Rathinam.)*

Arunkumar Rathinam, Abd El Rahman Shabayek, Anis Kacem, and Djamila Aouada are with the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, L- 1855 Kirchberg, Luxembourg (e-mail: arunkumar.rathinam@uni.lu).

Leo Pauly was with the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg.

Wassim Rharbaoui was with the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg. He is now with the XLIM institute, University of Poitiers, F- 87060 Limoges, France.

Vincent Gaudillière was with the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg. He is now with the Université de Lorraine, CNRS, Inria, Loria, F- 54000 Nancy, France.

Code is available at https://cvi2.uni.lu/ha-mlpd/.

Digital Object Identifier 10.1109/LRA.2024.3504296

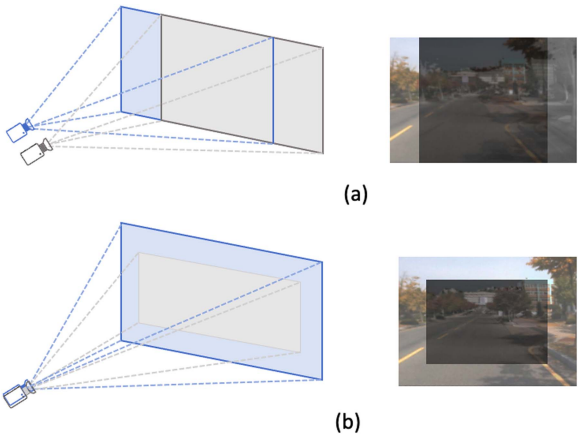

Fig. 1. Constraints on (a) stereo setup (extrinsic parameters) and (b) sensor resolution (intrinsic parameters) resulting in only partial overlap between RGB and thermal images.

regions in one modality during inference. In this letter, we adopt the term *blackout* to refer to areas in the union of pictures where data are absent from one of the modalities. Fig. 1 shows sample cases of *blackout* that can arise from sensor setups, Fig. 1(a) showing *sides blackout* from different camera extrinsics in stereo setup, Fig. 1(b) depicting *surrounding blackout* arising due to difference in camera intrinsics such as sensor resolution or focal length. To achieve robustness in such scenarios, we present a hybrid attention module, which reduces performance degradation irrespective of network architecture. We also consider a much lighter backbone compared to previous works for coping with hardware resource constraints in embedded systems. Our contributions are as follows:

- We introduce the Hybrid-Attention (HA) module, which combines self-attention and cross-attention, to mitigate performance degradation arising from modality-specific blackouts;
- We propose an improved RGB-T fusion algorithm, named Hybrid Attention-based Multi-Label Pedestrian Detector (HA-MLPD), robust against partial overlap and sensor failure encountered in real-world scenarios, while being resource-friendly;
- We provide experimental evidence that the proposed method prevents a performance drop and makes the fusion algorithm more robust and reliable irrespective of the network backbone architecture.

The remainder of the letter is structured as follows: Section II provides an overview of the related literature. In Section III, the HA-MLPD algorithm is introduced, with a detailed explanation of its implementation. Our approach is tested on the KAIST dataset [1] under various simulated blackout conditions, as discussed in Section IV. Finally, Section V concludes with a discussion on future research directions.

II. Related Works

The multispectral pedestrian detection problem has been widely studied in computer vision. Several classical methods use RGB and thermal images, relying on pixel difference values [9], local shape features [10], contour saliency maps [11], disparity maps [2] and HOG features [12].

In 2015, Hwang et al. introduced KAIST [1], a large-scale multispectral pedestrian detection dataset that contains RGB and thermal images with the corresponding pedestrian labels. The release of the KAIST dataset accompanied a renewed interest in the multispectral pedestrian detection problem, and several new methods have been proposed since then. For example, Liu et al. [13] proposed a deep learning-based Halfway Fusion model and presented comparative analyses with other early fusion architectures (input-level fusion) and late fusion architectures (decision-level fusion). In another study, Li et al. [8] demonstrated that the incorporation of an additional semantic segmentation task led to enhanced performance compared to the use of a model focused solely on detection. They introduced a combined architecture that included a multispectral proposal network to generate pedestrian proposals and a subsequent multispectral classification network to distinguish pedestrian instances from challenging negatives. The authors trained the integrated network by simultaneously optimising both pedestrian detection and semantic segmentation tasks. Zheng et al. [14] introduced Gated Fusion Units (GFU) which are designed to merge feature maps from the feature extraction layers of Single Shot MultiBox Detector (SSD) at various scales [15]. In their studies [16], [17], the authors explored the use of distinct subnetworks for individual modalities and incorporated illumination-adaptive weighting of these subnetworks to enhance the efficiency of multispectral pedestrian detection. This approach enabled the prioritization of information from the RGB modality in adequately illuminated images or from the thermal modality in low-light situations. Zhou et al. [18] introduced Modality Balance Network (MBNet), which simultaneously compensated for modality imbalance problems in illumination and at the feature levels. Chen et al. [19] presented a late fusion architecture by probabilistically ensembling decisions made individually from RGB and thermal images. Zhang et al. [20] presented a fusion mechanism under the guidance of the intermodal and intramodal attention modules, to learn to dynamically weigh and fuse multispectral features. Yang et al. [21] proposed an algorithm that uses cascaded information enhancement and fusion of cross-modal attention features, both of which rely on the attention mechanism.

The existing research has significantly enhanced the effectiveness of combining RGB and thermal images for pedestrian detection through multispectral fusion. Nonetheless, the algorithms typically operate under the assumption that both types of images are accessible during inference, neglecting scenarios where sensor failure may occur. Furthermore, potential sensor modifications during inference such as modifications in the stereo arrangement or lens settings, compared to the training dataset acquisition, are not taken into account. Such discrepancies may result in only partially-overlapping images leading to *blackout* regions (refer to Fig. 1). Analysis of existing algorithms shows a considerable degradation in performance under such conditions [6]. However, robustness to such realistic inference-time conditions is important for real-world deployment of algorithms and has not received much attention in the field. In this direction, Zhang et al. [7] proposed a region feature alignment module (RFA), which adaptively compensates for misalignment of feature maps in both modalities. Recently, Kim et al. [6] introduced the MLPD algorithm, which has shown robustness to sensor failures and inference-time partial overlap.

However, MLPD lacks interconnection between the feature extraction branches of the RGB and the thermal modalities.

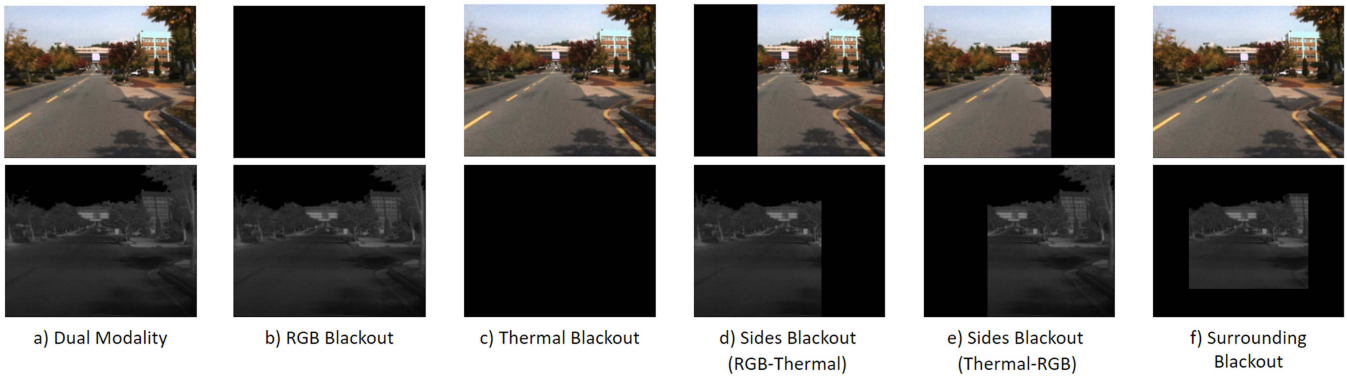

Fig. 2. Simulated inference-time conditions for assessing the robustness of HA-MLPD: (a) fully-overlapping RGB and thermal images; (b), (c) complete blackouts of one of the modalities (sensor failures); (d), (e), (f) partial overlaps (discrepancies in the extrinsic (d), (e) or intrinsic (f) camera parameters).

This could affect performance during blackout scenarios. Similarly, [22] used Multispectral Recalling (MSR) memory that can recall the missing features of multispectral modalities to detect pedestrians when one of the modalities is available. However, this architecture is not designed to handle both available simultaneous modalities.

Beyond the domain of multispectral pedestrian detection, the issue of partial image overlap has also been investigated in related fields, such as multi-view person tracking from multiple cameras [23]. In this context, proposed solutions range from hand-engineered methods, such as geometric transformations of images based on predefined ground plane homographies [24], [25], to more advanced deep learning-based approaches [26]. However, these methods require substantial adaptation to address the specific challenges posed by multispectral pedestrian detection, particularly due to the multimodal nature of the data and differences in problem formulation.

In this work, we propose an RGB and thermal fusion algorithm for pedestrian detection using the novel HA module, providing robustness to blackout scenarios caused by inference-time sensor failure or partial overlap. Our proposed HA module facilitates the flow of information and interconnects the features extracted between the modalities using cross-attention [28], [29] (during normal conditions) and self-attention [30], [31] (during blackouts).

III. Proposed HA-MLPD

A. Problem Statement

Our method, HA-MLPD, assumes that images from both modalities are automatically registered at test time and that pixel values in resulting non-overlapping regions are set to zero (i.e. blackout). Fig. 2 shows some examples. The masks of the overlapping regions are further leveraged to guide the network through cross- or self-attention to the features. For that, we use masks M_{rgb}, $M_{thermal}$ corresponding to blackout regions as what can be obtained by registering the images from the two modalities (using methods such as [32] for instance, or directly using the likely known stereo parameters). Note that the registration process is not within the scope of this letter. In detail, the masks M_{rgb} and $M_{thermal}$ are set to 1 modality-specific information is available at the location of the pixel, 0 otherwise.

TABLE I
MobileNetV2 [27] Architecture Description

Input	Operator	t	c	n	s	Blocks
$224^2 \times 3$	conv2d	-	32	1	2	
$112^2 \times 32$	bottleneck	1	16	1	1	B1
$112^2 \times 16$	bottleneck	6	24	2	2	
$56^2 \times 24$	bottleneck	6	32	3	2	
$28^2 \times 32$	bottleneck	6	64	4	2	B2
$14^2 \times 64$	bottleneck	6	96	3	1	
$14^2 \times 96$	bottleneck	6	160	3	2	B3
$7^2 \times 160$	bottleneck	6	320	1	1	
$7^2 \times 320$	conv2d 1x1	-	1280	1	1	N/A
$7^2 \times 1280$	avgpool 7x7	-	-	1	-	
$1 \times 1 \times 1280$	conv2d 1x1	-	-	-		-

Each line describes a sequence of 1 or more identical (modulo stride) layers, repeated n times. All layers in the same sequence have the same number c of output channels. The first layer of each sequence has a stride s and all others use stride 1; t denotes the expansion factor [28].

B. HA-MLPD Overview

Our model, HA-MLPD, consists of the feature extraction layers, HA module, fusion layer, and detection head. Fig. 3 presents the general network architecture for the proposed HA-MLPD algorithm with the MobilenetV2 [27] backbone. Two separate branches are used to extract feature maps from both RGB and thermal images. The HA module, placed after the first block of layers, is used to perform cross- or self-attention accross the two modalities. The attention mechanism varies according to the overlapping and non-overlapping image regions (as illustrated in Fig. 4). The extracted features from different levels are then concatenated and passed through a fusion layer to generate features used as input to the detection head. The fusion layer, designed to be as light as possible for real-time applications, is composed of one convolutional layer followed by batch normalization and ReLu activation. This is similar to MLPD, where this multi-level fusion strategy and fusion layer were introduced to cope with the loss of modality-specific information afer shared convolutional blocks. To strengthen this effect, we decoupled the modality-shared layers of MLPD to make them modality-specific in HA-MLPD.

C. Details of HA-MLPD Pipeline

Feature Extraction: Any standard feature extraction backbone, possibly followed by additional blocks of layers, can be

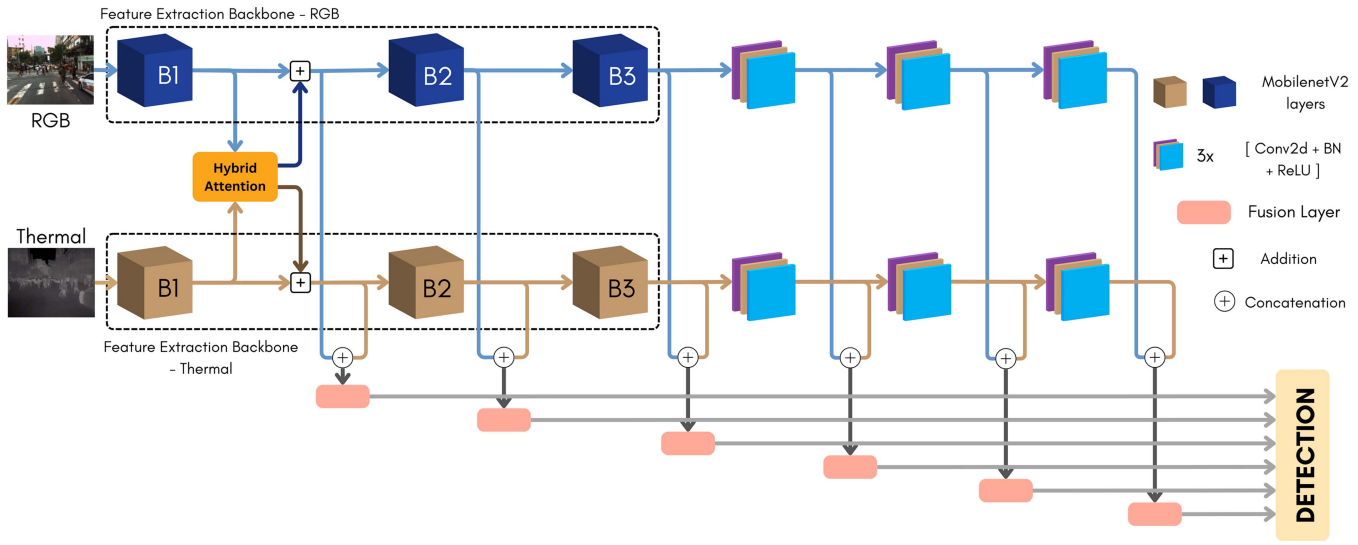

Fig. 3. Network Architecture for the proposed HA-MLPD. Modality-specific features are extracted from the respective image using a feature extraction backbone (here MobilenetV2 [27], and the blocks B1, B2 and B3 are described as in Table I). The HA module then enhances the features by attending only to the regions with useful information. This is followed by a sequence of shared convolutional blocks that extract features common to both modalities. Finally, the extracted multiscale features are fed to the SSD detector [15] to predict the pedestrian bounding boxes and their confidence scores.

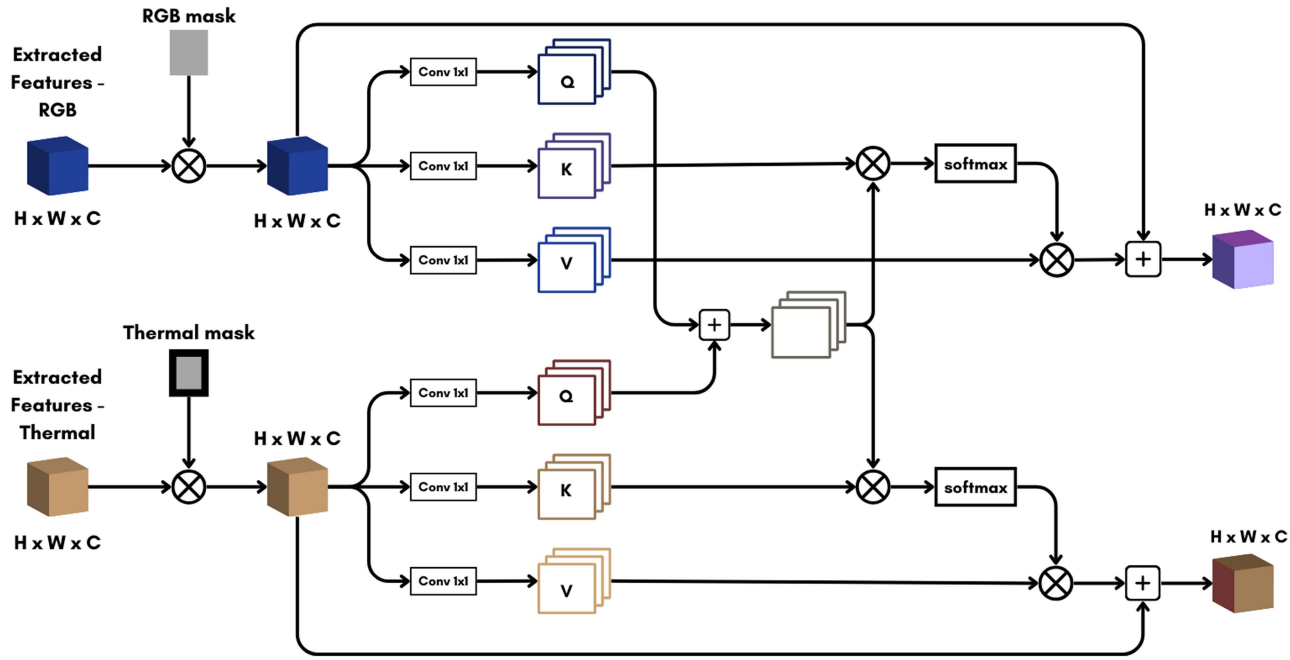

Fig. 4. Illustration of the novel HA module developed. The use of masks to remove blackout regions helps in *attending* to only regions with useful information. The results reported in Section IV show that the use of HA helps to reduce performance degradation during inference. Q,K,V depict Queries, Keys and Values.

used to extract features (F_{rgb}, $F_{thermal}$) from both images:

$$F_{\text{rgb}} = \phi_{\text{rgb}}(I_{\text{rgb}})$$
$$F_{\text{thermal}} = \phi_{\text{thermal}}(I_{\text{thermal}}) \quad (1)$$

where I_{rgb}, I_{thermal} represents RGB and thermal images respectively, and ϕ_{rgb}, ϕ_{thermal} the corresponding feature extractors. In our experiments, MobileNetV2 [27] was chosen for its compactness (see Fig. 3 for overview of the corresponding architecture), and VGG-16 [33] as the original MLPD backbone. These feature extractors are composed of B consecutive blocks of layers:

$$\phi_{\text{rgb}} = \phi_{\text{rgb}}^{(B)} \circ \cdots \circ \phi_{\text{rgb}}^{(2)} \circ \phi_{\text{rgb}}^{(1)}$$
$$\phi_{\text{thermal}} = \phi_{\text{thermal}}^{(B)} \circ \cdots \circ \phi_{\text{thermal}}^{(2)} \circ \phi_{\text{thermal}}^{(1)} \quad (2)$$

Features extracted after the first blocks $\phi_{\text{rgb}}^{(1)}$, $\phi_{\text{thermal}}^{(1)}$, denoted in what follows as F_{rgb}, F_{thermal} for simplification, are then fed to our novel HA module.

Hybrid-Attention Module: The HA module shown in Fig. 4 is the core of the proposed method. It enhances RGB and thermal features using the attention mechanism. The module switches between cross- and self-attention in response to the blackout regions in the input images. For that, we use masks M_{rgb}, M_{thermal} corresponding to blackout regions to filter out features originating from these regions ($\otimes$ denotes element-wise multiplication):

$$f_{\text{rgb}} = M_{\text{rgb}} \otimes F_{\text{rgb}},$$
$$f_{\text{thermal}} = M_{\text{thermal}} \otimes F_{\text{thermal}}. \quad (3)$$

The Keys, Queries and Values for each modality are then generated using 1×1 convolution layers:

$$Q_{\text{rgb}} = \text{Conv}_{1\times1}(f_{\text{rgb}}),$$
$$K_{\text{rgb}} = \text{Conv}_{1\times1}(f_{\text{rgb}}),$$
$$V_{\text{rgb}} = \text{Conv}_{1\times1}(f_{\text{rgb}}),$$
$$Q_{\text{thermal}} = \text{Conv}_{1\times1}(f_{\text{thermal}}),$$
$$K_{\text{thermal}} = \text{Conv}_{1\times1}(f_{\text{thermal}}),$$
$$V_{\text{thermal}} = \text{Conv}_{1\times1}(f_{\text{thermal}}). \quad (4)$$

The combined Query Q_{c} is then computed as the sum of the two modality-specific queries:

$$Q_{\text{c}} = Q_{\text{rgb}} + Q_{\text{thermal}}. \quad (5)$$

Removing the features corresponding to the blackout regions using the masks and then combining the queries will therefore make the modalities cross-attend where both modalities are available, and self-attend in the blackout regions. Following the standard practice, attended features f^*_{rgb} and f^*_{thermal} are calculated as:

$$f^*_{\text{rgb}} = \text{softmax}(Q_{\text{c}}^\top K_{\text{rgb}})V_{\text{rgb}},$$
$$f^*_{\text{thermal}} = \text{softmax}(Q_{\text{c}}^\top K_{\text{thermal}})V_{\text{thermal}}. \quad (6)$$

Finally, the enhanced features f'_{rgb} and f'_{thermal} for each modality are obtained as:

$$f'_{\text{rgb}} = f_{\text{rgb}} + f^*_{\text{rgb}},$$
$$f'_{\text{thermal}} = f_{\text{thermal}} + f^*_{\text{thermal}}. \quad (7)$$

Fusion Layer: The enhanced features are then processed by the remaining feature extraction layers, to extract higher-level dual-modality-guided information. After each block of layers, the features from both branches are concatenated then fused using shared network layers. These layers are composed of 2D convolutions followed by Batch Norm and ReLu, similar to the MLPD baseline (see Fig. 3).

Detector Head: The multi-level fused features are then passed through the detection head. In our architecture, and similar to MLPD, we use the SSD [15] model for object detection. However, it can be replaced with any other state-of-the-art detector in practice. The detector outputs pedestrian bounding box locations and confidence scores.

Loss function: The model is trained with regression loss on the bounding box locations $\mathscr{L}_{bbox}$ as in SSD [15] and multi-label loss $\mathscr{L}_{multilabel}$ from MLPD [6], balanced by a scaling factor λ :

$$\mathscr{L} = \mathscr{L}_{bbox} + \lambda\mathscr{L}_{multilabel}. \quad (8)$$

D. Masking Data Augmentation

In addition to the data augmentations from the MLPD baseline, we included data augmentations using masking in our training process to foster the resilience of our approach. Our method involves masking the complete RGB and thermal modalities (probability of 10% for each), as well as randomly masking patches of either modality again with a probability of 10% for RGB masking and 10% for thermal masking. Note that we avoid masking the same region in both modalities simultaneously. Also, these augmentations are implemented exclusively during the training phase and are not used during inference.

IV. Experiments and Results

A. Dataset

The KAIST Multispectral Pedestrian Dataset [1] consists of 95,328 RGB-Thermal pairs fully overlapped captured in an urban environment. The provided ground truth consists of 103,128 pedestrian bounding boxes in 1,182 instances. In our experiments, we sample 1 frame out of 2 so that 25,076 frames are used for training, as in [6]. For evaluation, we also follow the standard evaluation criterion, which consists of sampling 1 out of every 20 frames, so the results are evaluated in 2,252 frames where 1,455 frames were recorded during the day and 797 frames at night [6]. We use paired annotations for training [7] and sanitised annotations for evaluation [8].

To generate the complete blackout cases arising from sensor failures, the original pixels values are replaced with zeros for either of the modalities as shown in Fig. 2. For sides blackout cases, the original images are divided into three equal-sized vertical portions, and opposite side portions (e.g., left part in the RGB image and right part in the thermal image) are replaced with zero value pixels in each modality. For the surrounding blackout scenarios, we centre-crop the thermal image (96 pixels cropped at both the top and bottom and 120 pixels on the left and right sides) and replace the removed regions with zero pixels, while retaining the entire RGB image.

B. Training Details

HA-MLPD with MobileNetV2 backbone – We extended the MobileNetV2 architecture with further convolutional layers, as shown in Fig. 3. To initialise, we used ImageNet pre-trained weights for blocks (B1, B2, B3) and the remaining convolution kernels were initialised using values sampled from a normal distribution (std = 0.01). The network training process spanned 200 epochs, with an early stop callback in place to stop training if no improvement was observed for 50 epochs. The model was trained using SGD with an initial LR, momentum, and weight decay set at $1e^{-3}$, 0.9, and $4e^{-5}$, respectively. The LR was scheduled to decrease at the 150th and 190th epochs with a gamma value of 0.1.

HA-MLPD with VGG-16 backbone – Like the original MLPD, we use VGG16 pre-trained on ImageNet with batch normalisation, from $Conv1$ to $Conv5$, and the remaining convolution kernels are initialised with values drawn from the normal distribution (std = 0.01). The HA module is adopted at the output of $Conv4$ [6]. The model is trained by Stochastic Gradient Descent (SGD) with the initial learning rate (LR), momentum, and weight decay, as $1e^{-4}, 0.9$, and $5e^{-4}$, respectively. The mini-batch size is set to 8 and the input image size is resized to

TABLE II
EXPERIMENT RESULTS ON KAIST DATASET WITH TWO ("DUAL MODALITY") OR ONLY ONE ("RGB BLACKOUT", "THERMAL BLACKOUT") MODALITIES

Method	Dual Modality			RGB Blackout			Thermal Blackout			All
	MR (All)	MR (Day)	MR (Night)	MR (All)	MR (Day)	MR (Night)	MR (All)	MR (Day)	MR (Night)	Average difference
ACF [1]	47.32	42.57	56.17	-	-	-	-	-	-	-
Halfway Fusion [13]	25.75	24.88	26.59	-	-	-	-	-	-	-
FusionRPN+BF [34]	18.29	19.57	16.27	-	-	-	-	-	-	-
IAF R-CNN [16]	15.73	14.55	18.26	-	-	-	-	-	-	-
IATDNN+IASS [17]	14.95	14.67	15.72	-	-	-	-	-	-	-
CIAN [35]	14.12	14.77	11.13	-	-	-	-	-	-	-
MSDS-RCNN [8]	11.34	10.53	12.94	36.36	39.53	28.67	82.97	76.04	97.68	+27.60
AR-CNN [7]	9.34	9.94	8.38	17.70	21.95	8.64	77.03	67.54	97.85	+18.73
MBNet [18]	8.13	8.28	7.86	55.56	57.49	46.81	80.20	71.88	100	+32.01
SSD-RGB [15]	-	-	-	-	-	-	34.63	25.38	53.86	-
SSD-Thermal [15]	-	-	-	21.12	25.63	12.58	-	-	-	-
MLPD [6]	7.58	7.95	6.95	16.34	*20.07*	8.22	*23.95*	*16.88*	*39.37*	0
HA-MLPD (VGG-16)	*5.19*	*6.16*	*3.81*	*15.21*	20.41	*4.88*	**20.6**	**15.86**	**30.76**	**-2.29**
HA-MLPD (MobileNetv2)	**4.42**	**5.42**	**2.55**	**13.9**	**18.38**	**3.62**	24.29	16.92	40.83	*-1.75*

512 (H) × 640 (W). The network was trained for 40 epochs, and the LR was scheduled to decrease after 20 and 36 epochs. The models were trained on a single GPU accelerator node featuring 2× AMD Rome CPUs (32 cores @ 2.35GHz) and 4× NVIDIA A100-40 GPUs.

C. Adverse Inference-Time Conditions Simulation

To simulate sensor failure, we replaced one image of the RGB-Thermal pair by a full black image (zero pixel values). To simulate partial overlapping, we replace different parts of the images by black regions (see Fig. 2). Side blackouts simulate discrepancies in lateral fields of view, whereas surrounding blackouts simulate differences in either focal length or sensor resolution.

D. Metrics

Following the standard practice in the field, and especially the MLPD baseline [6], we report the log-average Miss Rate (MR) [36] at an Intersection-over-Union threshold of 0.5 to summarise the detector performance. This metric gives a stable and informative assessment of detector performance [36].

E. Results

Table II presents a comparison of the performance of our method with the existing literature in pedestrian detection scenarios using dual RGB and thermal modalities, especially in cases of sensor failure, such as RGB and thermal blackouts. Our method consistently shows competitive or superior results, demonstrating robustness in blackout conditions, particularly excelling in scenarios with RGB blackouts. It is important to highlight that even in complete blackouts of one modality, our method, in which HA then collapses in self-attention, outperforms both the RGB-only and Thermal-only models (referred to as SSD-RGB & SSD-Thermal). Furthermore, Table II includes comparisons with MLPD [6]. The colour and font coding highlights the best and second-best results in **bold red** and *italic blue*, respectively. HA-MLPD with the MobileNetv2 backbone achieves the highest performance in dual modality and RGB blackout scenarios, while HA-MLPD with the original VGG-16 backbone excels when RGB information is unavailable. In summary, our method demonstrates state-of-the-art performance, underscoring the effective RGB-Thermal fusion strategy in managing diverse real-world inference conditions.

Table III delves into the performance of the different methods in the context of inference-time misalignments involving extrinsic and intrinsic discrepancies in cameras. The scenarios considered include Sides Blackout with RGB-Thermal and Thermal-RGB misalignments, as well as Surrounding Blackout; see Fig. 2. Our method consistently exhibits the best performance across these misalignment scenarios, outperforming baseline methods such as MSDS-RCNN [8], AR-CNN [7], MBNet [18], SSD-RGB [15], SSD-Thermal [15] and MLPD [6]. Both backbones lead to second or best performance, with an average margin of more than 5 percentage points (pp) with respect to competitors. Overall, the table emphasises the robustness of HA-MLPD to inference-time blackouts, showcasing its efficacy in scenarios encountered in real-world applications. Fig. 5 summarizes the results from Tables II and III in a more easily interpretable plot.

F. Ablation Study

The ablation study in Table IV investigates the impact of our contributions in inference-time blackout scenarios, specifically Sides Blackout with RGB-Thermal, Sides Blackout with Thermal-RGB and surrounding blackout conditions. The table outlines the MR results for different configurations: (1) MLPD, (2) HA-MLPD without the HA mechanism (w/o HA) but with masking data augmentation, and (3) our proposed HA-MLPD. All three methods are with the VGG-16 backbone. The results demonstrate a systematic improvement with each modification, with lower MR values indicating better performance. The addition of data augmentation (Aug) results in a notable reduction in MR values across all scenarios ($-3.86pp$ on average). Furthermore, incorporation of HA contributes to an additional improvement, achieving the lowest MR values ($-5.34pp$ on average, compared to MLPD). Overall, the ablation study underscores the significance of both HA and data augmentation in improving the model's robustness under inference-time blackouts.

TABLE III
EXPERIMENT RESULTS ON KAIST DATASET WITH PARTIAL OVERLAP BETWEEN MODALITIES

Method	Sides Blackout (RGB-Thermal) MR (All)	MR (Day)	MR (Night)	Sides Blackout (Thermal-RGB) MR (All)	MR (Day)	MR (Night)	Surrounding Blackout MR (All)	MR (Day)	MR (Night)	All Avg. diff.
MSDS-RCNN [8]	43.00	-	-	59.42	-	-	47.22	-	-	+30.05
AR-CNN [7]	32.18	-	-	54.59	-	-	57.58	-	-	+28.28
MBNet [18]	56.65	-	-	63.81	-	-	46.99	-	-	+35.98
SSD-RGB [15]	63.75	-	-	51.08	-	-	34.63	-	-	+29.99
SSD-Thermal [15]	38.73	-	-	59.98	-	-	55.06	-	-	+31.42
MLPD* [6]	17.31	19.46	13.44	23.03	20.41	27.38	19.16	18.11	22.41	0
HA-MLPD (VGG-16)	*13.40*	*15.32*	**9.55**	**18.24**	**18.11**	**18.97**	*11.84*	**10.94**	*14.43*	*-5.34*
HA-MLPD (MobileNetv2)	**12.84**	**13.90**	*10.34*	*19.29*	*18.67*	*20.97*	**11.26**	*11.56*	**11.25**	**-5.37**

**: MR values for day and night images were computed from pre-trained model released by authors; not reported in [6].*

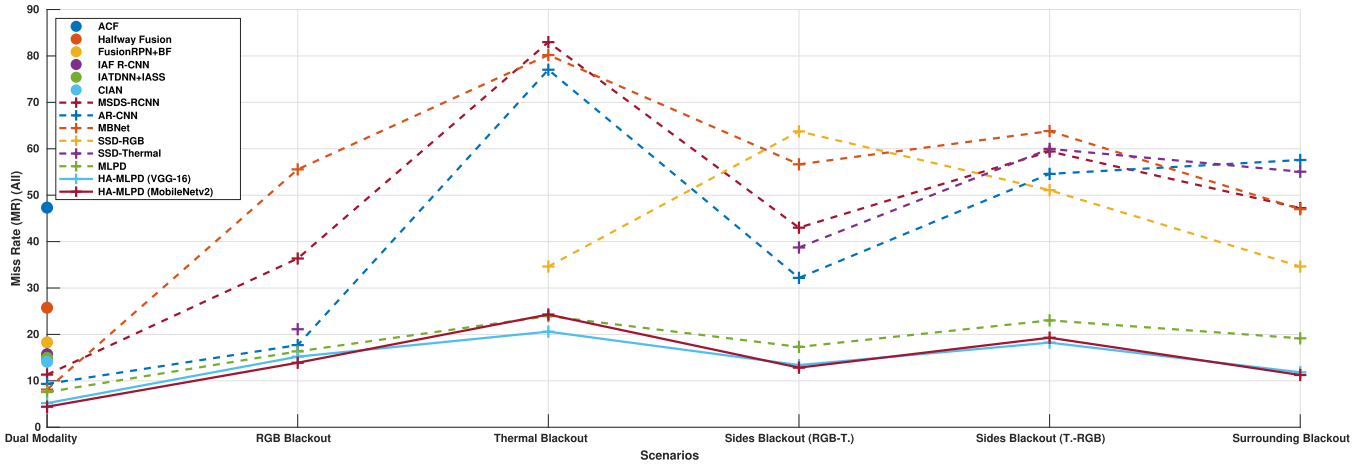

Fig. 5. Visualization of experiment results on KAIST Dataset.

TABLE IV
ABLATION STUDY AND COMPARISON WITH THE MLPD BASELINE

Method	Sides Blackout (RGB-Thermal) MR (All)	MR (Day)	MR (Night)	Sides Blackout (Thermal-RGB) MR (All)	MR (Day)	MR (Night)	Surrounding Blackout MR (All)	MR (Day)	MR (Night)	All Average difference
MLPD [6]	17.31	19.46	13.44	23.03	20.41	27.38	19.16	18.11	22.41	0
HA-MLPD w/o HA	*15.66*	*17.81*	*10.65*	*19.79*	*19.64*	*20.6*	*12.48*	*13.17*	**10.63**	*-3.86*
HA-MLPD	**13.40**	**15.32**	**9.55**	**18.24**	**18.11**	**18.97**	**11.84**	**10.94**	*14.43*	**-5.34**

G. Discussion

a) Dual modality performance: The results achieved in the dual modality scenario, i.e. when both modalities are fully overlapping and available, placed HA-MLPD in the second place (username: UniLu) on the leaderboard[1] at the time of submission. It is worth noting that this places our method above competitors that were specifically designed for dual modality conditions only, whence not robust to blackout scenarios. On the contrary, our method operates a necessary trade-off to cope with such challenging conditions, but still outperforms most other methods under the normal conditions.

b) Model efficiency: In our evaluation, we used two models, one with the VGG-16 backbone, which contains a total of 59.88 M parameters, and the other with the MobileNetV2 backbone, which contains a total of 13.96 M with a 4× reduction in the total number of parameters. Due to its larger size, the VGG-16 backbone required more time to train per epoch with a smaller batch size, while the smaller MobileNetV2 backbone allowed faster iteration through each epoch. However, MobileNetV2 required training for more number of epochs to achieve the best results. The overall duration of the training slightly favoured MobileNetV2. Given that the MobileNetV2 backbone is smaller, it can be easily accommodated on hardware with limited resources and can deliver a higher frame-per-second (FPS) rate during inference without compromising the performance. Note that our model takes approx. 20 ms (compared to 42 ms for MLPD baseline) to process an image during inference on a desktop workstation containing an Intel i7-11700 @ 2.50 GHz CPU and 1x NVIDIA GeForce RTX 3090 GPU.

c) Reliance on masks: Our approach relies on the masks of the blackout regions to guide the network through hybrid-attention mechanisms. The success of the proposed method depends on the accuracy of these masks. If the blackout regions are not

[1]https://eval.ai/web/challenges/challenge-page/1247/leaderboard/3137

correctly identified, the HA mechanism could be misguided, leading to suboptimal detection performance.

V. Conclusion

In this letter, we tackle the challenges of multispectral pedestrian detection in real-world scenarios, such as adverse inference-time configurations and hardware resource limitations. We introduced a novel HA mechanism to mitigate performance degradation arising from modality-specific lack of information. During training, the HA module enables learning both generalised and discriminative features through self- or cross-attention mechanisms. A mobile-friendly backbone is also used for higher energy efficiency. Extensive experimental comparisons demonstrate that the proposed method is robust to realistic stereo conditions. Furthermore, an ablation study highlighted that both introduced HA mechanism and data augmentation play crucial roles in improving the model's robustness. Future work will include deploying the method on edge device for real-condition testing, and evaluating the HA mechanism on other modalities (e.g., depth images).

Acknowledgment

The simulations were performed on the Luxembourg National Supercomputer MeluXina. The authors gratefully acknowledge the LuxProvide teams for their expert support.

References

[1] S. Hwang, J. Park, N. Kim, Y. Choi, and I. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baselines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.

[2] S. J. Krotosky and M. M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 4, pp. 619–629, Dec. 2007.

[3] Z. Guo, X. Li, Q. Xu, and Z. Sun, "Robust semantic segmentation based on RGB-thermal in variable lighting scenes," *Measurement*, vol. 186, 2021, Art. no. 110176. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224121010903

[4] A. González et al., "Pedestrian detection at day/night time with visible and FIR cameras: A comparison," *Sensors*, vol. 16, no. 6, 2016, Art. no. 820.

[5] T. F. LLC, "Free flir thermal dataset for algorithm training." 2022. Accessed: Jan. 2024. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/

[6] J. Kim, H. Kim, T. Kim, N. Kim, and Y. Choi, "MLPD: Multi-label pedestrian detector in multispectral domain," *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 7846–7853, Oct. 2021.

[7] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5126–5136.

[8] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proc. Brit. Mach. Vis. Conf.*, pp. 1–12, 2018.

[9] J. Lee et al., "Robust pedestrian detection by combining visible and thermal infrared cameras," *Sensors*, vol. 15, no. 5, pp. 10580–10615, 2015.

[10] L. Zhang, B. Wu, and R. Nevatia, "Pedestrian detection in infrared images based on local shape features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[11] J. Davis and V. Sharma, "Fusion-based background-subtraction using contour saliency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2005, pp. 11–11.

[12] Y. Yuan, X. Lu, and X. Chen, "Multi-spectral pedestrian detection," *Signal Process.*, vol. 110, pp. 94–100, 2015.

[13] J. Liu, S. Zhang, S. Wang, and D. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–13.

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

[15] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, vol. 9905, pp. 21–37.

[16] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, pp. 161–171, 2019.

[17] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Inf. Fusion*, vol. 50, pp. 148–157, 2019.

[18] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 787–803.

[19] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 139–158.

[20] H. Zhang, E. Fromont, S. Lefèvre, and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 72–80.

[21] Y. Yang, K. Xu, and K. Wang, "Cascaded information enhancement and cross-modal attention feature fusion for multispectral pedestrian detection," *Front. Phys.*, vol. 11, 2023, Art. no. 1121311.

[22] J. U. Kim, S. Park, and Y. M. Ro, "Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 1157–1165.

[23] N. Narayan, N. Sankaran, S. Setlur, and V. Govindaraju, "Learning deep features for online person tracking using non-overlapping cameras: A survey," *Image Vis. Comput.*, vol. 89, pp. 222–235, 2019.

[24] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[25] S. M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 133–146.

[26] L. V. Ma, T. T. D. Nguyen, B.-N. Vo, H. Jang, and M. Jeon, "Track initialization and re-identification for 3D multi-view multi-object tracking," *Inf. Fusion*, vol. 111, 2024, Art. no. 102496.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[28] C.-F. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 357–366.

[29] S. Cai, P. Li, E. Su, and L. Xie, "Auditory attention detection via cross-modal attention," *Front. Neurosci.*, vol. 15, 2021, Art. no. 652058.

[30] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, Eds., Curran Associates, Inc., vol. 30, 2017.

[31] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.

[32] M. Arar, Y. Ginger, D. Danon, A. Bermano, and D. Cohen-Or, "Unsupervised multi-modal image registration via geometry preserving image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13410–13419.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[34] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 243–250.

[35] L. Zhang et al., "Cross-modality interactive attention network for multispectral pedestrian detection," *Inf. Fusion*, vol. 50, pp. 20–29, 2019.

[36] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.