Illumination-Aware Hallucination-Based Domain Adaptation for Thermal Pedestrian Detection

Qian Xie, Ta-Ying Cheng, Zhuangzhuang Dai, Vu Tran, Niki Trigoni, and Andrew Markham

Abstract— Thermal imagery is emerging as a viable candidate for 24-7, all-weather pedestrian detection owning to thermal sensors' robust performance for pedestrian detection under different weather and illumination conditions. Despite the promising results obtained from combining visible (RGB) and thermal cameras in multi-spectral fusion techniques, the complex synchronization requirements, including alignment and calibration of sensors, impede their deployment in real-world scenarios. In this paper, we introduce a novel approach for domain adaptation to enhance the performance of pedestrian detection based solely on thermal images. Our proposed approach involves several stages. Firstly, we use both thermal and visible images as input during the training phase. Secondly, we leverage a thermal-to-visible hallucination network to generate feature maps that are similar to those generated by the visible branch. Finally, we design a transformer-based multi-modal fusion module to integrate the hallucinated visible and thermal information more effectively. The thermal-to-visible hallucination network acts as domain adaptation, allowing us to obtain pseudo-visual and thermal features using solely thermal input. Based on the experimental results, it is observed the mean average precision (mAP) increases by 4.72% and the miss rate decreases by 7.56% on the KAIST dataset when compared to the baseline model.

Index Terms— Pedestrian detection, thermal image, modality hallucination, transformed-based fusion.

I. INTRODUCTION

A. Background

PEDESTRIAN detection [1], [2], [3] is a vital component for the task of perception in autonomous driving [4], and it is an active research topic in computer vision with a wide range of other applications as well, such as security surveillance [5], [6]. Conventional pedestrian detection algorithms widely utilize visible (i.e., RGB) images as the input data source [7]. However, visible-based pedestrian detectors are usually prone to miss targets due to poor visibility at night
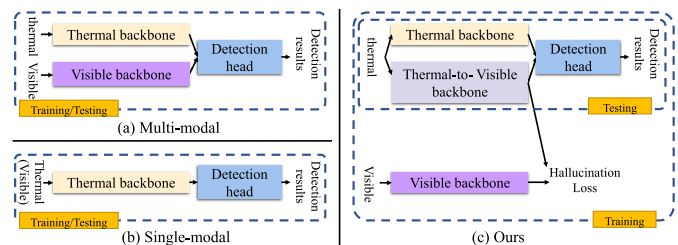
Fig. 1. Illustration of the proposed method, compared with multi-modal and single-modal methods. Only thermal images are needed for our method at the testing stage.

or under bad illumination conditions. Thus, thermal cameras are becoming more commonplace in the field of pedestrian detection [8], [9], [10], owing to their capability of capturing emitted thermal radiation rather than scene appearance.

Pedestrian detection methods using thermal images can be divided into two categories: thermal-only-based (i.e., single-modal) and thermal-visible fusion-based (i.e., multi-modal or multispectral), as illustrated in Fig. 1 (a) and (b). Thermal-visible fusion-based methods make use of information from both modalities. This category of methods surpasses the performance of visible-only and thermal-only-based approaches due to the advantageous complementary information provided by visible and thermal images. However, these methods necessitate the use of two types of sensors during both training and inference stages, incurring additional costs and complexities. Furthermore, these sensors must work simultaneously and under strict prerequisites for time synchronization, alignment, and calibration of multiple devices, further increasing the cost of deploying multi-modal data collection devices in practical applications [11]. Additionally, visible images are always taken as input even when their information may be contaminated by illumination variations and weather conditions. In contrast, thermal-only-based methods can work well across the challenging scenes discussed above. And thermal-only-based methods are easier to be employed since only thermal sensors are needed during inference, compared to those thermal-visible fusion-based methods which require both well-calibrated thermal and visible sensors. However, a notable drawback of thermal-only-based methods is their limited performance in certain conditions, as the absence of visible information can impede object detection, especially when the scene is well-illuminated and object appearance plays a significant role.
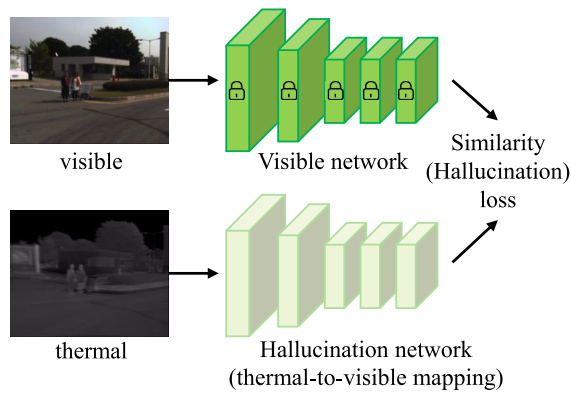
Fig. 2. Illustration of the hallucination mechanism [12] on thermal pedestrian detection. The hallucination network aims at regressing visible features from thermal image inputs. When the hallucination network is trained well, we can then generate visible-like features from thermal images.
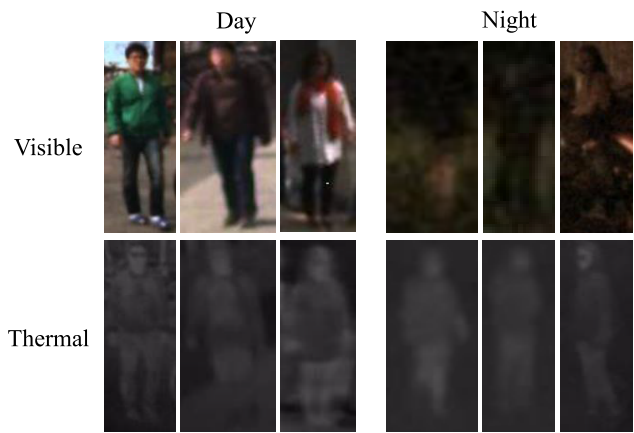

Fig. 3. Illustration of the domain inconsistency problem between thermal and visible domain. Similar thermal images could correspond to different visible images, which will hinder hallucination, as there is not a simple one-to-one mapping.

B. Motivation

Considering the availability of labeled datasets, there are more visible image pedestrian detection datasets than thermal image datasets. Hence, to utilize existing visible image datasets, thermal-only-based methods usually leverage domain adaptation to transfer knowledge from the visible domain to the thermal domain [13], [14], [15], [16]. Domain adaptation is a sub-category of transfer learning that aims at improving the performance of models on the target domain (thermal domain in this paper) by using the knowledge learned from the source domain (visible domain). A simple example of domain adaptation on thermal pedestrian detection is to initialize the thermal detection network using weights learning from visible images using the same network, rather than training thermal networks from scratch using random initialization.

Modality hallucination [12] is a concept of generating information for a secondary modality, given information from a primary modality, which can also be seen as one of the domain adaptation techniques. Specifically, hallucination stands for the procedure of learning the mapping from one domain (thermal domain in our paper) to another domain (visible domain), as shown in Fig. 2. Once the mapping (hallucination network) is learned, we can hallucinate visible features given thermal images as input. The features from the hallucination network are called hallucinated features. The hallucinated features come from thermal images, but they contain visible-related information. Hallucination mechanism has yielded impressive results on several multi-modal related tasks, such as visible-depth images-based segmentation [17], object detection [12] and localization [18]. These approaches typically use visible images to infer pseudo depth map through modality hallucination to deal with challenging scenes like bad weather or illumination where visible images do not work particularly well. Similarly, we propose that visible features could be learned from thermal images through modality hallucination, specifically for the task of pedestrian detection, which is an area that has not been extensively studied. This motivates us to investigate an effective modality hallucination framework for thermal-only pedestrian detection. However, implementing this idea naively could suffer from the domain inconsistency problem due to the significant differences between the thermal and visible domains. Specifically, while visible images can contain highly diverse content under different illumination and weather conditions, thermal images remain relatively consistent, as demonstrated in Fig. 3. Therefore, two similar thermal images may correspond to different visible images during modality hallucination, creating a domain inconsistency problem that can impede the learning of a useful mapping function from the thermal to visible domains. Subsequently, although several studies [19], [20] have focused on generating thermal images or features from visible images, none have addressed the problem of learning visible features from thermal images. We chose thermal images as the basis for generating visible information due to their illumination robustness. Thermal images remain in good condition regardless of lighting changes, providing reliable information. In contrast, visible images excel during the day but deteriorate at night. By leveraging the stability of thermal images, our approach ensures consistent performance and robustness across different lighting conditions.

C. Contributions

Our approach leverages a thermal-to-visible hallucination network to generate pseudo-visible information from thermal images. In summary, this work presents three main contributions:

- We introduce a novel approach for enhancing thermal-based pedestrian detection performance by utilizing thermal-visible image pairs during training and incorporating a multi-layer hallucination architecture to generate visible information from thermal input.
- We propose an illumination-aware hallucination loss to improve the traditional hallucination loss by weighting the loss based on illumination levels. This enhances the robustness of the thermal-to-visible hallucination procedure, enabling it to learn visible features from thermal images even under low illumination conditions.
- We design a transformer-based multi-modal fusion module that integrates the hallucinated visible

information with thermal information. The module includes spatial-wise and channel-wise fusion blocks, which employ self-attention and cross-attention operations to improve the spatial feature representation.

II. RELATED WORK

Our review will commence with an exploration of the literature surrounding pedestrian detection using thermal sensors, as it aligns with our method. Additionally, since the hallucination mechanism serves as the core concept of this paper, it can be considered a form of cross-modal transfer learning from the visible to the thermal domain. Therefore, we will also investigate related works on cross-modal transfer learning.

A. Thermal-Based Pedestrian Detection

Pedestrian detection has always been a research hotspot in recent years. Research on visible-only-based pedestrian detection based has achieved abundant results [21], [22], [23]. However, due to the inherent characteristics of visible images (i.e., ineffective in low-light conditions), research on pedestrian detection based on thermal imaging has gained increasing attention. As this paper primarily utilizes thermal imaging sensors for detection, only related research on pedestrian detection based on thermal imaging will be discussed. Efforts have been made to automatically detect pedestrians in open scenes by using thermal sensors [11], [14], [24], [25], [26]. These methods can be divided into two categories based on the input modalities: thermal-visible-based and thermal-only-based.

With multispectral input, methods are working on effectively fusing visible and thermal information [27], [28], [29], [30]. MSDS-RCNN [31] proposed to jointly optimize semantic segmentation and pedestrian detection tasks, and then integrate outputs from different branches to obtain the final detection. Based on the observation of the differences in illumination conditions during day and night, Guan et al. [32] and Li et al. [33] designed strategies to learn weights of thermal and visible modalities according to the illumination conditions. To address the modality imbalance problem, Zhou et al. [34] presented a Differential Modality Aware Fusion module to make the visible and thermal modalities complement each other. To more efficiently integrate the features from visible and thermal streams, Cao et al. [29] designed a multispectral channel feature fusion module to assign different attention values to visible and thermal features according to the illumination conditions. While detectors based on both thermal and visible sensors can leverage complementary information from both domains, their use of multiple devices can lead to complications, making deployment in real-world applications difficult. In contrast, our method utilizes only thermal sensors, allowing for easy deployment. Additionally, our approach can also extract visible information through the use of the hallucination branch, providing complementary information from both domains like thermal-visible-based methods.

With thermal-only input, most works focus on leveraging visible information to boost thermal-based pedestrian detection performance via domain adaptation. For instance,

Kieu et al. [25] proposed to synthesize realistic thermal versions of input RGB images through a Generative Adversarial Network (GAN) and then mixed real and improved fake thermal images as a way of data augmentation to relieve the problem of limited thermal image dataset in object detection. Similarly, Bongini et al. [35] also proposed to produce synthetic thermal data by rendering 3D models using a thermal shader in the Unity game engine, and then utilized GAN to improve fake thermal image realism. TC Det [16] established an auxiliary branch of day-and-night prediction to guide the domain adaptation from visible to thermal. Although thermal-based detectors are convenient to deploy, their detection performance is limited due to the lack of visible information. Our proposed method shares the same convenience as thermal-only-based methods in terms of model deployment as it only uses thermal sensors for inference. Additionally, the proposed method leverages the hallucination branch to obtain visible information from thermal images, thereby compensating for the deficiency of thermal-only-based methods in lacking visible information.

B. Cross-Modal Transfer Learning

Cross-modal transfer learning aims to find task correlations from one domain to another domain [36], [37], [38], [39]. The majority of research focuses on visible-to-depth domain transfer. Without depth images as input at test time, visible images are used to infer depth information to boost the performance of downstream tasks. For instance, Piasco et al. [40] proposed to reconstruct the depth map in outdoor, large-scale, image-based localization. For visible-thermal modality transfer, Devaguptapu et al. [41] proposed a 'pseudo-multimodal' object detector on the thermal domain by fusing information of thermal images and pseudo visible images generated using Cycle-GAN [42]. Xu et al. [19] proposed to reconstruct thermal image patches by learning a non-linear mapping between visible and thermal domains through a designed Region Reconstruction Network (PRN). With the learned network embedded as an additional branch, the final detection network can generate both visible and thermal features with only visible images as input. Taking inspiration from prior research, we introduce a thermal-to-visible feature hallucination network to learn the feature transformation from thermal to visible images. Our approach differs from existing methods, as the goal of our hallucination network is to directly produce pseudo-visible features rather than visible images. To further improve the transformation process, we incorporate multi-level feature similarity constraints.

III. METHOD

A. Main Architecture

The goal of this paper is to leverage the existing visible-thermal image pairs as the training set to get powerful pedestrian detection when supplied with thermal-only images as input at test time. Therefore, our model has different architectures at the training and testing stages. The proposed method can be implemented by adding multi-layer
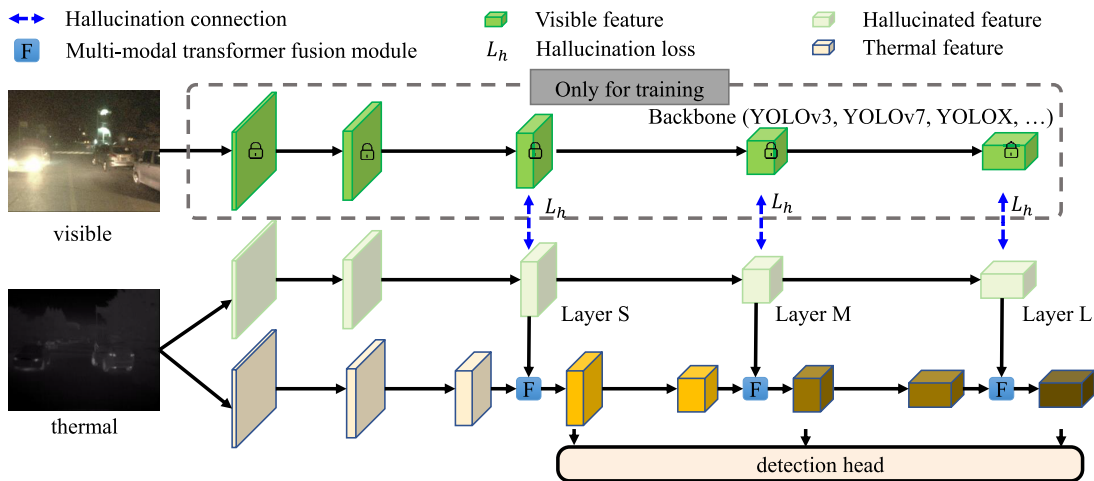
Fig. 4. The overall architecture of the proposed network. S, M, and L correspond to three feature maps responsible for detecting small, middle, and large objects. *Note that the visible branch is omitted at test time.*

hallucination links and feature fusion modules into existing detectors. Thus, our method can be applied to existing one-stage detection models, e.g., the YOLO [43] series detectors (YOLOv3 [44], YOLOX [45], YOLOv7 [46], et al.). The network architecture of our method is not limited to one specific architecture.

In Fig. 4, we give a general architecture of our method based on YOLO series detectors. In the following, we use YOLOv3 as the example to explain the procedure of design of our detection model. As shown in the figure, we have three branches to extract feature maps from input images at the training stage. All three backbones for feature extraction are based on DarkNet-53 (backbone in YOLOv3), except for the main branch in which three multi-modal transformer fusion modules are inserted to fuse features from the hallucination branch. The YOLO detection head takes multi-scale feature maps and individually predicts detection results in three different scales (i.e., small, medium, and large), in order to cope with the scale variation of targets. Based on the architecture of YOLOV3, our modifications are quite simple, mainly focusing on two parts: 1) the multi-layer hallucination network and 2) the multi-modal transformer fusion modules.

B. Multi-Layer Hallucination Network

The basic idea of this paper is to equip the network with the ability to generate pseudo-visible features from thermal inputs. An intuitive approach is to first generate visible images from the input thermal images and then use a relatively standard visual-thermal detector. However, instead of taking such a low-efficiency solution (i.e., the thermal image $I^T \rightarrow$ thermal feature $f^T \rightarrow$ visible image $I^V \rightarrow$ visible feature f^V chain), we propose to directly reconstruct the intermediate visible feature maps from the thermal input (i.e., the thermal image $I^T \rightarrow$ visible feature f^V chain). Since multi-modal features are integrated over three scales (i.e., small, medium, and large) in YOLOV3, we thus present a multi-layer hallucination module to better learn the visible features.

As illustrated in Fig. 4, at the training stage, we first encode the input visible and thermal images $I^V, I^T \in \mathbb{R}^{H \times W \times 3}$ into

feature maps through three branches, i.e., visible, thermal-to-visible (hallucination), and thermal branches in the order from top to bottom. H, W represent the height and width of images, and 3 is number of RGB channels. For each branch, we then take features at three different layers, corresponding to three scales. Thus, we can get three categories of features $f^V = [f_S^V, f_M^V, f_L^V]$, $f^{T-V} = [f_S^{T-V}, f_M^{T-V}, f_L^{T-V}]$, $f^V = [f_S^T, f_M^T, f_L^T]$. S, M, L indicate features are from small, medium, and large scales respectively. The hallucination process occurs between the visible and thermal-to-visible branches, which is trained by minimizing the difference d between visible features f^V and thermal-to-visible features f^{T-V}.

1) Illumination-Aware Hallucination Loss: In Fig. 3, we can observe that thermal images that appear similar can correspond to vastly different visible images, making it difficult for the hallucination network to learn an accurate mapping from the thermal to the visible domain when working with low-illumination images. To address this issue, we propose an illumination-aware hallucination loss that reduces the impact of visible images under low-illumination conditions during hallucination. This is accomplished by assigning adaptive weights to the thermal-visible pairs based on their respective illumination conditions, allowing us to treat low-illumination thermal-visible pairs as outliers.

First, we propose to encourage the hallucination optimization to focus more on thermal-visible image pairs under good illumination conditions by placing higher weights on their losses. The weights are determined by the following equation:

$$W_I = \begin{cases} \alpha & \text{for } b \leq t_b, \\ 1.0 & \text{otherwise,} \end{cases} \tag{1}$$

where b represents the brightness, which is the average intensity of all pixels that construct the input visible image. t_b is the brightness threshold. A smaller weight α will be assigned to those images whose brightness values are lower than t_b. In our experiments, α and t_b are set to be 0.2 and 70.

Second, we adopt the Huber loss [47] considering it is less sensitive to outliers than the generally used L2 loss. In all, our

illumination-aware hallucination loss is defined as follows:

$$L = W_I \frac{1}{n} \sum_{i=1}^{n} H\left(f_i^V, f_i^{T-V}\right),$$

$$H(,) = \begin{cases} \frac{1}{2}\left(f_i^V - f_i^{T-V}\right)^2 & \text{for } \left|f_i^V - f_i^{T-V}\right| \le \delta, \\ \delta\left(\left|f_i^V - f_i^{T-V}\right| - \frac{1}{2}\delta\right) & \text{otherwise,} \end{cases}$$

$$(2)$$

where n is the number of elements in feature maps f_i^V, f_i^{T-V} and δ is a threshold to make the loss function less sensitive to outliers, which is set to $\delta = 1.0$.

C. Multi-Modal Transformer Fusion Module

After thermal and hallucinated thermal-to-visible feature maps f^T, f^{T-V} are obtained through the backbone network and hallucination network, features from two branches should be integrated together $\Theta_{\textbf{Fusion}}$ to generate the fused feature f'', which can be formulated as:

$$f'' = \Theta_{\textbf{Fusion}}(f^T, f^{T-V}).$$

$$(3)$$

Note that there is no need for a specific design for feature matching or alignment because the hallucination branch and the thermal branch share the same network architecture. As a result, the hallucinated thermal-to-visible feature maintains the same dimension as the thermal feature. Instead of performing multi-modal feature fusion solely in a channel-wise way or in a spatial-wise way like existing methods, we propose a simple yet effective fusion module to integrate features from thermal and pseudo-visible (i.e., hallucination) streams in both spatial and channel views. In such a way, our network is capable of explicitly modeling the relationship between multi-modal features, rather than roughly concatenating or adding them. As illustrated in Fig. 5, the proposed fusion module consists of two sub-blocks: a spatial-wise fusion block $\Theta_{\textbf{S-Fusion}}$ and a channel-wise fusion block $\Theta_{\textbf{C-Fusion}}$. This module aims to enhance thermal features by exploring the spatial relationship in thermal feature maps and adaptively integrating the pseudo-visible features. The proposed fusion procedure can be summarized as:

$$f'^T, f'^{T-V} = \Theta_{\textbf{S-Fusion}}(f^T, f^{T-V}),$$
$$f'' = \Theta_{\textbf{C-Fusion}}(f'^T, f'^{T-V}).$$

$$(4)$$

1) Spatial-Wise Fusion Block: As shown in Fig. 5 (b), our spatial-wise fusion block consists of two attention modules: intra-modal attention and inter-modal attention module. We treat the thermal branch as the primary information and hallucinated features as side information because the thermal information is more reliable here. Thus, the intra-modal attention module takes thermal features $f_i^T, i \in \{S, M, L\}$ as input and captures the spatial relationship between thermal features themselves. The inter-modal attention module takes both thermal and thermal-to-visible features $f_i^{T-V}, i \in \{S, M, L\}$ as input and is designed to encode the relationship between the two different modalities.

Attention mechanisms have achieved great success in many visual tasks [48]. In this paper, we employ the commonly used attention mechanism [49]:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V},$$

$$(5)$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ represent query, key and value vectors, and d_k is the dimension of key features. In our method, for the thermal and thermal-to-visible features $f_i^T, f_i^{T-V} \in \mathbb{R}^{H_i \times W_i \times C_i}$ in the i-th scale, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ can be denoted as vectors with size of $\mathbb{R}^{(H_i W_i) \times C_i}$, where (H_i, W_i) is the resolution of the feature maps, C_i is the number of feature channels. The computation of the similarity matrix between query and key tokens results in $\mathbf{Q}\mathbf{K}^T \in \mathbb{R}^{(H_i W_i) \times (H_i W_i)}$, which is subsequently normalized by division with d_k, the dimension of the key matrix. This normalization enhances the stability of gradient values during training. In the intra-modal attention module, the three inputs are both from the thermal feature f_i^T. That is, $\mathbf{Q} = proj_Q(f_i^T), \mathbf{K} = proj_K(f_i^T), \mathbf{V} = proj_V(f_i^T)$, where $proj_Q, proj_K, proj_V$ are feature embedding operations implemented by Linear layers. However, in the inter-modal attention module, only the query matrix is generated from the thermal feature, while the key and value matrix are from thermal-to-visible features. Specifically, for the inter-modal attention module, $\mathbf{Q} = proj_Q(f_i^T), \mathbf{K} = proj_K(f_i^{T-V}), \mathbf{V} = proj_V(f_i^{T-V})$. In such a way, the proposed spatial-wise fusion block can not only encode the spatial relationship within thermal features through the self-attention mechanism but also capture the spatial relationship between thermal and thermal-to-visible features through the cross-attention mechanism. In all, this block can get the enhanced intermediate thermal and thermal-to-visible feature maps $f_i'^T, f_i'^{T-V}$ from the input thermal and thermal-to-visible feature maps f_i^T, f_i^{T-V}, as illustrated in Fig. 5 (b).

It is worth noting that the input thermal or thermal-to-visible feature maps are usually in high spatial resolution. Since producing attention maps $\mathbf{Q}\mathbf{K}^T \in \mathbb{R}^{(H_i W_i) \times (H_i W_i)}$ between two high spatial resolution feature maps is computationally expensive, we first adopt a global average pooling layer to downsample the feature maps to a lower resolution before feeding them to the fusion block. Here, we fix the lower resolution as 8×8. Finally, bilinear interpolation is performed to upsample the output features from attention modules to the original resolution.

2) Channel-Wise Fusion Block: As known, thermal and visible images should have different degrees of influence on detection results under varying illumination conditions. For instance, thermal images should contribute more to results at night, since they can capture better features of pedestrians. For this purpose, we introduce a channel-wise modality feature fusion block, which can adaptively adjust weights for two modalities. As illustrated in Fig. 5 (c), the block is fed with the intermediate thermal and thermal-to-visible feature maps $f_i'^T, f_i'^{T-V}$. Similar to [50], the feature maps are first squeezed into channel descriptor vectors with the size of $1 \times 1 \times C_i$, where C_i is the number of channels in the feature maps. The two squeezed vectors are then concatenated to form
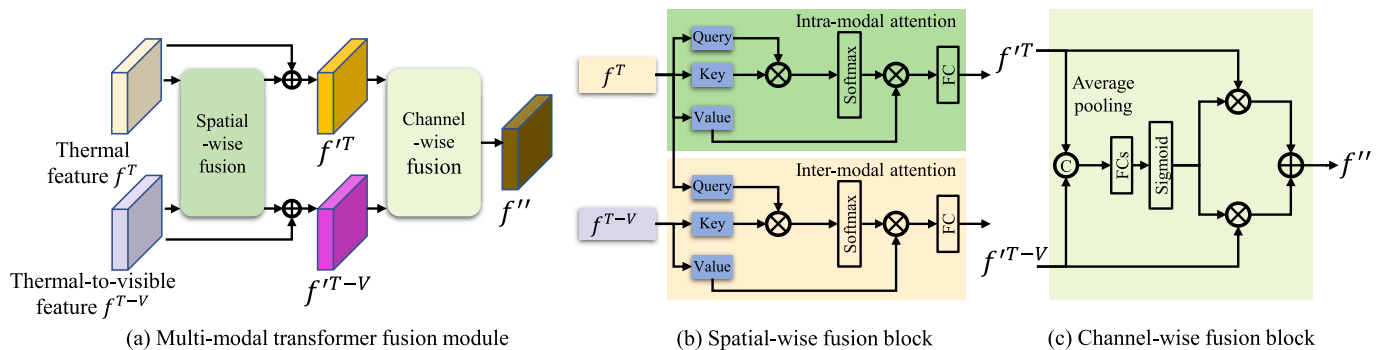
Fig. 5. The architecture of our multi-modal transformer fusion module (a), which consists of two blocks: spatial-wise fusion block (b) and channel-wise fusion block (c). In the spatial-wise fusion block, query, key, and value represent query, key, and value vectors in Eq. 5 respectively.

a unified vector of $1 \times 1 \times 2C_i$. After concatenation, channel-wise merging weights $W = (w_i^T; w_i^{T-V})$ are then learned through two FC layers ($2C_i \rightarrow C_i$, $C_i \rightarrow 2C_i$) with a sigmoid activation function in the latter FC layer. Subsequently, the input two feature maps $f_i^{'T}$ and $f_i^{'T-V}$ multiply corresponding scores to achieve channel weighting. The final fused feature map f_i'' for the i-th scale is obtained by adding different portions of input modality features.

IV. EXPERIMENTS

A. Datasets and Metrics

In order to validate the efficacy of our novel pedestrian detection approach, we conduct experiments on two widely-used public datasets: the KAIST and FLIR datasets. Both datasets consist of pairs of visible and thermal images captured during both daytime and nighttime conditions, presenting a challenge of domain inconsistency which can be addressed through the utilization of the proposed illumination-aware hallucination loss.

1) KAIST Dataset: KAIST dataset [51] is a multispectral pedestrian dataset with pixel-level aligned visible-thermal image pairs. It is captured in traffic scenes under different environments, including different lighting conditions from day to night. The original dataset contains 95,328 visible-thermal image pairs with 50,172 for training and 45,156 for testing, whose pixel resolution is 512×640. However, there are annotation errors (such as imprecise localization and misclassification) in the original dataset. As is common practice, we use the processed version dataset (with frame sampling and annotation sanitization) that consists of 7,601 image pairs for training and 2,252 pairs for testing. In the test set, there are 1,455 Day images and 797 Night images.

2) FLIR Dataset: FLIR dataset is also a well-known multispectral object detection dataset that captures street scenes via FLIR cameras in a car. It consists of both RGB and thermal domain images. However, the RGB and thermal image pairs are not all aligned well in the original dataset. Therefore, we use an aligned version [52] in which aligned image pairs are manually selected. Finally, 5142 well-aligned visible-thermal image pairs are remained, of which 4129 pairs for training and 1013 pairs for testing.

3) Metrics: To evaluate our method, we use the same evaluation metric for pedestrian detection proposed in [53], like most of other comparing methods. We adopt the standard *log-average miss rate (MR)* to summarize detection performance, calculated by averaging miss rate at nine *false positive per image (FPPI)* rates evenly spaced in a log-space in the range of $[10^{-2}, 10^0]$. Specifically, the normal miss rate is the ratio of false-negatives to all pedestrians:

$$MR = fn/(tp + fn) \tag{6}$$

And the *false positive per image (FPPI)* is the ratio of false positives to all the tested frames:

$$FPPI = fp/(number\ of\ tested\ frames) \tag{7}$$

Following [16], we also give results on three sub-metrics, i.e., MR_All for all the test images, MR_Day for day images and MR_Night for night images. Moreover, we also report mAP (mean Average Precision) which is also commonly used in object detection.

B. Implementation Details

As stated in Sec. III-A, our method can be applied to many existing one-stage detection models, e.g., the YOLO series detector. Here, we implement our method based on three YOLO-based detectors (i.e., YOLOv3 [44], YOLOX [45] and YOLOv7 [46]). Among them, YOLOv3 was proposed in 2018 to improve the original YOLOv1 [43] and YOLOv2 [54]. And YOLOX and YOLOv7 were presented in the recent two years, representing the state-of-the-art YOLO series detectors. Thus, these three implementations can demonstrate the generalization ability of our method. For our YOLOv3-based network, we implement it from scratch using PyTorch. For the YOLOX- and YOLOv7-based networks, we build them based on the code provided in this repository.[1] For each of our networks, we use the same scheme to implement our method, that is, using three backbones and adding multi-layer hallucination connections and three fusion modules, as in Fig. 4.

Our networks are trained in two stages. In the first stage, we train a multi-modal network without the hallucination branch, as in Fig. 1 (a). Thus, visible features are directly fused into the main branch of the thermal stream. In the

---

[1] https://github.com/bubbliiiing/yolov7-pytorch

Fig. 6. Visual comparison results between the baseline model (YOLOv3) and our method (YOLOv3-based network). *Note that visible images here are not used during testing*.

second stage, we insert the hallucination network and fix weights of the visible branch that have been trained in the first stage. In this way, the hallucination network, with thermal images as input, can learn hallucinated visible features by simultaneously optimizing the hallucination loss and the final detection loss. To accelerate the training of the whole network, the hallucination network is initialized using the parameters of the pre-trained visible backbone. At test time, the visible branch is totally eliminated, allowing operation with thermal-only devices.

The size of input images is 640×512. As in [16], we use weights pre-trained on MS COCO as a starting point in our experiments. For the YOLOv3-based network, we set the initial learning rate to 0.0001, and divide it by 10 when the loss turns stagnant. Training stops after two divisions on the learning rate. The batch size is set to 8. All other hyper-parameters are identical to the original YOLOV3. For the YOLOX- and YOLOv7-based networks, we use the default hyper-parameters in their repositories.

TABLE I
COMPARED RESULTS WITH OUR METHOD AND BASELINE MODEL. v AND t REPRESENT HALLUCINATION BRANCH WEIGHT INITIALIZATION SCHEMES, VISIBLE AND THERMAL, RESPECTIVELY

Datasets	Models	mAP↑	MR_All↓	MR_Day↓	MR_Night↓
KAIST	Baseline	58.36	31.05	37.81	15.40
	Ours (t)	62.42	26.12	33.47	10.50
	Ours (v)	63.08	23.49	30.11	9.64
FLIR	Baseline	73.07	39.87	-	-
	Ours (t)	75.53	28.28	-	-
	Ours (v)	76.38	27.29	-	-

C. Comparisons

1) Comparison With the Baseline Model: To verify the effectiveness of the proposed method, we first compare our method with a baseline model. Here, we use YOLOv3 as the baseline detector and compare our YOLOv3-based network with it. The reason why we use YOLOv3 as the baseline model here is that YOLOV3 is a simple yet efficient model without

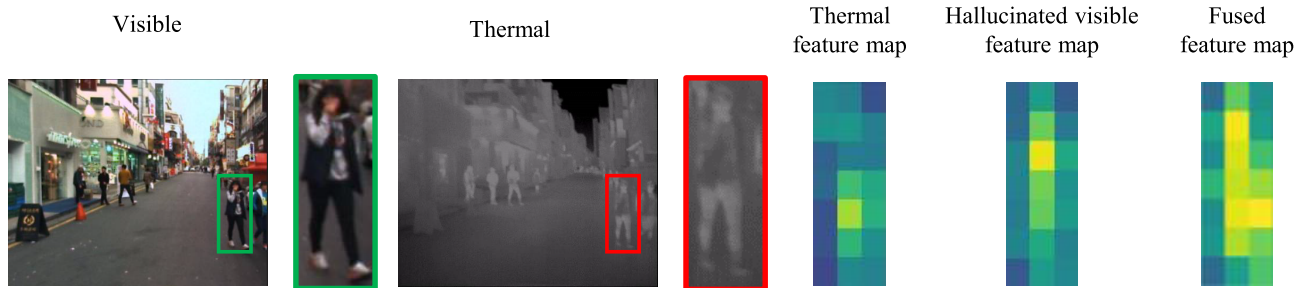| Visible	Thermal		Thermal feature map	Hallucinated visible feature map	Fused feature map

Fig. 7. Feature maps at Scale M. During the testing stage, the thermal feature map and the hallucinated visible feature map are fused together. *Note that visible images are not used during testing.*

TABLE II

COMPARISON OF DOMAIN ADAPTATION THERMAL-BASED PEDESTRIAN DETECTION METHODS. *Note That All the Methods Use Visible Images for Domain Adaptation During Training*

Detectors	MR_all↓	MR_day↓	MR_night↓
SSD300 [13]	69.81	-	-
VGG16-two-stage [14]	46.30	53.37	31.63
ResNet101-two-stage [14]	42.65	49.59	26.70
Bottom-up [15]	35.20	40.00	20.50
TC Det [16]	27.11	34.81	10.31
Kim et al. [55]	19.16	24.70	8.26
Kim et al. [56]	15.87	20.26	6.48
Ours (YOLOv3)	23.49	30.11	9.64
Ours (YOLOX)	17.83	22.12	7.14
Ours (YOLOv7)	**14.65**	**18.88**	**5.92**

many intricate tricks. We do not expect complex architectures, like YOLOX and YOLOv7, to impact the validation of the effectiveness of the proposed modules. To get the optimal result of the baseline model on thermal images, we first pre-train the baseline model on visible images and save the weights after convergence. We then use these weights to initialize the network and finetune it on thermal images. As can be seen in Tab. I, our method achieves 63.08 and 23.49 in terms of mAP and MR_All, outperforming the baseline by 4.72 and 7.56 on KAIST dataset, which demonstrates the effectiveness of the proposed hallucination strategy in thermal-based pedestrian detection task. A similar trend can also be seen in the FLIR dataset, which demonstrates the pretty generalization ability of our network on various datasets. In addition to the quantitative results, some visual results are also given in Fig. 6. Our method shows a clear improvement in detecting more correct targets and generating less false detection, as compared to the baseline model. The reason is that our method can generate hallucinated visible information to improve detection performance even without visible images as input during testing. It is worth noting that, in the fourth row of the figure, false positive detections are both in the result images of the baseline model and our method. This is because the baseline model and our method have no customized structure to handle crowded pedestrians, which could be one of our future works.

As stated in [12], different initializations of the hallucination network also affect the convergence. We further evaluate the detection performance of our hallucination network initialized with weights from thermal and visible branches at the first training stage. It can be noted that both weight initialization strategies surpass the baseline model, while the model initialized with visible weights (i.e., ours (v)) achieves better performance, consistent with the result in [12] that using weights from the branch to be hallucinated is a better choice.

2) Feature Map Visualizations: In order to gain a more comprehensive understanding of the hallucination process, we provide visual representations of the feature maps generated by each branch of our network in Fig. 7. The results demonstrate that the hallucination network is capable of producing visible feature maps that cannot be captured by the thermal feature maps alone. This supports the notion that the generated pseudo-visible features can supplement the thermal domain branch when actual visible data is either absent or unreliable.

3) Comparison With State of the Art: To demonstrate the superiority of our method over the existing methods, we also provide a detailed comparison to state-of-the-art domain adaptation methods on thermal-based pedestrian detection in Tab. II. As stated in Sec. IV-B, we implement three networks (i.e., *Ours (YOLOv3), Ours (YOLOX) and Ours (YOLOv7)* in Tab. II) for our method based on different YOLO detectors. As can be seen, the proposed *Ours (YOLOv3)* model surpass most of the comparing methods [13], [14], [15], [16], except for the two recent methods [55], [56]. The reason could be that YOLOv3 is a one-stage detector proposed in the year of 2018, which is too old compared to these two methods [55], [56] proposed in recent two years. Thus, when we replace the YOLOv3 architecture with the latest YOLO series architectures, the detection performance increases accordingly. Finally, the YOLOv7-based network (*Ours (YOLOv7)*) outperforms all the competitors both on day and night images on the KAIST dataset. Overall, the best performance of our method reaches 14.65 miss rate for all images (MR_All), exceeding current state-of-the-art method [56] by 1.22. It is worth noting that all the comparative methods use visible images to pre-train their network, i.e., they have access to the same input information (aligned thermal-visible image pairs) as our method.

D. Ablation Study

1) Multi-Layer vs Single-Layer vs Zero-Layer Hallucination: One of our contributions in this paper is the multi-layer hallucination scheme. Conventional hallucination networks usually add the hallucination operation after a

(a) Visible image   (b) Hallucinated feature with Huber loss   (c) Hallucinated feature with Hallucination loss   (d) Visible feature
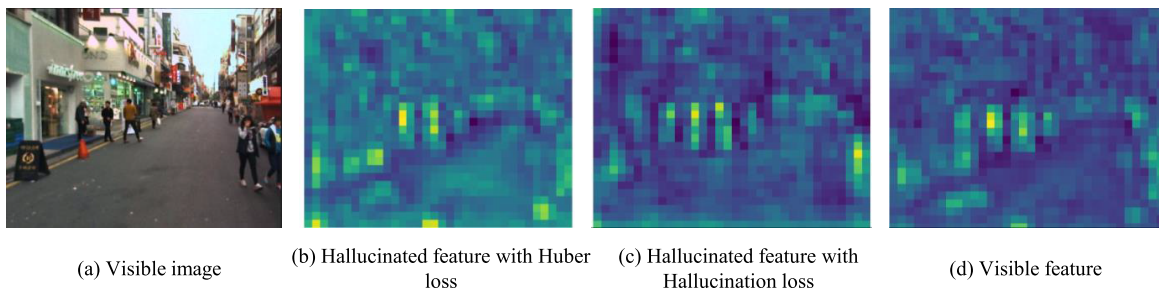
Fig. 8. Hallucinated visible feature maps from scale M. As can be seen, the hallucinated visible feature with the proposed hallucination loss (c) contains more meaningful information, compared to the visible feature with the one with Huber loss (b).

fixed single layer, i.e., single-layer hallucination networks. However, we are trying to produce hallucinated features in multiple layers, after which feature fusions are directly performed. To verify the effectiveness of the proposed multi-layer hallucination scheme over the conventional single-layer hallucination scheme, we implement several single-layer hallucination networks on the basis of both *Ours (YOLOv3)* and *Ours (YOLOv7)* architectures. Specifically, we keep just one of the three hallucination links in Fig. 4, forming three single-layer hallucination architectures, i.e., Single-S, Single-M, and Single-L. From the ablation experimental results in Tab. III, we can see the same performance trend for *Ours (YOLOv3)* and *Ours (YOLOv7)* architectures. That is, a single-layer hallucination scheme can improve the result from the zero-layer model, which indicates the hallucination link between the hallucination branch and the visible branch can actually force the hallucination branch to learn visible information and thus boost detection performance. Moreover, our multi-layer hallucination scheme achieves the highest detection performance compared to all three single-layer hallucination networks, demonstrating our multi-scale hallucination strategy is a more efficient way to learn thermal-to-visible mapping for pedestrian detection.

We also create a zero-layer model (Zero) by deleting all the hallucination links (i.e., hallucination losses) between the visible branch and the thermal-to-visible (hallucination) branch. In that case, we just use one more branch to extract thermal features. The worst results in Tab. III demonstrate that the performance improvement actually comes from the hallucination losses because the performance drops significantly when all the hallucination losses are deleted (Zero). In other words, the performance improvement is not brought by the wider network architecture in our method because the models in Tab. III all have the same network architecture width.

2) Illumination-Aware Hallucination Loss vs Huber Loss: To verify the effectiveness of the proposed illumination-aware hallucination loss, we investigate the performance with and without the proposed hallucination loss. As seen in Tab. IV, illumination-aware hallucination loss improves the miss rate by 2.20 against vanilla Huber loss. Another simple idea to solve the domain inconsistency problem is merely performing hallucination on day images, whose results are reported as Huber (day) in Tab. IV. As seen, the results of Huber (day) are even poorer than using both day and night images (i.e., Huber), which further demonstrates the superiority of our illumination-aware weighting loss scheme. There are two

TABLE III
DETECTION RESULTS OF SINGLE-LAYER AND MULTI-LAYER HALLUCINATION SCHEMES

Base model	#Hal	MR_All	MR_Day	MR_Night
YOLOv3	Zero	29.71	37.23	13.76
	Single-S	26.12	33.47	10.50
	Single-M	24.31	30.68	11.02
	Single-L	26.17	32.36	12.72
	Multiple(3 Links)	23.49	30.11	9.64
YOLOv7	Zero	18.61	24.81	9.56
	Single-S	16.38	20.15	8.74
	Single-M	15.16	18.57	6.47
	Single-L	16.99	20.75	7.82
	Multiple(3 Links)	14.65	18.88	5.92

TABLE IV
DETECTION RESULTS WITH DIFFERENT HALLUCINATION LOSSES

Loss function	mAP ↑	MR_All↓	MR_Day↓	MR_Night↓
Huber	62.13	25.69	32.61	11.00
Huber (day)	59.37	25.99	32.26	12.42
Illumination-aware	63.08	23.49	30.11	9.64

TABLE V
EFFECTS OF TWO PARAMETERS α AND t_b ON DETECTION PERFORMANCE

α	MR_All↓		t_b	MR_All↓
0.10	24.03		50	24.67
0.15	24.01		55	24.28
0.20	23.49		60	24.23
0.25	23.55		65	24.06
0.30	23.68		70	23.49
0.35	24.11		75	23.51
0.40	23.95		80	23.80

potential reasons for the lower performance of our method on the day dataset. Firstly, some of the day images may also suffer from poor illumination conditions. Secondly, since the hallucination process during training does not involve any night-time visible images, the pseudo-visible features generated from night thermal images may not be reliable. To gain a deeper understanding of the proposed hallucination loss, we visualize the feature maps obtained using both the Huber loss and our proposed loss in Fig. 8. As demonstrated, the hallucinated feature maps generated using our proposed loss contain fewer background noises and are more similar to the visible feature maps, indicating the effectiveness of the proposed loss in generating useful feature maps.

We further conduct experiments on the effects of two fixed parameters in Eq. 1. The results are shown in Tab. V. We can observe that the network achieves the best performance when α and t_b are equal to 0.2 and 70 respectively.

TABLE VI

DETECTION RESULTS OF THREE FUSION STRATEGIES OF THERMAL AND VISIBLE INFORMATION

Fusion scheme	mAP ↑	MR_All↓	MR_Day↓	MR_Night↓
Add	60.79	26.63	32.80	12.36
Concat	61.69	26.01	33.36	10.48
Ours	63.08	23.49	30.11	9.64

TABLE VII

COMPARISON BETWEEN OUR METHOD AND THE THERMAL-VISIBLE FUSION METHOD

Method	MR_All↓	MR_Day↓	MR_Night↓
Thermal-Visible Fusion	17.75	22.39	8.36
Thermal-Hallucination (Ours)	23.49	30.11	9.64
Thermal-only	31.05	37.81	15.40

3) Transformer-Based Multi-Modal Fusion vs Add or Concat Fusion: To validate the superiority of the proposed multi-modal fusion module, we conduct ablation experiments with different fusion strategies, including addition and concatenation. For concatenation, we use a convolutional layer to reduce the dimension to its original state, so that the feature map can be properly fed into the subsequent network. Results are presented in Tab. VI. It can be observed that the proposed fusion module achieves the best result. Our fusion module captures the complementary modality features in a more explicit way. We believe that the superiority comes from the carefully designed fusion module, which is more suitable for hallucinated visible and thermal features fusion.

E. Discussion on Performance

1) Comparison With the Thermal-Visible Fusion Model: As mentioned above, the motivation of this paper is to narrow the performance gap between thermal-only and thermal-visible detectors. Thus, we further conduct comparison experiments between our method and a thermal-visible fusion method on the basis of YOLOv3. Specifically, the original YOLOv3 and *Ours (YOLOv3)* are acting as the thermal-only detector and our method respectively. For the thermal-visible fusion model, we implement it by replacing the hallucination branch with the visible branch in Fig. 4. The comparison results are given Tab. VII. As shown, among the three methods, the thermal-visible fusion model achieves the best performance undoubtedly. Even though there still exists a certain gap (17.75 to 23.49, in terms of MR_All) between our method and the thermal-visible fusion model, the difference is much smaller than the difference between thermal-only and thermal-visible models (17.75 to 31.05), which demonstrates that our method can actually narrow the gap between thermal-only and thermal-visible detectors. Another notable point is that, compared to the thermal-visible fusion model, the performance of our method for day images (22.39 to 30.11) decrease more than for night images (8.36 to 9.64). Moreover, without visible data at testing time, the detection accuracy for night images of our method (8.36 to 9.64) does not decrease as significantly as the thermal-only model (8.36 to 15.40). That means the hallucination mechanism has a better performance boost at
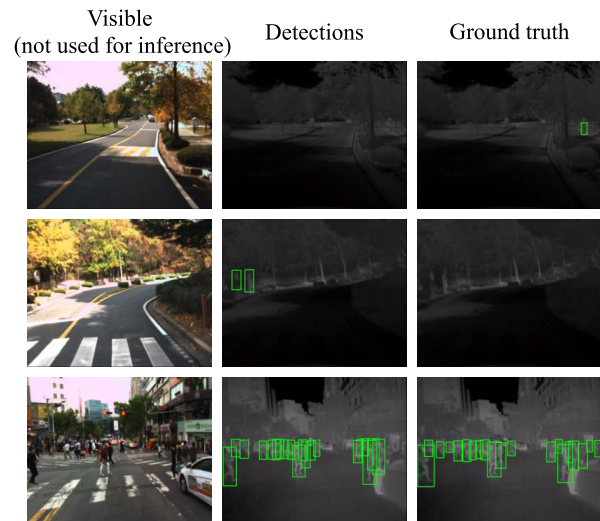

Fig. 9. Examples of failure detection results under long-distance, misrecognized, and crowded scenes.

night, and our method achieves a comparable detection performance to a thermal-visible fusion-based detector in terms of MR_Night. A reasonable reason could be that night-time thermal images are more reliable than day-time thermal images which could be contaminated by sunlight, as showcased below.

2) Failure Cases: Although our method has shown improvements in boosting thermal-only detectors by hallucinating visible-like features, the characteristics of thermal sensors can still result in failure detection cases. The hallucination branch in our method relies on thermal input to generate visible-like features, which may be incorrect when the input thermal images are contaminated. For instance, as demonstrated in Fig. 9, the detector misses the target in the first row, as it is not clear in the thermal image when far away from the thermal sensor. In the second row, tree trunks dissipate heat like pedestrians due to direct sunlight, leading to misclassification by the detector. Moreover, since our method does not include optimization for occluded cases, incorrect detections may occur in crowded scenes, as shown in the third row of Fig. 9.

V. CONCLUSION

In this paper, we propose a novel thermal-only pedestrian detector to narrow the performance gap between single-modality and multi-modality approaches. Specifically, we utilize the multi-layer modality hallucination strategy to produce visible related features from thermal images. In this way, the proposed network is capable of incorporating information from the visible domain even without visible images as input at testing time. We design an illumination-aware hallucination loss to relieve the domain inconsistency problem in the hallucination process from thermal to visible domains. To efficiently integrate thermal and hallucinated visible features, we present a novel multi-modal fusion module that can adaptively fuse features from two modal streams both in spatial and channel-wise ways. It is worth noting that our method needs paired thermal and visible images as training data. That means complicated synchronization setting

of visible and thermal sensors is still required before model training. However, only a one-time sensor setting is needed to collect training data. After the network is trained, it can be easily deployed into applications with only thermal sensors. Hence, the advantage of our method lies in the high efficiency of deployment for practical applications using the trained network.

Future work. Our proposed method demonstrates superior performance compared to state-of-the-art methods when using only thermal images as input during testing. However, this performance can be further improved in situations where visible images are available during testing. Thus, one potential avenue for future work is to explore the benefits of combining visible, hallucinated, and thermal features to further enhance detection accuracy. Another limitation of our approach is the fixed threshold used to determine illuminance weights in Equation 1, which is vulnerable to changes in illumination. Additionally, this simplistic threshold-based approach may erroneously classify some overexposed scenes as having desirable illumination, introducing poor-quality visible features into the hallucination process. To address this limitation, future work could incorporate a light sub-network to evaluate the quality of visible images, enabling the adaptive determination of hallucination weights. Furthermore, incorporating the hallucination branch into the network increases the number of parameters, thereby increasing its complexity compared to single-modal detectors. However, this tradeoff is necessary to achieve improved performance.

REFERENCES

[1] B. Han, Y. Wang, Z. Yang, and X. Gao, "Small-scale pedestrian detection based on deep neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 7, pp. 3046–3055, Jul. 2020.

[2] X. Liu, K.-A. Toh, and J. P. Allebach, "Pedestrian detection using pixel difference matrix projection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1441–1454, Apr. 2020.

[3] P. Yang, G. Zhang, L. Wang, L. Xu, Q. Deng, and M.-H. Yang, "A part-aware multi-scale fully convolutional network for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1125–1137, Feb. 2021.

[4] H. Gao, D. Fang, J. Xiao, W. Hussain, and J. Y. Kim, "CAMRL: A joint method of channel attention and multidimensional regression loss for 3D object detection in automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 8, pp. 8831–8845, Aug. 2023.

[5] X. Wang, M. Wang, and W. Li, "Scene-specific pedestrian detection for static video surveillance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 361–374, Feb. 2014.

[6] M. Bilal, A. Khan, M. U. Karim Khan, and C.-M. Kyung, "A low-complexity pedestrian detection framework for smart video surveillance systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2260–2273, Oct. 2017.

[7] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[8] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 988–997.

[9] R. Lahmyed, M. El Ansari, and A. Ellahyani, "A new thermal infrared and visible spectrum images-based pedestrian detection system," *Multimedia Tools Appl.*, vol. 78, no. 12, pp. 15861–15885, Jun. 2019.

[10] L. Ding, Y. Wang, R. Laganière, D. Huang, and S. Fu, "Convolutional neural networks for multispectral pedestrian detection," *Signal Process., Image Commun.*, vol. 82, Mar. 2020, Art. no. 115764.

[11] C. Lu, S. Zhang, and M. Liu, "Pedestrian detection based on center, temperature, scale and ratio prediction in thermal imagery," in *Proc. 40th Chin. Control Conf. (CCC)*, Jul. 2021, pp. 7288–7293.

[12] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 826–834.

[13] C. Herrmann, M. Ruf, and J. Beyerer, "CNN-based thermal infrared person detection by domain adaptation," *Proc. SPIE*, vol. 10643, Jan. 2018, Art. no. 1064308.

[14] T. Guo, C. P. Huynh, and M. Solh, "Domain-adaptive pedestrian detection in thermal images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1660–1664.

[15] M. Kieu, A. D. Bagdanov, M. Bertini, and A. D. Bimbo, "Domain adaptation for privacy-preserving pedestrian detection in thermal imagery," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2019, pp. 203–213.

[16] M. Kieu, A. D. Bagdanov, M. Bertini, and A. D. Bimbo, "Task-conditioned domain adaptation for pedestrian detection in thermal imagery," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 1–12.

[17] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. Lau, and T. S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2864–2873.

[18] N. Piasco, D. Sidibé, V. Gouet-Brunet, and C. Demonceaux, "Improving image description with auxiliary modality for visual localization in challenging conditions," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 185–202, Jan. 2021.

[19] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4236–4244.

[20] V. V. Kniaz, V. A. Knyaz, J. Hladuvka, W. G. Kropatsch, and V. Mizginov, "ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 1–10.

[21] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5182–5191.

[22] G. Li, Y. Yang, and X. Qu, "Deep learning approaches on pedestrian detection in hazy weather," *IEEE Trans. Ind. Electron.*, vol. 67, no. 10, pp. 8889–8899, Oct. 2020.

[23] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4913–4934, Sep. 2022.

[24] M. A. Marnissi, H. Fradi, A. Sahbani, and N. E. B. Amara, "Unsupervised thermal-to-visible domain adaptation method for pedestrian detection," *Pattern Recognit. Lett.*, vol. 153, pp. 222–231, Jan. 2022.

[25] M. Kieu, L. Berlincioni, L. Galteri, M. Bertini, A. D. Bagdanov, and A. del Bimbo, "Robust pedestrian detection in thermal imagery using synthesized images," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8804–8811.

[26] M. Kieu, A. D. Bagdanov, and M. Bertini, "Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1, pp. 1–19, Feb. 2021.

[27] A. Wolpert, M. Teutsch, M. S. Sarfraz, and R. Stiefelhagen, "Anchor-free small-scale multispectral pedestrian detection," in *Proc. Brit. Mach. Vis. Conf., (BMVC)*, Sep. 2020, pp. 1–12.

[28] Y. Wang, X. Wei, X. Tang, H. Shen, and H. Zhang, "Adaptive fusion CNN features for RGBT object tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7831–7840, Jul. 2022.

[29] Z. Cao, H. Yang, J. Zhao, S. Guo, and L. Li, "Attention fusion for one-stage multispectral pedestrian detection," *Sensors*, vol. 21, no. 12, p. 4184, Jun. 2021.

[30] P. Wang, L. Zhou, M. Xiao, and P. Zhang, "Multi-spectral fusion network for full-time robust pedestrian detection," in *Proc. Int. Conf. Electron. Inf. Eng. Comput. Technol. (EIECT)*, Dec. 2021, pp. 159–169.

[31] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2018, pp. 1–15.

[32] 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," *Inf. Fusion*, vol. 50, pp. 148–157, Oct. 2019.

[33] 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, Jan. 2019.

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

[35] F. Bongini, L. Berlincioni, M. Bertini, and A. D. Bimbo, "Partially fake it till you make it: Mixing real and fake thermal images for improved object detection," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5482–5490.

[36] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 935–943.

[37] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4068–4076.

[38] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2827–2836.

[39] M. R. U. Saputra et al., "DeepTIO: A deep thermal-inertial odometry with visual hallucination," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1672–1679, Apr. 2020.

[40] N. Piasco, D. Sidibé, V. Gouet-Brunet, and C. Demonceaux, "Learning scene geometry for visual localization in challenging conditions," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 9094–9100.

[41] C. Devaguptapu, N. Akolekar, M. M. Sharma, and V. N. Balasubramanian, "Borrow from anywhere: Pseudo multimodal object detection in thermal imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1029–1038.

[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[44] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[45] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[46] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[47] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in Statistics*. Cham, Switzerland: Springer, 1992, pp. 492–518.

[48] H. Gao, J. Xiao, Y. Yin, T. Liu, and J. Shi, "A mutually supervised graph attention network for few-shot segmentation: The perspective of fully utilizing limited samples," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 14, 2022, doi: 10.1109/TNNLS.2022.3155486.

[49] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[50] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

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

[52] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 276–280.

[53] 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.

[54] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[55] J. U. Kim, S. Park, and Y. M. Ro, "Robust small-scale pedestrian detection with cued recall via memory learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3030–3039.

[56] J. U. Kim, S. Park, and Y. M. Ro, "Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, pp. 1–12.

Qian Xie received the B.Sc. and Ph.D. degrees in electrical engineering from the Nanjing University of Aeronautics and Astronautics (NUAA), China, in 2015 and 2021, respectively. He is currently a Research Associate with the Department of Computer Science, University of Oxford, U.K. Prior to the University of Oxford, he went to Cardiff University, U.K., as a joint-trained Ph.D. Student, in 2019, for 18 months. His current research interests include 3D vision, point cloud processing, deep learning, and scene understanding.

Ta-Ying Cheng received the B.Eng. degree in computer science from HKUST. He is currently pursuing the D.Phil. degree in computer science with the University of Oxford, U.K. Prior to this, he was a Research Assistant with the Computer Vision Laboratory, Academia Sinica. His current research interests include deep learning approaches for 3D computer vision tasks, with a strong passion for single and multi-view reconstructions under difficult settings.

Zhuangzhuang Dai is currently a Lecturer of computer science with Aston University. Before joining Aston University, he was a NIST Software Engineer with the Cyber-Physical Systems Group, University of Oxford, working on simultaneous localization and mapping (SLAM) systems for emergency responders and robots. His current research interests include sensor fusion, embedded systems, machine learning, computer vision, propagation modeling, the IoT, and urban data science.

Vu Tran received the bachelor's and M.Eng. degrees from the Ho Chi Minh University of Technology in 2009 and 2012, respectively, and the Ph.D. degree in computer science from Singapore Management University in 2020. He was a Post-Doctoral Research Associate with the Cyber-Physical Systems Group, Department of Computer Science, University of Oxford. His current research interests include mobile and wearable sensing, wireless communication and sensing, and indoor localization.

Niki Trigoni received the D.Phil. degree from the University of Cambridge in 2001. She is currently a Professor with the Department of Computer Science, Oxford University, and a fellow of the Kellogg College. She became a Post-Doctoral Researcher with Cornell University from 2002 to 2004 and a Lecturer with the Birkbeck College from 2004 to 2007. At Oxford University, she is currently the Director of the EPSRC Centre for Doctoral Training on Autonomous Intelligent Machines and Systems, a program that combines machine learning, robotics, sensor systems, and verification/control. She also leads the Cyber-Physical Systems Group, which is focusing on intelligent and autonomous sensor systems with applications in positioning, healthcare, environmental monitoring, and smart cities.

Andrew Markham received the Ph.D. degree from the University of Cape Town, South Africa, in 2008, researching the design and implementation of wildlife tracking systems, using heterogeneous wireless sensor networks. He is currently a professor working on sensing systems, with applications from wildlife tracking to indoor robotics to checking that bridges are safe. He works with the Cyber-Physical Systems Group. He designed novel sensors, investigated new algorithms (increasingly deep and reinforcement learning-based), and applied these innovations to solving new problems. Previously, he was an EPSRC Post-Doctoral Research Fellow, working on the UnderTracker Project.