Beyond Fusion: Modality Hallucination-based Multispectral Fusion for Pedestrian Detection

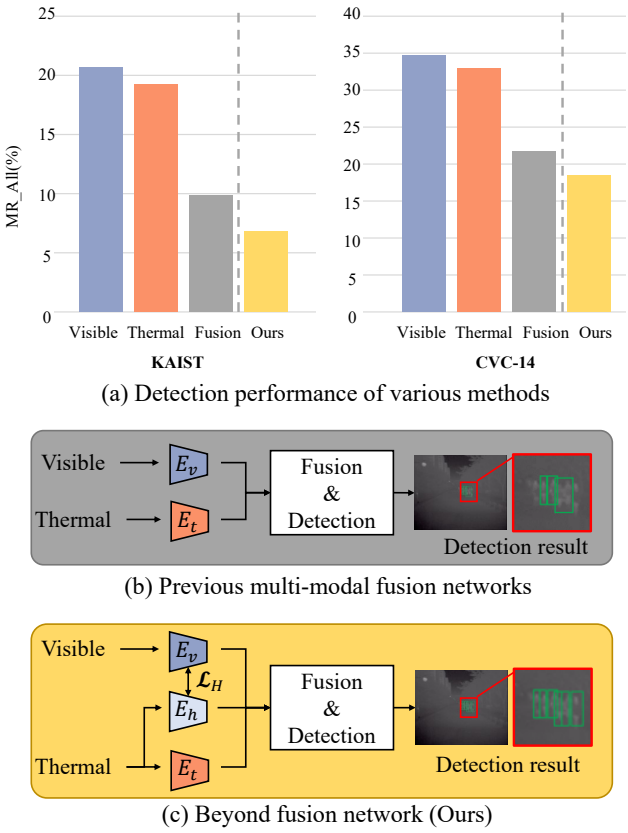
Qian Xie, Ta-Ying Cheng, Jia-Xing Zhong, Kaichen Zhou, Andrew Markham, Niki Trigoni
Department of Computer Science, University of Oxford, Oxford, UK
{qian.xie,ta-ying.cheng,jiaxing.zhong,rui.zhou,andrew.markham,niki.trigoni}@cs.ox.ac.uk

Abstract

*Pedestrian detection is a fundamental task for many downstream applications. Visible and thermal images, as the two most important data types, are usually used to detect pedestrians under various environmental conditions. Many state-of-the-art works have been proposed to use two-stream (*i.e.*, two-branch) architectures to combine visible and thermal information to improve detection performance. However, conventional visible-thermal fusion-based methods have no ability to obtain useful information from the visible branch under poor visibility conditions. The visible branch could even sometimes bring noise into the combined features. In this paper, we present a novel thermal and visible fusion architecture for pedestrian detection. Instead of simply using two branches to separately extract thermal and visible features and then fusing them, we introduce a hallucination branch to learn the mapping from the thermal to the visible domain, forming a novel three-branch feature extraction module. We then adaptively fuse feature maps from all three branches (*i.e.*, thermal, visible, and hallucination). With this new integrated hallucination branch, our network can still get relatively good visible feature maps under challenging low-visibility conditions, thus boosting the overall detection performance. Finally, we experimentally demonstrate the superiority of the proposed architecture over conventional fusion methods.*

1. Introduction

In the field of 2D pedestrian detection, visible image-based techniques have been well explored [1, 2, 20, 25, 26, 37]. However, an inevitable weakness of using only visible images is that they are easily affected by illumination changes, and detection performance drops heavily under poor illumination conditions. As a complementary data source, thermal images are more robust in detecting pedestrians in badly illuminated scenes by capturing the temperature information of objects instead of color information.



(a) Detection performance of various methods

(b) Previous multi-modal fusion networks

(c) Beyond fusion network (Ours)

Figure 1. **Comparison of our pedestrian detection method to thermal, visible, and fusion-based methods**, on two datasets of visible-thermal pair images. As shown, the detection performance (a) of our method (c) goes beyond the previous fusion architecture (b). E_v, E_t, E_h denote visible, thermal, and hallucinated feature encoders respectively, and $\mathcal{L}_H$ denotes the proposed hallucination loss.

Combining thermal and visible images, recent work on multi-spectral detectors has tackled the challenge of complex environments resulting in promising detection performance [17, 39]. Most thermal- and visible-based fusion architectures focus on boosting the effectiveness of multi-

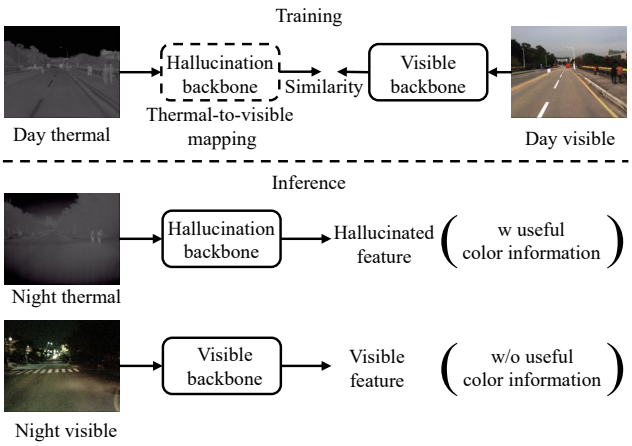


Figure 2. **Illustration of the modality hallucination technique.** Examples of day and night thermal-visible image pairs are shown. In the day image pair, pedestrians are easily recognized on both thermal and visible images. In this way, we can learn good thermal-to-visible mapping through the hallucination network. In the night image pair, we can still generate useful visible features (*i.e.*, hallucinated features) using the learned mapping.

modal integration. For instance, previous methods usually assign lower fusion weights to bad visible features to reduce their impact on the fused features. However, visible images at night usually contain no meaningful color information for pedestrian detection, as illustrated in Fig. 2. Even with a very effective fusion approach, the overall detection performance is still adversely impacted by darkness and other challenging illumination conditions (*e.g.*, overexposure).

Instead of mitigating the negative impact of meaningless visible features, we propose to actively generate good visible-like features from thermal images, since thermal images are always informative for pedestrian detection. Based on this observation, we advocate using the modality hallucination technique proposed in [13] to generate visible-like features (*i.e.*, hallucinated features) from thermal image inputs. The generated visible-like features can then be fused with thermal features and improve the detection performance when visible images are uninformative under poor visibility conditions. As illustrated in Fig. 2, the basic idea is that we can borrow color information from good visible-good thermal image pairs to improve detection performance for bad visible-good thermal image pairs. Through the hallucination network, we first learn a good mapping from the thermal to the visible domain. In the testing stage, we then use the learned thermal-to-visible mapping to generate relatively good visible-like features from inputs of good thermal images. The generated hallucinated feature can be seen as complementary color information for uninformative visible images. With the help of the relatively good visible-like features, our network can achieve better detection performance than conventional two-branch fusion methods which use bad visible features, as shown in Fig. 1.

To this end, we design a novel fusion architecture, which is simple yet effective. First, we have two backbones for feature extraction from thermal and visible images separately, like most of the previous methods. In addition, we introduce another backbone, the hallucination network, to learn the thermal-to-visible mapping. This hallucination backbone and the visible backbone are further connected using feature similarity losses (*i.e.*, the hallucination losses). We can thus train the network so that the hallucinated feature maps look similar to the visible feature maps. Finally, a multi-modality fusion module is introduced to adaptively combine feature maps from all three branches.

The key issue here is how we can learn a good thermal-to-visible mapping function through the hallucination loss between the hallucination backbone and the visible backbone. The basic idea is to force the hallucinated feature maps to mimic the visible feature maps as much as possible by narrowing the distance between these two kinds of feature maps. While we found thermal images are usually in good conditions regardless of illumination changes, visible images excel during the day and are usually in bad conditions at night. For night-time visible images, color information is lost, and thus the thermal-to-visible mapping will be misled by still forcing the hallucination feature maps to mimic the visible feature maps that contain less or no useful color information. This problem caused by the difference between good-condition and bad-condition visible images is defined as the domain inconsistency problem. To tackle this problem, we further design an illumination-aware hallucination loss to prevent information transfer from bad-condition visible images. That is, we propose to assign lower weights to the hallucination loss when visible images are in bad condition.

In summary, the contributions of this work are threefold:

- We go beyond the limitation of previous fusion methods and propose a modality hallucination-based multispectral fusion network comprising three feature extraction branches (*i.e.*, thermal, hallucination, and visible branches).
- We present a Hierarchical Multi-modal Feature Fusion module to effectively combine features from the three branches, and an illumination-aware hallucination loss to relieve the domain inconsistency problem.
- Extensive experiments demonstrate the benefits of hallucination-based three-branches fusion architecture. The proposed network outperforms conventional fusion methods on both KAIST and CVC-14 datasets.

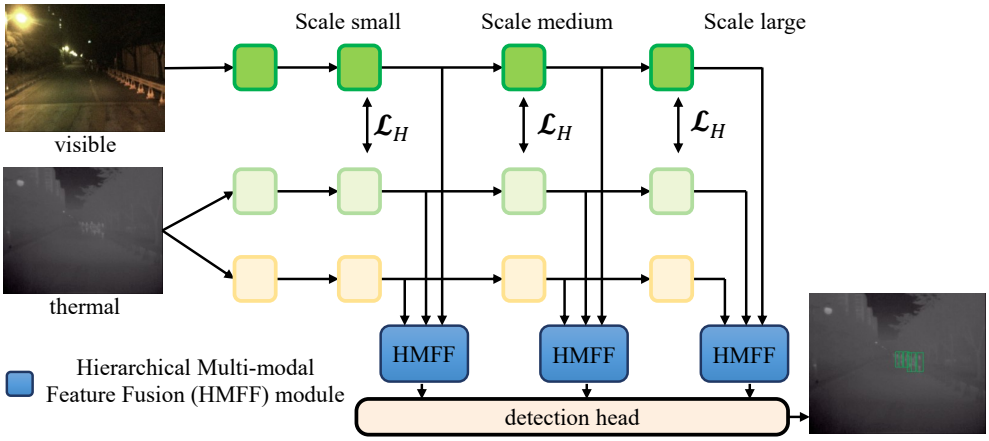


Figure 3. **The overall architecture of our proposed multi-modal fusion network for pedestrian detection.** A hallucination backbone is added to the traditional two-stream fusion network, and a Hierarchical Multi-modal Feature Fusion (HMFF) module is further designed to fuse features from all three branches at multiple scales.

2. Related work

2.1. Multispectral Fusion for Pedestrian Detection

Most studies on multispectral pedestrian detection have focused on developing two-branches architectures extracting and fusing information from visible and thermal images [3, 4, 17, 23, 24, 33, 34, 39].

Li *et al*. [23] presented a two-branches pedestrian detector based on Faster R-CNN [28]. Two VGG-16 [30] backbones are used to extract features from color and thermal images separately. To fuse the detection results from two modalities more effectively, they proposed an illumination-aware network as a side branch to estimate the illumination value from visible images and an illumination-aware weighting layer to get the fusion weights for fusion. Chen *et al*. [5] proposed a probabilistic ensembling technique to smartly fuse detection results from two modalities. Instead of fusing detection results, most of the related works focus on designing mid-feature integration modules. For instance, Zhou *et al*. [39] designed a Modality Balance Network (MBNet) to tackle the modality imbalance problems when fusing visible and thermal features. An illumination-aware feature alignment module was also proposed to address the misalignments between visible and thermal modalities according to the illumination conditions. Zhang *et al*. [33] designed a Cyclic Fuse-and-Refine module to iteratively refine separate modality features using fused features. They also utilized an auxiliary segmentation task to better learn each refined modality feature. To effectively fuse features from those modalities that are not fully aligned, Kim *et al*. [17] leveraged the multi-label learning strategy to learn more discriminative features even when one domain of the input data has some problems. Zhang *et al*. [34] proposed the inter- and intra-modality attention modules to improve the modality feature fusion efficiency under the guidance of ground truth segmentation masks.

Unlike the above research working on improving feature integration efficiency from only two modalities, we propose to extract color information from good-condition visible image input and then generate more color information for bad-condition visible image input, by adding a hallucinated medi-modality (thermal-to-visible) as the third modality. To the best of our knowledge, our model is the first work leveraging the hallucination mechanism to improve the fusion efficiency for multispectral pedestrian detection.

2.2. Modality Hallucination

The concept of modality hallucination is first presented in [13]. Taking into the fact that depth information can contribute a lot to object detection but sometimes is not readily available, Hoffman *et al*. [13] designed a modality hallucination network to produce depth-related features from the input RGB images by mimicking the real depth mid-level features at training time. The learned depth-related features can then be combined with RGB features to boost the object detection performance at testing time when only RGB images are taken as input. Since then, the modality hallucination mechanism has been widely applied to various tasks, such as hand pose estimation [6], video action classification [10], object detection in indoor scenes [38], visual odometry [29] and much more [21, 32]. As a follow-up, Jiao *et al*. [16] also proposed a two-branch network to jointly learn semantic and geometry (*i.e*., depth) information from only RGB images for the semantic segmentation task. Crasto *et al*. [8] designed a network to hallucinate motion features from RGB frames for the action recognition task. Recently, Saputra *et al*. [29] utilized a visual hallucination network to predict fake RGB latent features from

thermal images, and then employed selective fusion to combine features from thermal, hallucination, and inertial features to finally perform thermal-inertial odometry.

Even though the big idea of mimicking one modality feature from another modality input in this paper is similar to the above methods, they did not consider the domain inconsistency problem in these methods. In this paper, instead of simply reducing the distance between features from two modalities like these methods, we further explore how to effectively learn the mapping function between two modalities under the problem of domain inconsistency.

3. Methods

The proposed network architecture is illustrated in Fig. 3. Our network takes a pair of visible image I_V and thermal image I_T as input, and outputs pedestrian detection bounding boxes. The input images are first sent to the backbone component (Sec. 3.1) for multi-modality feature extraction. Different from traditional multi-modality fusion networks, a thermal-to-visible hallucination feature extraction branch (Sec. 3.2) is added to the feature extraction component, forming a three-branch feature extraction backbone. Then, a Hierarchical Multi-modal Feature Fusion (HMFF) module (Sec. 3.3) is presented to adaptively combine features from the three branches. The fused feature is fed into the detection head to generate detection results. Finally, the designed hallucination loss is detailed in Sec. 3.4.

3.1. Backbone

Our network is built upon the state-of-the-art detector, YOLOv7 [31] because of its recent excellent performance in both detection accuracy and inference speed. Before the feature fusion, there are three feature extraction branches. The upper branch is responsible for visible features F_V extraction and the bottom branch is responsible for thermal features F_T extraction. As for the middle branch, it takes thermal images as input and outputs hallucinated features F_H. The hallucination loss between the visible branch and the middle branch (*i.e.*, the hallucination branch) makes the hallucinated features mimic the visible features from the input thermal images. In fact, the hallucination branch can be seen as a thermal-to-visible mapping function that transfers thermal modality input into the visible modality feature space. For each feature extraction branch, we use the same backbone architecture in YOLOv7. More architecture details can be found in [31].

3.2. Thermal to Visible Hallucination

The hallucinated feature extraction branch aims to learn a mapping function $f(\cdot)$, which can map the thermal image input to the visible feature space. In such a way, we can then generate visible-like features (called hallucinated features F_H) from the input thermal images I_T when the mapping function is learned well. To achieve this, we propose to establish hallucination connections between the visible and hallucination branches. Specifically, three hallucination connections are added in the three outputs of backbone networks, which are responsible for detecting large, medium, and small pedestrian targets respectively, as illustrated in Figure 3. Within the hallucination connections, a hallucination loss is introduced to measure the similarity between visible features F_V and hallucinated feature F_H, which is described in detail in Sec. 3.4. By iteratively decreasing the hallucination loss during the training stage, the hallucination branch can then gradually generate visible-like features from input thermal images.

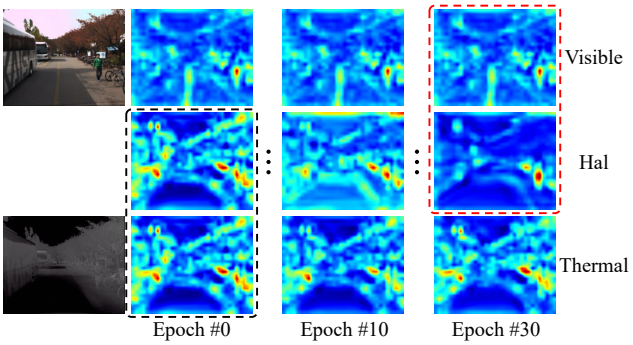


Figure 4. **Illustration of how the hallucinated feature map mimics the visible feature map.** As the training epoch goes on, the hallucinated feature map looks more and more similar to the well-lighted visible feature map. Visible feature maps stay consistent over epochs because the visible backbone is fixed during the last training stage, as detailed in Sec. 4.1.

To have an intuitive understanding of the hallucination procedure, we further visualize a set of feature maps to show the change of hallucinated features as the training process deepens. As shown in Figure 4, the hallucinated feature map looks more like to the thermal feature map at the beginning (see the black dotted-line box). However, as the training goes on, the hallucinated feature map becomes more similar to the visible feature map (see the red dotted-line box) than to the thermal feature map.

3.3. Hierarchical Multi-modal Feature Fusion

The features from the three branches need to be fused before being fed to the detection head module. Unlike most existing visible-thermal pedestrian detection methods which have only features from two branches to be fused, we have three streams of features to deal with. To make the fusion module easily learn the relationship between each modality, we propose a Hierarchical Multi-modal Feature Fusion (HMFF) module to gradually fuse features from the three branches in a two-step fashion, as illustrated in Fig. 5. Two different fusing strategies are employed according to

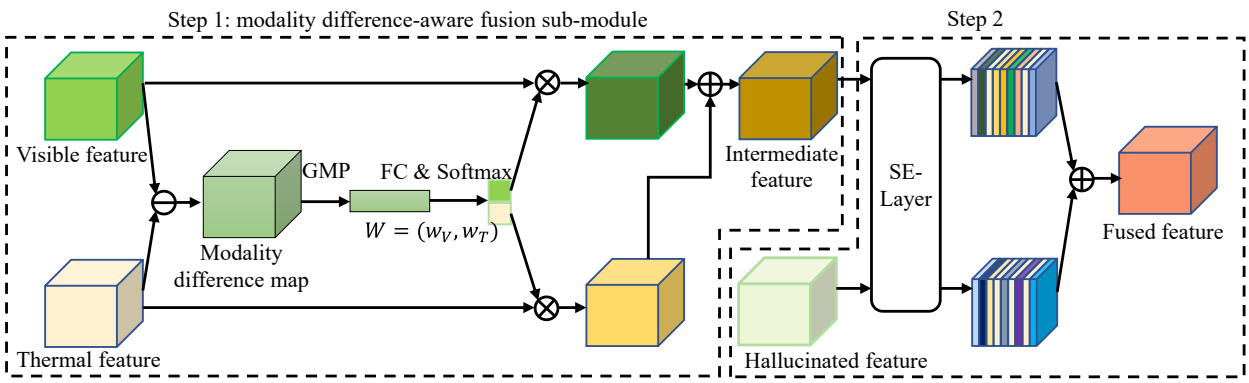


Figure 5. **Architecture of the proposed HMFF module**, which first combines visible and thermal features by the proposed modality difference-aware fusion sub-module and then integrates the combined feature with the hallucinated feature to obtain the final fused feature. GMP: global max pooling; FC: fully connected layer. SE-Layer: attention layer from [14].

the relationship between modalities in the two steps.

Specifically, inspired by the idea of assigning different fusion weights to different modality features under different conditions in many other multi-modality fusion works, we design a modality difference-aware fusion sub-module for the feature integration from the visible and thermal branches in the first step. The combined feature of the first step and the hallucinated feature is fed into a SE-Layer before finally being added to form the fused feature.

In the modality difference-aware fusion sub-module, let $W^l = (w_V^l, w_T^l)$ be the weighting vector of visible and thermal branches in scale l (*i.e.*, small, medium, and large in Fig. 3). As in Fig. 5, this weighting vector is obtained from the modality difference map between visible feature F_V and thermal feature F_T. The Global Max Pooling (GMP) operation is then performed on this modality difference map to get a modality difference-aware vector, which is finally transferred to a two-dimensional weight vector via a fully connected (FC) layer with a Softmax activation function. Overall, the fusion weight calculation procedure can be formulated as:

$$W^l = Softmax(FC(GMP(F_V^l - F_T^l))). \quad (1)$$

The fused feature in the first step is the weighted sum of features from the visible and thermal using W^l. The intuition of this weighting strategy is that the modality weights should be intrinsically related to the difference between the two modalities. That is, when the features from the two modalities are similar to each other, the fusion weights of these two modalities do have not much impact on the detection results. In contrast, when the difference between the two modalities is large, it is crucial to determine which modality is more reliable.

After the first fusion step, we can now ensure that the thermal-visible combined feature and the hallucinated feature are both informative. Then, we can perform a further fine-grained fusion strategy in the second step. To achieve this, we employ the SE-Layer to first channel-wisely enhance the two features and then add them to get the final fused feature. This fused feature is then fed to the detection head to get the final detection results.

There certainly exist more efficient and elaborate ways to design the fusion strategy in the second step. However, the core idea in HMFF is the first-time application of different fusion strategies for specific characteristics of different modalities in a hierarchical way. We encourage the community to further extend our current method.

3.4. Hallucination Loss

During the training stage, we simultaneously predict bounding boxes and perform modality hallucination by minimizing the detection loss $\mathcal{L}_D$ and the hallucination loss $\mathcal{L}_H$. The goal of hallucination loss is to make the hallucinated feature F_H mimic the visible feature F_V. We adopt the ℓ_1-loss as the basic hallucination loss first. Furthermore, we found that visible images captured at night are usually in bad conditions, containing less meaningful color information, as shown in Fig. 2. In that case, it is meaningless and would even bring noise into the mapping function learning procedure if we still enforce the hallucinated feature to strongly mimic the visible feature. Thus, to avoid negative feature hallucination, we propose an illumination-aware loss which assigns a lower weight to the hallucination loss of images captured at night. Moreover, to increase the hallucination efficiency, we enforce the hallucination procedure to focus on the pedestrian targets by further introducing a mask weight w_m to the hallucination loss. Overall, the final hallucination loss is defined as a weighted ℓ_1-loss, which is formulated as:

$$\mathcal{L}_H = w_i \cdot \sum_{j=1}^{n} w_m^j \cdot \left\| f_V^{ij} - f_H^{ij} \right\|, \quad (2)$$

where n is the number of pixels f_V^i, f_H^i in feature maps F_V^i, F_H^i for i-th image pair. w_m^j is 1 if the j-th feature map pixel f_V^{ij} is within the ground truth boxes, whereas, w_m^j is 0. w_i represents the illumination weight for i-th image pair, which is adaptively generated based on the illumination conditions. To achieve this, we simply adopt a two-layer MLP to predict the day or night classification of visible images using the final output vector from the visible branch. w_i is then set to be the day classification score, which is in the range of $[0, 1]$. In that way, day image pairs would get higher illumination weights than night image pairs. That is, the dissimilarity between visible and hallucinated features in good-light conditions is penalized more than dissimilarity in bad-light conditions. We set the day weight large than the night weight to encourage the hallucination network to focus on mimicking good visible features.

The overall loss is then the sum of the two losses:

$$\mathcal{L}_{total} = \mathcal{L}_D + \lambda \mathcal{L}_H, \tag{3}$$

where λ is used to scale the hallucination loss to the same scale as the detection loss, which is set to 10.0 in our experiments. For the detection loss, we just use the same loss as YOLOv7, which consists of three terms: regression loss, objectness loss, and classification loss.

4. Experiments

4.1. Experimental Setup

Datasets. The KAIST [15] dataset consists of 7,601 and 2,252 well-aligned thermal and visible image pairs for training and testing, respectively. For fair comparisons, we follow previous work [17] and use the train dataset annotations from [36] and sanitized test dataset annotations from [22]. There are 1,455 daytime images and 797 nighttime images in the test set. Note that some papers follow the raw KAIST dataset comprising 95,328 images [3,4], which contains imperfect annotations. For fairness, we only compare with methods trained using the cleaned 7,601 image pairs. The CVC-14 [11] dataset also encompasses both visible and thermal images captured in the driving environments at day and night time. We use the same training and testing division as in [18, 36], 7,085 for training and 1,433 for testing.

Metrics. The widely used miss rate (MR) averaged over the false positive per image (FPPI) with the range of $[10^{-2}, 10^{0}]$ in [9] is adopted as our evaluation metric. With this metric, lower values represent better detection performance. Following recent methods, we also separately measure miss rates for day (MR_Day) and night (MR_Night) images and then report miss rates for all images (MR_All).

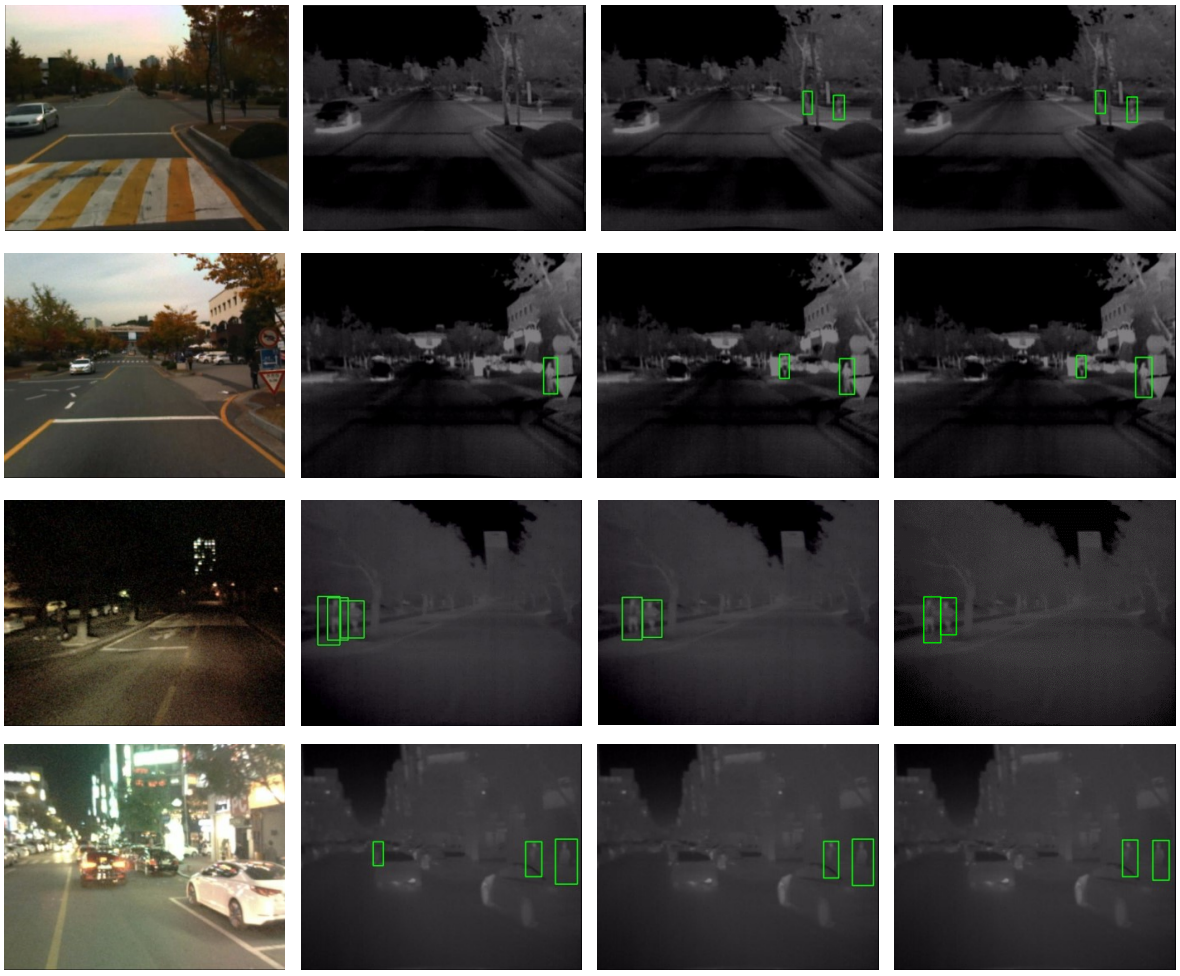
Training. The training procedure is divided into three stages in our paper. In the first stage, we train two one-branch networks for thermal and visible images respectively. In the second stage, we train a two-branch fusion network. This two-branch network has two backbones to separately extract thermal and visible features. Moreover, for this two-branch network, the feature fusion module is modified from the HMFF module by deleting the second-stage fusion step. In this training stage, we use the backbone weights trained in the first stage to initialize the two backbones here. The weights of the detection head are initialized by the weights of its counterpart in the thermal one-branch network from the first stage. The fusion module is initialized using the He initialization technique [12]. In the last training stage, we train the complete network proposed in our paper using the network weights obtained in the second stage. Specifically, compared to the network in the second stage, a hallucination branch is added to the complete network, and weights in the hallucination backbone are initialized using weights of the visible backbone in the second stage. This initialization strategy can guarantee our network a better detection performance, as demonstrated in the ablation study (Sec. 4.3). Moreover, the weights in the visible backbone are fixed at this stage.

Methods	Day	Night	All
CAIN [35]	14.77	11.13	14.12
MSDS-RCNN [22]	10.53	12.94	11.34
AR-CNN [36]	9.94	8.38	9.34
MBNet [39]	8.28	7.86	8.13
UGC [19]	8.18	6.96	7.89
MLPD [17]	7.95	6.95	7.58
ProbEn [5]	9.93	5.41	8.50
ProbEn_3 [5]	9.07	4.89	7.66
Beyond Fusion (Ours)	**5.89**	**3.27**	**5.01**

Table 1. Pedestrian detection results on KAIST dataset in terms of missing rate (MR).

4.2. Comparison with State of the Art

We present our pedestrian detection results on the KAIST dataset in Tab. 1. For this dataset, the proposed model achieves the best detection performance in terms of MR_All, 5.01. For the CVC-14 dataset, our method also achieves the leading performance, as shown in Tab. 2. Except for quantitative results, we also report some visual comparisons to demonstrate the superiority of the proposed method. Fig. 6 gives several qualitative results of our method. As observed, our method can generate more accurate bounding boxes and correctly detect pedestrians under difficult situations compared to the baseline model. Note that these shown cases in Fig. 6 are all in bad light conditions. It is difficult to rely on these visible images to detect pedestrians in these images. However, the hallucination branch in our method can still generate useful color information from thermal image inputs. Thus, the detection performance is boosted compared to the baseline model without the hallucination mechanism.



(a) Visible image (b) Baseline (c) Ours (d) Ground truth

Figure 6. **Visual comparison of the baseline and our approach.** We delete the hallucination backbone to create the baseline model. As seen, our method has better detection performance when visible images are in bad conditions thanks to the visible-like information provided by the hallucination branch.

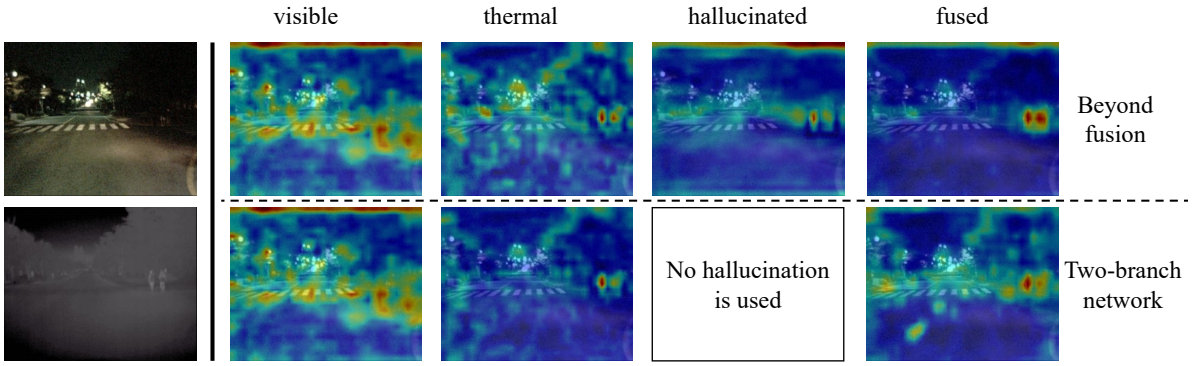


Figure 7. **Comparison visualization of feature maps with and without hallucination.** With the hallucinated feature map, the background interference, brought by the bad-condition visible image, is significantly attenuated and the fused feature map is more focused on pedestrians.

Methods	Day	Night	All
MACF [27]	72.63	65.43	69.71
Choi et al. [7]	63.39	63.99	63.34
Halfway Fusion [27]	36.29	26.29	31.99
Park et al. [27]	28.67	23.48	26.29
AR-CNN [36]	24.7	18.1	22.1
MBNet [39]	24.7	13.5	21.1
MLPD [17]	24.18	17.97	21.33
LG-FAPF [3]	22.5	12.2	18.2
Beyond Fusion (Ours)	**20.18**	**9.86**	**16.62**

Table 2. Pedestrian detection results on CVC-14 in terms of MR.

methods		Day	Night	All
one-branch	visible	18.88	23.83	20.68
	thermal	25.12	8.14	19.23
two-branch	visible+thermal	11.12	6.88	9.87
three-branch	w/o $\mathcal{L}_H$	10.70	6.09	9.28
	w $\mathcal{L}_H$(w/o weights)	7.40	5.34	6.91
	w $\mathcal{L}_H$(w weights)	5.89	3.27	5.01

Table 3. Analysis of the effectiveness of the proposed hallucination backbone and the illumination-aware hallucination loss.

4.3. Ablation study

To verify the effectiveness of hallucination, we design another network (w/o $\mathcal{L}_H$ in Tab. 3) by deleting all the connections between hallucination and visible branches. For reference, we also create a two-branch baseline network by deleting the whole hallucination branch, and we also report the results of two one-branch networks which take only single modalities as input. We train the new networks under the exact same super-parameter and weight initialization settings. The comparison results are given in Tab. 3.

w *vs* w/o hallucination. As shown, two- and three-branch networks undoubtedly outperform one-branch networks since thermal+visible image pairs provide more information. Moreover, thanks to more parameters and a wider network, three-branch networks outperform the two-branch network. For three-branch networks, the network without hallucination loss $\mathcal{L}_H$ has the same architecture and number of parameters as the proposed network. However, the w/o $\mathcal{L}_H$ network underperforms both of the networks with $\mathcal{L}_H$, which can demonstrate the effectiveness of the hallucination connections.

w *vs* w/o weights in $\mathcal{L}_H$. Thanks to the weighting strategy in Equation 2, the proposed illumination-aware hallucination loss also contributes to performance improvement by relieving the domain inconsistency problem. In the last two rows, both of these two networks have hallucination connections. However, the network with weights in $\mathcal{L}_H$ outperforms the network without weights by 1.9, verifying that the proposed illumination-aware hallucination loss can further boost the detection performance.

Fusion visualization w and w/o hallucination. Apart from the quantitative results, we visualize feature maps to demonstrate the effectiveness of the hallucinated features.

	Day	Night	All
Add	8.41	5.38	7.22
Concat	8.35	5.27	6.87
HMFF	5.89	3.27	5.01

Table 4. Analysis of the effectiveness of HMFF module.

As in Fig. 7, it is difficult to recognize two pedestrians in the visible image, owing to the low illumination condition. The corresponding visible feature map also contains no meaningful activation, while the thermal and hallucinated feature maps look more meaningful. Moreover, in the first row, the fused feature map with hallucination is much better than without hallucination in the second row.

w *vs* w/o HMFF. The proposed HMFF module plays a key role in the performance boost. To verify this, we compare our HMFF with the other two feature fusion strategies, *Add* and *Concat*. The *Add* operation simply adds three feature maps in an element-wise way, and the *Concat* represents first concatenating all the three features and then applying a 1×1 Conv operation to restore the dimension of the original features. As shown in Tab. 4, the proposed fusion strategy outperforms the comparing strategies, demonstrating the superiority of the HMFF module.

5. Conclusions

In this paper, we propose a novel multi-modal fusion architecture for pedestrian detection by leveraging the modality hallucination mechanism. Instead of simply combining thermal and visible features using a two-branch architecture like most previous works, a novel three-branch network is designed. In this way, the detection performance can be boosted by generating good visible features through the added hallucination backbone when visible images are in bad conditions. Moreover, an HMFF module is proposed to selectively combine features from three different modalities, *i.e.*, thermal, hallucination, and visible features. An illumination-aware hallucination loss is further presented to avoid negative modality hallucination by assigning lower weights to the hallucination losses of night image pairs.

Limitations. One assumption here is that thermal images are in good condition so that useful hallucinated visible features can be generated from thermal images. However, thermal images could be contaminated by temperature noise. In that way, generated hallucinated features would also be affected. Hence, one limitation is that our method relies on thermal images so much that the hallucination procedure would fail to generate reasonable visible-like features when input thermal images are also in bad condition.

Acknowledgement. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Program "ACE-OPS: From Autonomy to Cognitive assistance in Emergency OPerationS" under Grant EP/S030832/1.

References

[1] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *European Conference on Computer Vision*, pages 613–627. Springer, 2014. 1

[2] Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta, and Vitoantonio Bevilacqua. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, 300:17–33, 2018. 1

[3] Yanpeng Cao, Xing Luo, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection. *Information Fusion*, 88:1–11, 2022. 3, 6, 8

[4] Yi-Ting Chen, Jinghao Shi, Christoph Mertz, Shu Kong, and Deva Ramanan. Multimodal object detection via bayesian fusion. *arXiv preprint arXiv:2104.02904*, 2021. 3, 6

[5] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 139–158. Springer, 2022. 3, 6

[6] Chiho Choi, Sangpil Kim, and Karthik Ramani. Learning hand articulations by hallucinating heat distribution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3104–3113, 2017. 3

[7] Hangil Choi, Seungryong Kim, Kihong Park, and Kwanghoon Sohn. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 621–626. IEEE, 2016. 8

[8] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019. 3

[9] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011. 6

[10] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Learning with privileged information via adversarial discriminative modality distillation. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2581–2593, 2019. 3

[11] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M López. Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, 16(6):820, 2016. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 6

[13] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 826–834, 2016. 2, 3

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[15] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 6

[16] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson WH Lau, and Thomas S Huang. Geometry-aware distillation for indoor semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2869–2878, 2019. 3

[17] Jiwon Kim, Hyeongjun Kim, Taejoo Kim, Namil Kim, and Yukyung Choi. Mlpd: Multi-label pedestrian detector in multispectral domain. *IEEE Robotics and Automation Letters*, 6(4):7846–7853, 2021. 1, 3, 6, 8

[18] Jung Uk Kim, Sungjune Park, and Yong Man Ro. Robust small-scale pedestrian detection with cued recall via memory learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3050–3059, 2021. 6

[19] Jung Uk Kim, Sungjune Park, and Yong Man Ro. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1510–1523, 2021. 6

[20] Wenbo Lan, Jianwu Dang, Yangping Wang, and Song Wang. Pedestrian detection based on yolo network model. In *2018 IEEE international conference on mechatronics and automation (ICMA)*, pages 1547–1551. IEEE, 2018. 1

[21] Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. Learning with privileged information for efficient image super-resolution. In *European Conference on Computer Vision*, pages 465–482. Springer, 2020. 3

[22] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference (BMVC)*, September 2018. 6

[23] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019. 3

[24] Qing Li, Changqing Zhang, Qinghua Hu, Huazhu Fu, and Pengfei Zhu. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 2022. 3

[25] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5187–5196, 2019. 1

[26] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF inter-*

national conference on computer vision, pages 4967–4975, 2019. 1

[27] Kihong Park, Seungryong Kim, and Kwanghoon Sohn. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognition*, 80:143–155, 2018. 8

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[29] Muhamad Risqi U Saputra, Pedro PB de Gusmao, Chris Xiaoxuan Lu, Yasin Almalioglu, Stefano Rosa, Changhao Chen, Johan Wahlström, Wei Wang, Andrew Markham, and Niki Trigoni. Deeptio: A deep thermal-inertial odometry with visual hallucination. *IEEE Robotics and Automation Letters*, 5(2):1672–1679, 2020. 3

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[31] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 4

[32] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5363–5371, 2017. 3

[33] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 276–280. IEEE, 2020. 3

[34] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 72–80, 2021. 3

[35] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019. 6

[36] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5127–5137, 2019. 6, 8

[37] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018. 1

[38] Zhijie Zhang, Yan Liu, Junjie Chen, Li Niu, and Liqing Zhang. Depth privileged object detection in indoor scenes via deformation hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3456–3464, 2021. 3

[39] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *European Conference on Computer Vision*, pages 787–803. Springer, 2020. 1, 3, 6, 8