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.

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 $L_H$ 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-
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 three-fold:

- 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.
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 testing 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.

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
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 \( W^l = (w^l_V, w^l_T) \) 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:

\[
W^l = \text{Softmax}(FC(GMP(F^l_V - F^l_T))).
\] (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 \( \ell_1 \)-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 \( \ell_1 \)-loss, which is formulated as:

\[
\mathcal{L}_H = w_1 \cdot \sum_{j=1}^{n} w_m^j \cdot \| f^j_V - f^j_H \|_1,
\] (2)
where \( n \) is the number of pixels \( f^i_V, f^i_H \) in feature maps \( F^i_V, F^i_H \) for \( i \)-th image pair. \( w^i_m \) is 1 if the \( j \)-th feature map pixel \( f^m_{i,j} \) is within the ground truth boxes, whereas, \( w^i_v \) 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:

\[
L_{total} = L_D + \lambda L_H, \tag{3}
\]

where \( \lambda \) 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.

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.
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.
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.

**4.3. Ablation study**

To verify the effectiveness of hallucination, we design another network (w/o $L_H$ 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 $L_H$ has the same architecture and number of parameters as the proposed network. However, the w/o $L_H$ network underperforms both of the networks with $L_H$, which can demonstrate the effectiveness of the hallucination connections.

**w vs w/o weights in $L_H$.** 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 $L_H$ 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.

**5. Conclusions**

As in Fig. 7, it is difficult to recognize two pedestrians in the visible image, owing to the low illumination condition. The corresponding 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 $1 \times 1$ 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.

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


