# Guided Attentive Feature Fusion for Multispectral Pedestrian Detection

Heng ZHANG[1,3], Elisa FROMONT[1,4], Sébastien LEFEVRE[2] and Bruno AVIGNON[3]
[1]Univ Rennes 1, IRISA, France [2]Univ Bretagne Sud, IRISA, France
[3]ATERMES company, France [4]IUF, Inria.

## Abstract

*Multispectral image pairs can provide complementary visual information, making pedestrian detection systems more robust and reliable. To benefit from both RGB and thermal IR modalities, we introduce a novel attentive multispectral feature fusion approach. Under the guidance of the inter- and intra-modality attention modules, our deep learning architecture learns to dynamically weigh and fuse the multispectral features. Experiments on two public multispectral object detection datasets demonstrate that the proposed approach significantly improves the detection accuracy at a low computation cost.*

## 1. Introduction

Real world pedestrian detection applications require accurate detection performance under various conditions, such as darkness, rain, fog, etc. In these conditions, it is difficult to perform precise detection using only standard RGB cameras. Instead, multispectral systems try to combine the information coming from e.g. thermal and visible cameras to improve the reliability of the detections.

Deep learning-based methods, more specifically, two-stream convolutional neural networks, nowadays largely dominate the field of multispectral pedestrian detection [6, 9, 10, 11, 14, 18, 19, 20]. As illustrated in Fig. 1, a typical two-stream pedestrian detection network consists of two separate spectra-specific feature extraction branches, a multispectral feature fusion module and a pedestrian detection network operating on the fused features. The system uses some aligned thermal-visible image pairs as input and outputs the joint detection results on each image pair.

Thermal and visible cameras have different imaging characteristics under different conditions. As shown in Fig. 2, visible cameras provide precise visual details (such as color and texture) in a well-lit environment, while thermal cameras are sensitive to temperature changes, which is extremely useful for nighttime or shadow detection. An adaptive fusion of thermal and visible features should take such differences into account, and should identify and leverage the information from the most relevant modality.

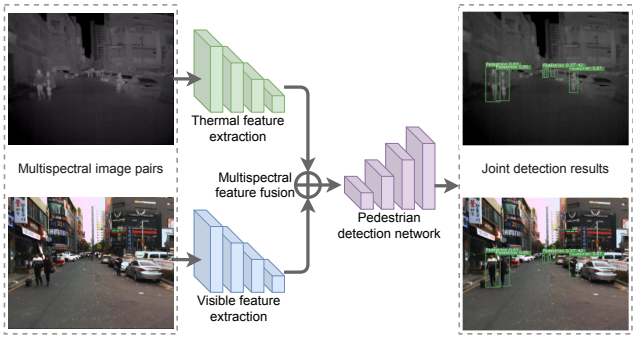



Figure 1: Multispectral pedestrian detection via a two-stream convolutional neural network.

Figure 2: Typical examples of thermal-visible image pairs captured during the day (first two rows) and night (bottom row). For each pair, the thermal image is on the left and the RGB image is on the right.

An intuitive solution to adapt the feature fusion to the different weather and lighting conditions is to manually identify multiple usage scenarios and design a specific solution for each scenario. For example, [6] proposes an illumination-aware network consisting of a day illumination sub-network and a night illumination sub-network. The

detection results from the two sub-networks are then fused according to the prediction of the illumination context. Such a kind of hand-crafted fusion mechanism improves the resilience of the model to a certain extent, nonetheless, there are still two limitations: firstly, cherry-picked scenarios may not cover all conditions, e.g., different illumination/season/weather conditions; Secondly, the situation may be completely different even in the same usage scenario, e.g., at nighttime, lighting conditions in urban areas are different from those in rural areas.

In this paper, we propose a novel and fully adaptive multispectral feature fusion approach, named *Guided Attentive Feature Fusion* (GAFF). By combining the intra- and inter-modality attention modules, the proposed approach allows the network to learn the adaptive weighing and fusion of multispectral features. These two attention mechanisms are guided by the prediction and comparison of the pedestrian masks in the multispectral feature fusion stage. Specifically, at each spatial position, thermal or visible features are enhanced when they are located in the area of a pedestrian (intra-modality attention) or when they possess a higher quality than in the other modality (inter-modality attention). To the best of our knowledge, GAFF is the first work that regards the multispectral feature fusion as a sub-task in the network optimization and that introduces a specific guidance in this task to improve the multispectral pedestrian detection. Extensive experiments on KAIST multispectral pedestrian detection dataset [8] and FLIR ADAS dataset [1] demonstrate that, compared with common feature fusion methods (such as addition or concatenation), GAFF brings important accuracy gains at a low computational cost.

This paper is organized as follows: Section 2 reviews some representative work applying static/adaptive feature fusion for multispectral pedestrian detection; Section 3 introduces implementation details on how to integrate GAFF into a typical two-stream convolutional neural network; In Section 4, we evaluate our methods on two public multispectral object detection datasets [8, 1], then we provide an extensive ablation study and visualization results to discuss the reasons of the accuracy improvements; Section 5 concludes the paper.

## 2. Related Work

### 2.1. Static multispectral feature fusion

KAIST released the first large-scale multispectral pedestrian detection dataset [8], which contains approximately 95k well-aligned and manually annotated thermal-visible image pairs captured during daytime and nighttime. Some example image pairs are shown in Fig. 2. Then [18] demonstrated the first application of deep learning-based solutions in multispectral pedestrian detection. They compared the early and late fusion architectures and found that the late fusion architecture is superior to the early one and the traditional ACF method [4]. This late-stage fusion architecture can be regarded as a prototype of a two-stream neural network, in which multispectral features are fused through concatenation operations. Both [14] and [9] adapted Faster R-CNN [16] to a two-stream network architecture for multispectral pedestrian detection. They compare different multispectral fusion stages and came to the conclusion that the fusion in the middle stage outperforms the fusion in the early or late stage. Based on this, MSDS-RCNN [10] adopted a two-stream middle-level fusion architecture and combined the pedestrian detection task and the semantic segmentation task to further improve the detection accuracy.

### 2.2. Adaptive multispectral feature fusion

As mentioned in Section 1, thermal and visible cameras have different imaging characteristics and the adaptive multispectral fusion can improve the resilience and the detection accuracy of the system. This has become the main focus of the multispectral pedestrian detection research in recent years. Both [11] and [6] use the illumination information as a clue for the adaptive fusion: they train a separate network to estimate the illumination value from a given image pair, then [11] uses the predicted illumination value to weigh the detection results from both the thermal and visible images. [6] uses the illumination value to weigh the detection results from a day illumination sub-network and a night illumination sub-network. As mentioned in the previous section, such a handcrafted weighing scheme is limited and produces sub-optimal performance. CIAN [20] applies the channel-level attention in the multispectral feature fusion stage to model the cross-modality interaction and weigh each feature map extracted from the different spectrum. This network realizes a fully adaptive fusion of thermal and visible features, however, in this approach, the fusion module is optimized directly while solving the pedestrian detection task which means that the network uses information about what (pedestrian or background) and where (bounding box) relevant elements are in the images but it does not use the fact that some features may contain more relevant information than others. We believe and we show that with these additional information (that we include in our method through the guidance mechanism), we can improve the detection precision.

## 3. Proposed approach

The proposed *Guided Attentive Feature Fusion* (GAFF), shown in Fig. 3, takes place in the multispectral feature fusion stage of a two-stream convolutional neural network. It consists of two components: an intra-modality attention module and an inter-modality one.

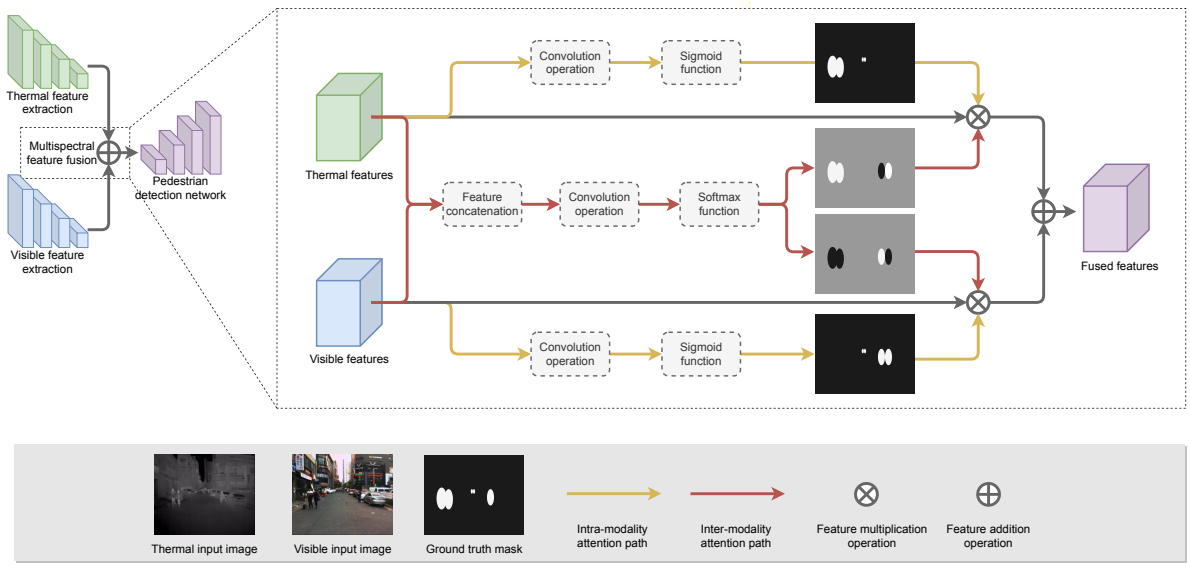



Figure 3: The overall architecture of *Guided Attentive Feature Fusion* (GAFF). Green, blue and purple blocks represent thermal, visible and fused features. Yellow and red paths represent the intra- and inter-modality attention modules, respectively.

### 3.1. Intra-modality attention module

The intra-modality attention module aims at enhancing the thermal or visible features in a monospectral view. Specifically, as illustrated by the yellow paths on Fig. 3, features of an area with a pedestrian are highlighted by multiplying the learnt features with the predicted pedestrian mask. Moreover, in order to avoid directly affecting the thermal or visible features, the highlighted features are added as a residual to enhance the mono-spectral features. This procedure can be formalized as:

$$\begin{aligned} f^t_{intra} &= f^t \otimes (1 + m^t_{intra}) \\ f^v_{intra} &= f^v \otimes (1 + m^v_{intra}) \end{aligned} \tag{1}$$

where

$$\begin{aligned} m^t_{intra} &= \sigma(\mathcal{F}^t_{intra}(f^t)) \\ m^v_{intra} &= \sigma(\mathcal{F}^v_{intra}(f^v)) \end{aligned} \tag{2}$$

Superscripts ($t$ or $v$) denote the thermal ($t$) or visible ($v$) modality; $\otimes$ denotes the element-wise multiplication; $\sigma$ represents the sigmoid function; $\mathcal{F}_{intra}$ represents a convolution operation to predict the intra-modality attention masks (pedestrian masks) $m_{intra}$; $f$ and $f_{intra}$ represent the original and enhanced features, respectively.

The prediction of the pedestrian masks is supervised by the semantic segmentation loss, where the ground truth mask ($m^{gt}_{intra}$) is converted from the object detection annotations. As illustrated in Fig. 3 the bounding box annotations are transformed into some filled ellipses to approximate the shape of the true pedestrians.

### 3.2. Inter-modality attention module

Thermal and visible cameras have their own imaging characteristics, and under certain conditions, one sensor has superior imaging quality (i.e. is more relevant for the considered task) than the other. To leverage both modalities, we propose the inter-modality attention module, which adaptively selects thermal or visible features according to the dynamic comparison of their feature quality. Concretely, an inter-modality attention mask is predicted based on the combination of thermal and visible features. This predicted mask has two values for each pixel, corresponding to the weights for thermal and visible features (summing to 1). This attention module is illustrated as the red paths in Fig. 3. It can be formulated as:

$$\begin{aligned} f^t_{inter} &= f^t \otimes (1 + m^t_{inter}) \\ f^v_{inter} &= f^v \otimes (1 + m^v_{inter}) \end{aligned} \tag{3}$$

where

$$m^t_{inter}, m^v_{inter} = \delta(\mathcal{F}_{inter}([f^t, f^v])) \tag{4}$$

Here, $\delta$ denotes the softmax function; $[\cdot]$ denotes the feature concatenation operation; $\mathcal{F}_{inter}$ represents a convolution operation to predict the inter-modality attention mask $m_{inter}$. At each spatial position of the mask, the sum of $m^t_{inter}$ and $m^v_{inter}$ equals to 1. Note that this formalization could theoretically allow for more than two modalities to be fuse following the same principles.

The inter-modality attention module allows the network to adaptively select the most reliable modality. However, in

order to train this module, we should need a costly ground truth information about the best pixel-level modality quality. Our solution to relieve the annotation cost is to assign labels according to the prediction of the pedestrian masks from the intra-modality attention module, i.e., we force the network to select one modality if its intra-modality mask prediction is better (i.e. closer to the ground truth pedestrian mask) than the other. Specifically, we first calculate an error mask for each spectrum with the following formula:

$$\begin{aligned} e^t_{intra} &= \mid m^t_{intra} - m^{gt}_{intra} \mid \\ e^v_{intra} &= \mid m^v_{intra} - m^{gt}_{intra} \mid \end{aligned} \tag{5}$$

then the label for the modality selection is defined as:

$$m^{gt}_{inter} = \begin{cases} 1,0 & if\ (e^v_{intra} - e^t_{intra}) > margin \\ 0,1 & if\ (e^t_{intra} - e^v_{intra}) > margin \\ ignored & otherwise \end{cases} \tag{6}$$

Here, $|\cdot|$ denotes the absolute function; $e_{intra}$ represents the error mask, defined by the L1 distance between the predicted intra-modality mask $m_{intra}$ and the ground truth intra-modality mask $m^{gt}_{intra}$; $m^{gt}_{inter}$ is the ground truth mask for inter-modality attention (2 values at each mask position); $margin$ is a hyper-parameter to be tuned.

An example of the label assignment for the inter-modality attention mask is shown in Fig. 3. If the intra-modality pedestrian masks are predicted as shown in the yellow paths, the inter-modality (weak) ground truth masks are then defined as the ones shown on the red paths, where white, black and gray areas denote the classification labels 1,0 and *ignored*, respectively. Here, the thermal features produce a better intra-modality mask prediction for the pedestrians on the left side of the input images in Fig. 3. Therefore, according to Eq. 6, the label for the inter-modality mask on this area is assigned as 1,0 (1 for the thermal mask and 0 for the visible mask). For regions where the two intra-modality masks have comparable prediction qualities (i.e., the difference between prediction errors is smaller than the predefined margin), the optimization of the inter-modality attention mask prediction on these areas are ignored (i.e., do not participate in the loss calculation).

### 3.3. Combining intra- and inter-modality attention

The intra-modality attention module enhances features on areas with pedestrians and the inter-modality attention module adaptively selects features from the most reliable modality. When these two modules are combined, the fused features are obtained by:

$$f^{fused} = \frac{f^t_{hybrid} + f^v_{hybrid}}{2} \tag{7}$$

where

$$\begin{aligned} f^t_{hybrid} &= f^t \otimes (1 + m^t_{intra}) \otimes (1 + m^t_{inter}) \\ f^v_{hybrid} &= f^v \otimes (1 + m^v_{intra}) \otimes (1 + m^v_{inter}) \end{aligned} \tag{8}$$

Here, $m_{intra}$ and $m_{inter}$ are predicted intra- and inter-modality attention masks from Eq. 2 and Eq. 4; $f_{hybrid}$ represents features enhanced by both attention modules; $f^{fused}$ represents the final fused features.

As mentioned in Section 2, the optimization of the multispectral feature fusion task may not benefit enough from the sole optimization of the object detection task (as done e.g. in [20]). In GAFF, we propose two specific feature fusion losses, including the pedestrian segmentation loss for the intra-modality attention and the modality selection loss for the inter-modality attention, to guide the multispectral feature fusion task. These losses are jointly optimized with the object detection loss. The final training loss $\mathcal{L}_{total}$ is:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \mathcal{L}_{intra} + \mathcal{L}_{inter} \tag{9}$$

where, $\mathcal{L}_{det}$, $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ are the pedestrian detection, the intra- and inter-modality attention loss, respectively.

## 4. Experiments

In this section, we conduct experiments on KAIST Multispectral Pedestrian Detection Dataset [8] and FLIR ADAS Dataset [1] to evaluate the effectiveness of the proposed method. Moreover, we attempt to interpret the reasons for improvements by visualizing the predicted attention masks. Finally, we provide inference speed analysis on two different target platforms.

### 4.1. Datasets

KAIST dataset contains 7,601 training image pairs and 2,252 pairs testing ones. Some example image pairs from this dataset are shown in Fig. 2. [10] proposes a "sanitized" version of the annotations, where numerous annotation errors are removed. Our experiments are conducted with the original as well as the "sanitized" version of annotations for fair comparisons with our competitors. We found out that the "sanitized" annotations substantially improve the detection accuracy for different network architectures. All models are evaluated with the improved testing annotations from [14] and the usual pedestrian detection metric: log-average Miss Rate over the range of $[10^{-2}, 10^{0}]$ false positives per image (FPPI) under a "reasonable" setting [5], i.e., only pedestrians taller than 50 pixels under no or partial occlusions are considered [1].

[1] We use the evaluation code provided by [10]: https://github.com/Li-Chengyang/MSDS-RCNN/tree/master/lib/datasets/KAISTdevkit-matlab-wrapper

We also conduct experiments on FLIR ADAS Dataset [1]. [19] proposed an "aligned" version of the dataset for multispectral object detection. This new version contains 5,142 well-aligned multispectral image pairs (4,129 pairs for training and 1,013 pairs for testing). FLIR covers three object categories: "person", "car" and "bicycle". Models are evaluated with the usual object detection metric introduced with MS-COCO[13]: the mean Average Precision (mAP) averaged over ten different IoU thresholds.

### 4.2. Implementation details

The proposed GAFF module can be included in any type of two-stream convolutional neural networks. In these experiments, we choose RetinaNet [12] as our base detector. It is transformed into a two-stream convolutional neural network by adding an additional branch for the extraction of thermal features. A ResNet18 [7] or a VGG16 [17] network is pre-trained on ImageNet [2], then adopted as our backbone network. The input image resolution is fixed to $640 \times 512$ for training and evaluation. Our baseline detector applies the basic addition as the multispectral feature fusion method. GAFF is implemented by adding the intra- and inter-modality attention modules, corresponding to the yellow and the red branches in Fig. 3. Focal loss [12] and Balanced L1 loss [15] are adopted as the classification loss and the bounding box regression loss to optimize the object detection task. In order to introduce our specific guidance, we adopt the DICE [3] loss as the pedestrian segmentation loss ($\mathcal{L}_{intra}$ in Eq. 9) and the cross-entropy loss as the modality selection loss ($\mathcal{L}_{inter}$ in Eq. 9).

### 4.3. Ablation study

| $Margin$ | Miss Rate | | |
|---|---|---|---|
| | All | Day | Night |
| 0.05 | 6.92% | 8.47% | 3.68% |
| 0.1 | 6.48% | 8.35% | 3.46% |
| 0.2 | 7.47% | 9.31% | 4.22% |

Table 1: Detection results of GAFF with different $margin$ values in the inter-modality attention module.

**Hyper-parameter tuning.** As reported in Table 1, we conduct experiments with different $margin$ values in the inter-modality attention module on KAIST dataset [8] with "sanitized" annotations. The Miss Rate scores on the Reasonable-all, Reasonable-day and Reasonable-night subsets are listed. We observe that the optimal Miss Rate is achieved when $margin = 0.1$. Thus, we use $margin = 0.1$ for all the following experiments.

**Residual attention.** As mentioned in Section 3, attention enhanced features are added as residual to avoid directly affecting the thermal or visible features. We verify this choice by comparing in Table 2 the Miss Rate of GAFF where the attention masks are directly applied to mono-spectral features ($f_{intra} = f \otimes m_{intra}$ and $f_{inter} = f \otimes m_{inter}$) or added as residual (as in Eq. 1 and Eq. 3).

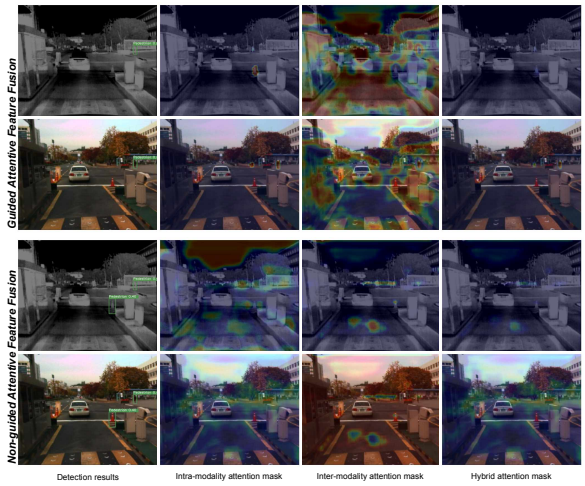
| Residual | Miss Rate | | |
|---|---|---|---|
| | All | Day | Night |
| | 7.46% | 8.88% | 4.85% |
| ✓ | 6.48% | 8.35% | 3.46% |

Table 2: Detection results of GAFF where the attention masks are directly applied or added as residual.

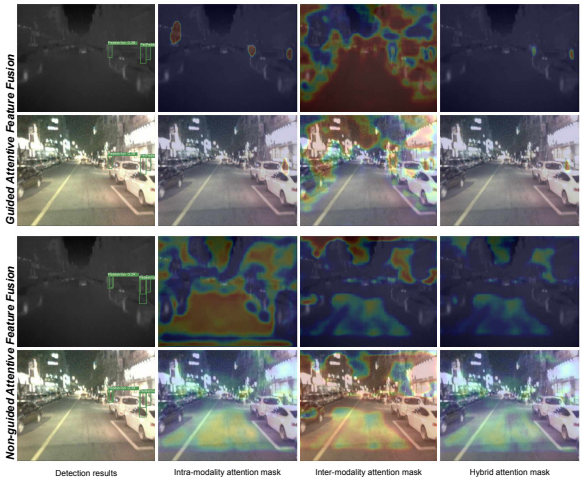
**Necessity of attention.** We compare in Tab. 3 the detection accuracy on KAIST dataset with different attention settings, different backbone networks, and different annotation settings (original and "sanitized"). When conducting experiments with inter-modality but without intra-modality attention, the pedestrian masks are predicted but are not multiplied with the corresponding mono-spectral features. For each backbone network or annotation setting, both intra- and inter-modality attention modules consistently improve the baseline detection accuracy, and their combination leads to the lowest overall Miss Rate under all experimental settings. The present findings confirm the effectiveness of the proposed guided attentive feature fusion modules.

**Necessity of guidance.** To explore the effect of the proposed multispectral feature fusion guidance, we compare our guided approach to one with a similar architecture as ours but where the optimization of the specific fusion losses ($\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ in Eq. 9) are removed from the training process, i.e., the fusion is only supervised by the object detection loss (as done with [20]). We report in Tab. 4 the detection performance with and without guidance, under different backbone networks and annotations settings. The results confirm our assumption that the object detection loss is not relevant enough for the multispectral feature fusion task: even though the non-guided attentive fusion module improves the baseline Miss Rate to some degree (e.g., with the "sanitized" annotations and VGG16 backbone, non-guided model improves the base detector's Miss Rate from 9.28% to 8.38%), it could be further improved when the specific fusion guidance is added (from 8.38% to 6.48%).

**Attention mask interpretation.** Fig. 4 provides the visualization results of the intra-modality, the inter-modality and the hybrid attention masks during daytime and nighttime. For each figure, the top and bottom two rows of images are visualization results of guided and non-guided attentive feature fusions, respectively. We can see on the




(a) Daytime




(b) Nighttime

Figure 4: Visualization examples of attention masks on KAIST dataset. Zoom in to see details.

| Backbone | GAFF | | Miss Rate | | |
|---|---|---|---|---|---|
| | Intra. | Inter. | All | Day | Night |
| ResNet18 | | | 13.04% | 13.83% | 11.60% |
| | ✓ | | 12.13% | 11.97% | 11.99% |
| | | ✓ | 11.15% | 10.68% | 11.67% |
| | ✓ | ✓ | 10.74% | 10.46% | 11.10% |
| VGG16 | | | 12.72% | 11.37% | 15.57% |
| | ✓ | | 11.78% | 11.45% | 12.50% |
| | | ✓ | 11.03% | 10.99% | 11.44% |
| | ✓ | ✓ | 10.62% | 10.82% | 10.14% |

(a) Original annotations

| Backbone | GAFF | | Miss Rate | | |
|---|---|---|---|---|---|
| | Intra. | Inter. | All | Day | Night |
| ResNet18 | | | 9.98% | 12.46% | 5.29% |
| | ✓ | | 9.26% | 11.51% | 5.32% |
| | | ✓ | 9.29% | 11.97% | 5.14% |
| | ✓ | ✓ | 7.93% | 9.79% | 4.33% |
| VGG16 | | | 9.28% | 11.73% | 5.17% |
| | ✓ | | 8.70% | 11.42% | 3.55% |
| | | ✓ | 7.73% | 10.35% | 2.81% |
| | ✓ | ✓ | 6.48% | 8.35% | 3.46% |

(b) "Sanitized" annotations

Table 3: Ablation study of two attentive fusion modules on KAIST dataset [8] with original (top) or "sanitized" (bottom) annotations.

intra-modality attention masks that the guided attention mechanism focuses on pedestrian areas, even though, sometimes, it is not accurate from a single mono-spectral view. For example, the traffic cone is misclassified as a pedestrian

| Backbone | Guidance | Miss Rate | | |
|---|---|---|---|---|
| | | All | Day | Night |
| ResNet18 | | 13.15% | 13.71% | 11.54% |
| | ✓ | 10.74% | 10.46% | 11.10% |
| VGG16 | | 13.67% | 13.19% | 14.51% |
| | ✓ | 10.62% | 10.82% | 10.14% |

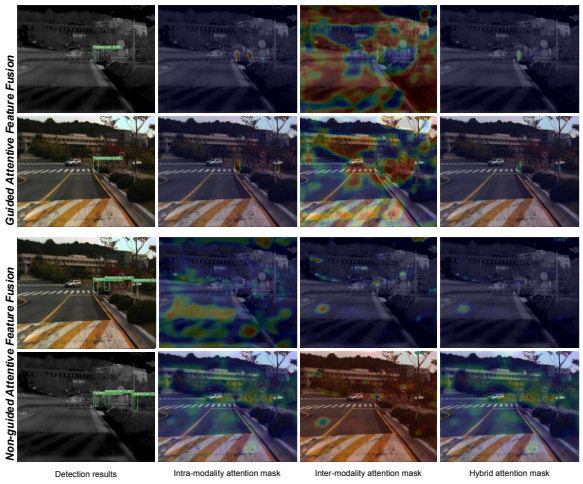
(a) Original annotations

| Backbone | Guidance | Miss Rate | | |
|---|---|---|---|---|
| | | All | Day | Night |
| ResNet18 | | 9.05% | 10.63% | 6.01% |
| | ✓ | 7.93% | 9.79% | 4.33% |
| VGG16 | | 8.38% | 10.39% | 4.44% |
| | ✓ | 6.48% | 8.35% | 3.46% |

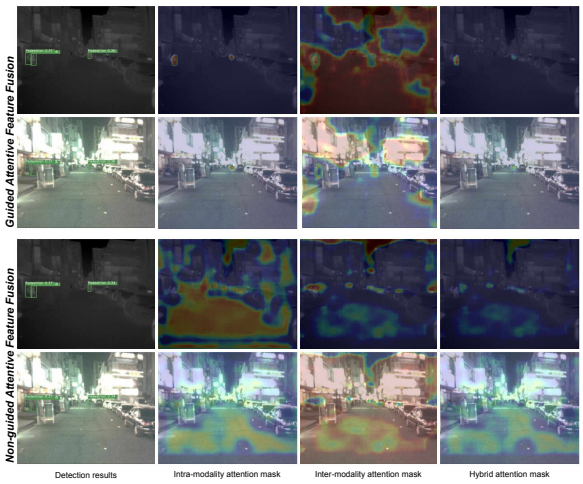
(b) "Sanitized" annotations

Table 4: Comparison between guided and non-guided models on KAIST dataset [8] with both annotation settings.

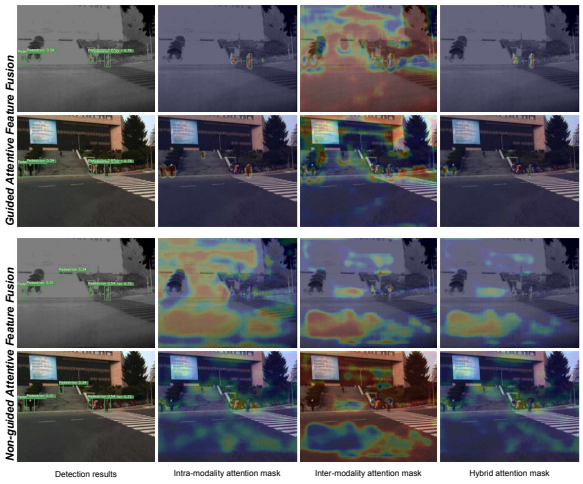
due to its human-like shape on the thermal image of Fig. 4a, and the pedestrian in the middle right position is missed due to insufficient lighting on the RGB image of Fig. 4b. For inter-modality attention masks, it appears that the guided attentive fusion tends to select visible features on well-lit areas (such as upside of images in Fig. 4b) and brightly coloured areas (e.g., traffic cone, road sign, speed bump, car tail light, etc), and to select thermal features on dark areas and uniform areas (such as sky and road). Note that these attention preferences are automatically learnt via our inter-modality attention guidance. On the contrary, despite the fact that the non-guided attention mechanism brings some accuracy improvements, the predicted attention masks are




(a) Daytime




(b) Nighttime




(c) Error case

Figure 5: More visualization examples of attention masks on KAIST dataset. Zoom in to see details.

quite difficult to interpret. More visualization results are shown in Fig. 5. Besides, an interesting error case is shown in Fig. 5c, where the pedestrian on the steps is not detected with the guided model but detected with the non-guided model. As mentioned earlier, GAFF selects thermal features on uniform areas, which is intuitive since thermal cameras are sensitive to temperature change and there exist few objects on uniform areas of the thermal image. However, in this particular case, the pedestrian is not captured on the thermal image, which leads to the final detection error.

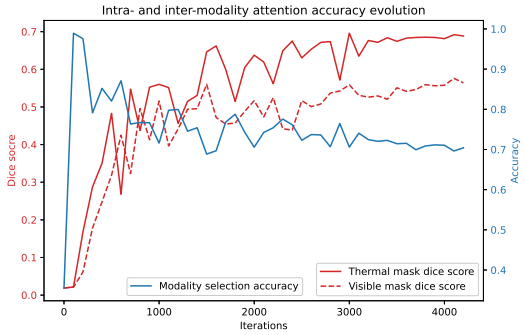



Figure 6: Intra- and inter-modality attention accuracy evolution during training.

**Attention accuracy evolution** We plot in Fig. 6 the evolution of intra- and inter-modality attention accuracy during training. Specifically, red solid and dashed lines represent the pedestrian segmentation accuracy (via DICE score [3] $Dice = \frac{2|A \cap B|}{|A|+|B|}$) from thermal and visible features in intra-modality attention module; blue line indicates the modality selection accuracy in inter-modality attention module. From the plot, we can conclude that thermal images are generally better for recognition than visible images. This observation is consistent with our mono-spectral experiments, where thermal-only model reaches 18.8% of Miss Rate while visible-only model achieves 20.74% (both trained with "sanitized" annotations). Interestingly, as the segmentation accuracy increases for both images, the modality selection task becomes more and more challenging. Note that this accuracy is irrelevant at the beginning of the training, where predicted pedestrian masks are almost zero for both thermal and visible features, thus the difference between their error masks is minor and the set of $margin$ makes most areas ignored for modality selection optimization. Such mechanism avoids the "cold start" problem.

**Runtime analysis** In Tab. 5 we report the total number of learnable parameters and the average inference runtime on two different computation platforms. Specifically, the models are implemented with Pytorch (TensorRT) framework

| Backbone | GAFF | Param. | Runtime | |
|---|---|---|---|---|
| | | | 1080Ti | TX2 |
| ResNet18 | | 23,751,725 | 10.31ms | 10.5ms |
| | ✓ | 23,765,553 | 10.85ms | 12.1ms |
| VGG16 | | 31,403,053 | 8.87ms | 10.3ms |
| | ✓ | 31,430,705 | 9.34ms | 11.6ms |

Table 5: Runtime on different computing platforms.

| Methods | Miss Rate | | |
|---|---|---|---|
| | All | Day | Night |
| ACF+T+THOG [8] | 47.24% | 42.44% | 56.17% |
| Halfway Fusion [14] | 26.15% | 24.85% | 27.59% |
| Fusion RPN+BF [14] | 16.53% | 16.39% | 18.16% |
| IAF R-CNN [11] | 16.22% | 13.94% | 18.28% |
| IATDNN+IASS [6] | 15.78% | 15.08% | 17.22% |
| CIAN [20] | 14.12% | 14.77% | 11.13% |
| MSDS-RCNN [10] | 11.63% | 10.60% | 13.73% |
| CFR [19] | 10.05% | 9.72% | 10.80% |
| GAFF (ours) | 10.62% | 10.82% | 10.14% |

(a) Original annotations

| Methods | Miss Rate | | |
|---|---|---|---|
| | All | Day | Night |
| MSDS-RCNN [10] | 7.49% | 8.09% | 5.92% |
| CFR [19] | 6.13% | 7.68% | 3.19% |
| GAFF(ours) | 6.48% | 8.35% | 3.46% |

(b) "Sanitized" annotations

Table 6: Detection results on KAIST dataset [8] with original (top) or "sanitized" (bottom) annotations.

for an inference time testing on the Nvidia GTX 1080Ti (Nvidia TX2) platform. Since GAFF only involves 3 convolution layers, the additional parameters and computation cost is low, i.e., it represents less than 0.1% of additional parameters and around 0.5ms (1.5ms) of inference time on 1080Ti (TX2). Note that the time for post-processing treatments (such as Non-Maximum Suppression) is not taken into account for the benchmarking. Our model meets the requirement of real-time treatment on embedded devices, which is essential for many applications.

### 4.4. Comparison with State-of-the-art Multispectral Pedestrian Detection Methods

**KAIST Dataset** Tab. 6 shows the detection results of existing methods and our GAFF with the original and "sanitized" annotations on KAIST. It can be observed that GAFF achieves state-of-the-art performance on this dataset (it is slightly less accurate than CFR [19], which applies cascaded Fuse-and-Refine blocks for sequential feature enhancement and needs more computation than GAFF (see Table 7). According to Tab. 7, thanks to the lightweight design of GAFF, our model has substantial advantage in terms of inference speed compared to e.g. [19].

| Methods | Platform | Runtime |
|---|---|---|
| ACF+T+THOG [8] | MATLAB | 2730ms |
| Halfway Fusion [14] | Titan X | 430ms |
| Fusion RPN+BF [14] | MATLAB | 800ms |
| IAF R-CNN [11] | Titan X | 210ms |
| IATDNN+IASS [6] | Titan X | 250ms |
| CIAN [20] | 1080Ti | 70ms |
| MSDS-RCNN [10] | Titan X | 220ms |
| CFR [19] | 1080Ti | 50ms |
| GAFF (ours) | 1080Ti | 9.34ms |

Table 7: Runtime comparisons with different methods on KAIST dataset [8].

| Backbone | GAFF | mAP | AP75 | AP50 |
|---|---|---|---|---|
| ResNet18 | | 36.6% | 31.9% | 72.8% |
| | ✓ | 37.5% | 32.9% | 72.9% |
| VGG16 | | 36.3% | 30.2% | 71.9% |
| | ✓ | 37.3% | 30.9% | 72.7% |

Table 8: Detection results on FLIR dataset [1].

**FLIR Dataset** Tab. 8 reports the detection results with and without GAFF on FLIR dataset. We can observe that the average precision is improved for all IoU thresholds with GAFF (around 1% of mAP improvement for both backbone networks), which shows that our method can generalize well to different types of images. For comparison, the more costly CFR [19] reaches 72.39% of AP50 on this dataset, whereas our best result is 72.9%.

## 5. Conclusion

We argue that the lack guidance is a limitation for efficient and effective multispectral feature fusion, and we propose *Guided Attentive Feature Fusion* (GAFF) to guide this fusion process. Without hand-crafted assumptions or additional annotations, GAFF realizes a fully adaptive fusion of thermal and visible features. Experiments on KAIST and FLIR datasets demonstrate the effectiveness of GAFF and the necessity of attention and guidance in the feature fusion stage. We noticed that certain thermal-visible image pairs are slightly misaligned in the above datasets, such a problem could be more critical in real life applications. Our future research is devoted to the development of a real-time feature calibration module based on the predicted attention masks from GAFF.

## References


[1] Free flir thermal dataset for algorithm training. https://www.flir.com/oem/adas/adas-dataset-form/.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[3] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[4] Piotr Dollar, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(8):1532–1545, Aug. 2014.

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

[6] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[8] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 243–250, 2017.

[10] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 225, 2018.

[11] 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.

[12] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007, 2017.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft COCO: Common objects in context. In *ECCV*. European Conference on Computer Vision, September 2014.

[14] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas. Multispectral deep neural networks for pedestrian detection. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.

[15] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[16] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[18] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27-29, 2016*, 2016.

[19] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks. In *ICIP 2020 - IEEE International Conference on Image Processing*, pages 1–5, Abou Dabi, United Arab Emirates, Oct. 2020.

[20] 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.