# DaFF: Dual Attentive Feature Fusion for Multispectral Pedestrian Detection

Afnan Althoupety, Li-Yun Wang, Wu-chi Feng, Banafsheh Rekabdar
Portland State University, USA
{afnan2, liyuwang, wuchi, rekabdar}@pdx.edu

## Abstract

*Inspired by how humans perceive and interpret the world using multiple senses, multi-modal learning involves integrating information from multiple modalities to improve understanding and performance in various tasks. Aligning with that notion, our key intuition is to utilize multi-model learning to solve the domain shift problem in nighttime pedestrian detection.*

*In this paper, we show that pairing RGB and infrared (IR) image features increases the robustness of pedestrian detection at night. Indeed, this solution is unbiased towards a specific time of the day as the IR domain reduces the reliance on lighting and serves as complementary information to the RGB domain. Our work aims at exploiting the power of attention mechanisms to guide a multi-modal framework in feature fusing from RGB and IR modalities. Our novel fusion approach, named dual attentive feature fusion (DaFF), leverages the duality of the transformer and channel-wise global attentions. To demonstrate the effectiveness of DaFF, we conducted experiments on two real-world multispectral pedestrian datasets. Extensive experimental results reveal the superiority of DaFF. We believe that combining the complementary properties of RGB and IR modalities is an effective remedy to mitigate the domain shift problem in pedestrian detection.*

## 1. Introduction

Developing a robust pedestrian detection system at nighttime has become an important problem in computer vision in recent years due to applications of Advanced Driver Assistance Systems (ADAS) and Video Surveillance [1, 12]. Many pedestrian detection studies have shown that based on RGB images, deep learning-based pedestrian detectors enhance detection performance on daytime images but encounter degraded detection performance on nighttime images. Since these images have limited illumination, the detectors struggle to accurately detect pedestrians in the nighttime images. The degraded detection performance can raise safety issues in the ADAS and video surveillance applications. Unlike conventional pedestrian detection techniques that only consider mono-modalities as inputs to the pedestrian detectors, multi-modal frameworks integrate data from different sources to enhance the performance in various applications [6, 11, 18, 26, 30, 39]. Multi-modal learning is motivated by how humans use multiple senses to perceive and interpret the world. Accordingly, we tackle the domain shift problem in nighttime pedestrian detection by applying multi-model learning.

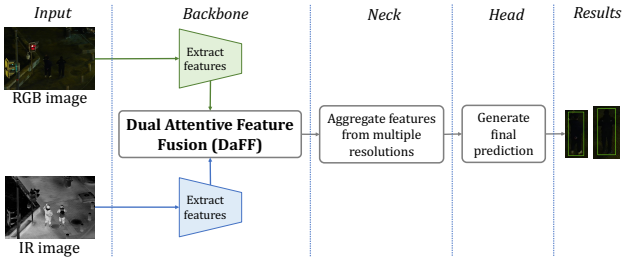


Figure 1. Pipeline of the multi-modal framework guided by DaFF, dual attentive feature fusion, for multispectral pedestrian detection. The framework takes the same scene from two sensors (RGB and IR) as inputs. Green and blue represent two separated backbones with the same network architecture for the RGB and IR images, respectively. DaFF fuses features during the feature extraction. Fused features from RGB and IR branches are passed to the neck module then head module for final detection. Green boxes in the right-hand side are detected bounding boxes of pedestrians.

In this paper, we introduce a novel multi-modality feature fusion method, namely *dual attentive feature fusion (DaFF)*, to improve the detection performance of nighttime images for multispectral pedestrian detection. Our proposed DaFF method fuses visual textures and object contours in the RGB and IR images at the feature extraction stage. The proposed DaFF method combines two forms of global attention. In particular, DaFF includes spatial and channel-wise attention mechanisms to guide the multi-modal framework. Exploiting attention mechanisms provides a better feature representation and improves the fusion quality. Thus, the DaFF method can enhance detection performance by exploiting the dual attention mechanisms.

To effectively fuse the complementary information in the RGB and IR images, we propose a duality of spatial and channel-wise attention mechanisms. For the spatial attention, we leverage transformers to attain informative features across different locations [4, 22, 34]. Additionally, we obtain important features across different channels utilizing a channel-wise attention module with learnable parameters. Therefore, the duality in DaFF enhances features by learning the global context at different locations and channels.

Figure 1 shows the whole pipeline of our proposed approach for multispectral pedestrian detection. We first utilize two backbones to extract various levels of features from the RGB and IR images. Then, we feed two mono-modality features into the DaFF module to generate multi-modality fused features. Fused features by DaFF are passed through a neck module for aggregation. Finally, a head module outputs the final detection bounding boxes.

The main contribution of this paper can be summarized in threefold:

- We present a novel DaFF method for multispectral pedestrian detection, which fuses the complementary information from RGB and IR images at the feature level utilizing global attention that operates on both spatial locations and feature channels.
- We show the effectiveness of each module in our framework through several qualitative and quantitative ablation studies.
- We compare our proposed feature fusion method, DaFF, to state-of-the-art fusion approaches on two real-world multispectral pedestrian datasets.

## 2. Related Work

Combining information from different sources to obtain a unified picture is advantageous in various disciplines [14]. In particular, for autonomous driving and video surveillance, exploiting RGB and IR sensors accommodates different weather conditions and time of the day. RGB sensors provide color and texture information under good illumination while IR sensors are resilient to illumination variation. Pedestrian detection as a central part of many applications can benefit from the concept of multi-sensor data, especially for nighttime robustness. Adapting such a concept is usually known as multi-modal learning, where features are extracted from each mono-modality and then it gets fused together for better feature representation. For multispectral pedestrian detection, feature fusion methods are concerned with two aspects: when to fuse? and how to fuse?

Corresponding to when to fuse, Wagner *et al.* [31] performed the first study to exploit RGB and IR images using a deep object detector, RCNN. The study evaluates two fusion stages: fusion at the pixel-level and fusion at the feature level and reveals the superiority of the latter. To further explore the potential of multispectral pedestrian detection, Liu *et al.* [20] analyze the performance of Faster R-CNN considering four fusion stages: early stage (after the first convolutional layers), halfway fusion (after the fourth convolutional layers), late fusion (at the last fully connected layers), and score fusion (detection scores are combined). Halfway fusion achieves the best detection performance and the author claims this is because the low-level features at the early stage are irrelevant and the high-level features at the late stage are semantic. Therefore, halfway fusion produces the best balance between the fine visual details and the semantic meanings.

Corresponding to how to fuse, Zhang *et al.* [37] suggest a cyclic fuse-and-refine approach to progressively improve the complementary and the alignment of multispectral features. Seeking illumination aware fusion mechanisms, Li *et al.* [17] present an illumination-aware network consisting of a day illumination sub-network and a night illumination sub-network. Specifically, the illumination value is used to adaptively weigh feature fusion. MBNet [42] was suggested to address the modality imbalance problem for multispectral pedestrian detection. MBNet uses two modules to first differentiate between RGB and thermal features and then adaptively align both features utilizing the illumination conditions. Probabilistic ensembling (ProbEn) [5] is a late fusion approach based on Bayes' theorem. ProbEn assumes conditional independence across modalities and finds the final detection results by fusing the score from each modality. PIAFusion [28] is a progressive image fusion framework with the guidance of illumination-aware loss. As a task-driven fusion approach, Sun *et al.* propose DetFusion [27], an object-aware image fusion network that exploits a priori information of object locations to guide the fusion process. Considering attention-based fusion, [38] proposes Guided Attentive Feature Fusion (GAFF) that utilizes an attention mask highlighting pedestrian objects to weigh features' importance before fusion. Qingyun *et al.* [21] propose a cross-modality fusion transformer (CFT) that leverages the self-attention mechanism of the transformer to adaptively learn the correlation between RGB and IR modalities. Shen *et al.* [25] introduce the Iterative Cross-Attention Guided Feature Fusion (ICAFusion) framework for multispectral object detection tasks. ICAFusion is based on a cross-attention fusion transformer and uses a new iterative learning strategy to share parameters between transformer blocks for efficiency purposes.

While prior feature fusion strategies are constrained by hand-crafted weighing schemes, limited local-range feature interactions, or one-dimensional attention modules, DaFF provides a comprehensive attention-based feature fusion method benefiting from both channel and spatial axes to enrich the quality of the feature fusion for multispectral pedestrian detection. In the following sections, we will describe DaFF, motivate the integration of channel and spatial atten-

tions, and show its advantages over other fusion methods in the context of multispectral pedestrian detection.

## 3. Proposed Approach

**Why attention?** The concept of attention draws inspiration from human perception. Initially introduced in the context of neural networks, attention has evolved from simple mechanisms to complex, adaptive architectures. The foundation of modern attention mechanisms, self-attention enables a model to weigh the importance of different positions in a sequence, facilitating contextual understanding and information aggregation from a global perspective.

The intuition behind DaFF's design choice is to use the full potential of global attention in guiding the process of combining RGB and IR features. In other words, we leverage the power of attention mechanisms on channel and spatial axes to learn the global context of RGB and IR modalities for an effective feature fusion process. Our key idea is to incorporate two attention modules in a multi-modal framework to obtain a learned feature representation for multispectral pedestrian detection tasks.

We argue that straightforward fusion mechanisms such as element-wise addition and concatenation that rely on the local spatial correlation between features are vulnerable to multispectal image misalignment. Additionally, straightforward fusion mechanisms fuse redundant and irrelevant information. Furthermore, we contend that one-dimensional attentive fusion mechanisms constrain the power of learning global feature interactions. To efficiently borrow complementary information from other modalities and thoroughly learn the importance weights of RGB-IR features, we propose a dual attentive feature fusion, DaFF, that combines two global attentions.

### 3.1. Overall Framework

Figure 2 shows the overall multi-modal framework guided by the proposed fusion approach DaFF. The fusion takes place after the third convolution (halfway fusion) as it is proven to achieve the best results [20]. In addition, we choose YOLOv5 as a base model for our proposed DaFF approach due to its competitive performance according to [12, 21, 25].

As illustrated in Figure 2, the multi-modal framework takes a pair of RGB and IR aligned images that passes through a separate mono-modality to extract RGB and IR features. In the halfway of the feature extraction stage, we start fusing RGB and IR features to gain the complementary properties of both modalities. To address the limitation of the local receptive field in convolutional neural networks (CNNs), we guide the multi-modality by adding two attention modules. The main objective of these add-on modules is to emphasize informative features and suppress the less useful ones, leading to improved multispectral pedestrian detection performance. Specifically, to fuse the features from each mono-modality, we apply transformer and channel-wise attention (CWA) modules, dually learning meaningful features.

### 3.2. Transformer Attention Module

**Why transformer attention?** Vaswani *et al.* [29] first introduced the transformer architecture and it has become a cornerstone in natural language processing (NLP) due to its efficiency and scalability [7, 15, 23, 35]. The basic notion of transformers is to transform input sequences into output sequences by integrating self-attention and feedforward neural networks. Inspired by the massive successes of transformers in NLP, the vision transformer (ViT) [8] was proposed as the first transformer-based model for computer vision that treats the sequence of image patches the same way as the sequence of words in NLP. With the advent of transformer architectures, researchers have extended it to multi-modal learning where attention weights can be learned to emphasize relevant information across modalities, facilitating more effective fusion. Many studies have applied transformers to different multi-modal applications, including object detection [21, 25], image segmentation [36, 41], image and text matching [33], video object segmentation [2]. Because the efficiency of transformers in multi-modal learning, we choose it as our spatial attention and explore its effectiveness for multispectral pedestrian detection tasks.

Figure 3 illustrates the building blocks of the transformer attention module. Similar to [21], we insert transformer modules to the multi-modal framework. First, given two feature maps denoted as $\boldsymbol{F}_{RGB}$ and $\boldsymbol{F}_{I}$ from RGB and IR branches respectively, we flatten each feature map and then concatenate them together. Next, we add a learnable positional embedding, which is a trainable parameter to encode spatial information between different tokens. Let $\boldsymbol{I}$ be the input sequence to the transformer module. $\boldsymbol{I}$ is projected onto three separate matrices to compute a set of queries, keys, and values ($\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$) as follows:

$$\boldsymbol{Q} = \boldsymbol{I}\boldsymbol{W}^{Q}, \boldsymbol{K} = \boldsymbol{I}\boldsymbol{W}^{K}, \boldsymbol{V} = \boldsymbol{I}\boldsymbol{W}^{V} \tag{1}$$

where $\boldsymbol{W}^{Q}$, $\boldsymbol{W}^{K}$, and $\boldsymbol{W}^{V}$ represent the weight matrices.

Second, the attention weights are computed using the dot-product operation between queries and keys, and then multiplied by the values to obtain output $\boldsymbol{Z}$ as follows:

$$\boldsymbol{Z} = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{D_{\text{k}}}}\right)\boldsymbol{V} \tag{2}$$

where $\frac{1}{\sqrt{D_{\text{k}}}}$ is a scaling factor to prevent the gradient vanishing problem when the softmax function returns extremely small gradients.

The multi-head attention is employed with 8 parallel heads.

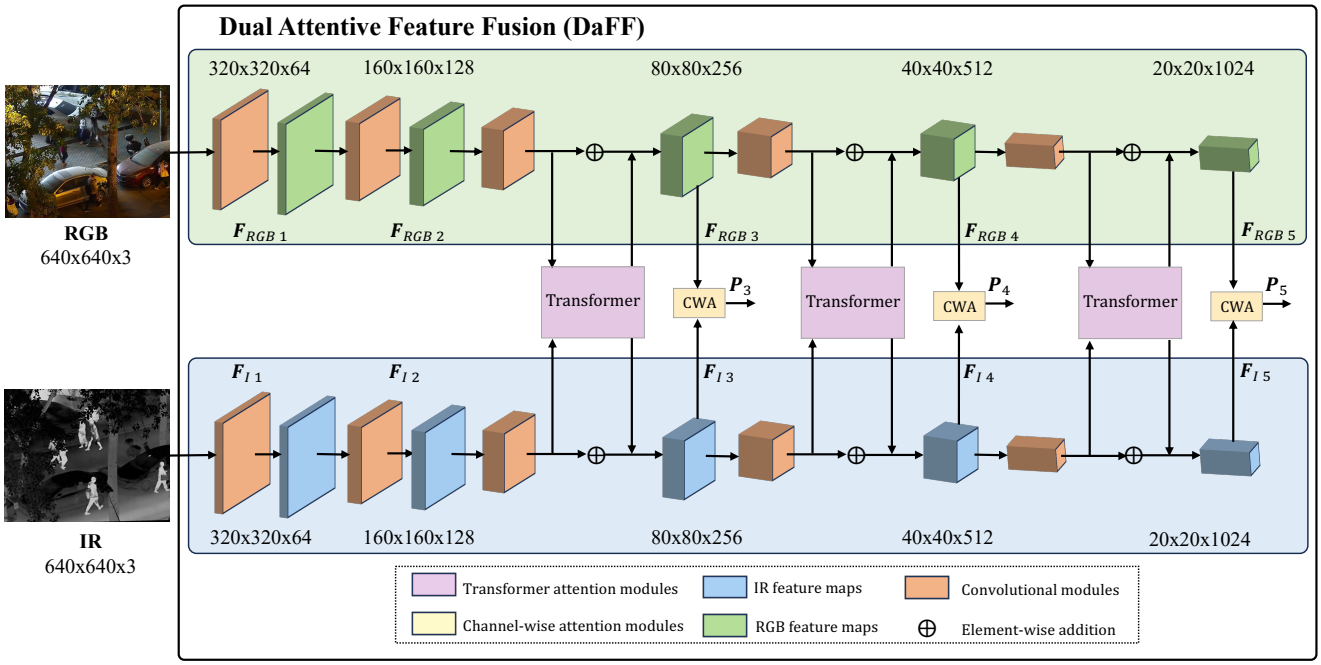


Figure 2. Proposed DaFF method based on YOLOv5. The multi-modal framework takes a pair of RGB and IR images and each image progresses through a separate mono-modality to extract features. The transformer and channel-wise attention modules take place after the third convolution modules, producing P3, P4, and P5.

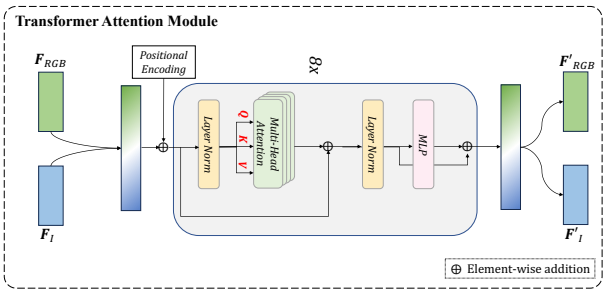


Figure 3. The structure of the transformer attention module in Figure 2.

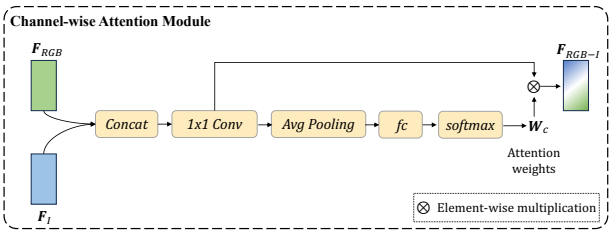


Figure 4. The structure of the channel-wise attention (CWA) module in Figure 2.

Third, the output sequence $\boldsymbol{O}$ is obtained using two fully-connected layers as follows:

$$\boldsymbol{O} = \mathrm{MLP}(\boldsymbol{Z}) + \boldsymbol{I} \tag{3}$$

Finally, the resulting output sequence $\boldsymbol{O}$ is converted by inversing the first step, producing $\boldsymbol{F'}_{RGB}$ and $\boldsymbol{F'}_{I}$, which is added to the corresponding modality branch.

### 3.3. Channel-wise Attention Module

**Why channel-wise attention?** Unlike traditional attention mechanisms that focus on specific spatial or temporal elements within data sequences, channel-wise attention operates at the channel level in feature maps. It allows neural networks to selectively emphasize or suppress certain channels, enabling the model to adaptively learn and utilize information across different channels in a feature map. One of the pioneering works in the channel-wise attention domain is Squeeze-and-Excitation Networks (SENet) [9]. SENet proposes an add-on module to increase representation power by using global information between channels and reweighing features. Influenced by SENet, many researchers have investigated channel-wise attention capabilities and applications [3, 19, 32].

Motivated by [9], we propose to adopt channel-wise attention and expand it to a multi-modal framework for multispectral pedestrian detection. Figure 4 demonstrates the building blocks of the CWA module. Given two feature maps denoted as $\boldsymbol{F}_{RGB}$ and $\boldsymbol{F}_{I}$ from the RGB and IR branches respectively, we concatenate both features and employ a $1\times1$ convolutional layer to reduce the number of

channels. Second, we perform an average pooling operation for channel-wise global coherence. Then, the channel-wise statistics get aggregated through two fully connected layers followed by the softmax function. The output is channel-wise attention weights denoted as $\boldsymbol{W}_c$. The final step includes multiplying those weights by the output of the convolution operation. In this manner, the features from both mono-modalities get fused considering the non-mutually-exclusive relationship. In other words, fusing RGB and IR feature maps utilizing CWA captures the channel-wise dependencies rather than relying only on spatial ones.

## 4. Experimental Setup

In this section, we describe implementation details, including datasets, base model, and evaluation protocol.

### 4.1. Dataset Preparation

To validate our feature fusion method, we use two paired RGB-IR datasets, namely KAIST [10] and LLVIP [12].

*KAIST*. In 2015, Hwang *et al.* introduced KAIST, a popular multi-modal benchmark for pedestrian detection. KAIST has RGB-IR aligned image pairs captured during daytime and nighttime. The KAIST dataset approximately contains 95k image pairs: 50k for training and 45k for testing. Since the original annotations were problematic, we leverage the sanitized version of the training set [16], and the test set [40]. Following the common procedure with KAIST, we use a sampling skip of every 2nd frame for the training set as in [13, 17, 20], and every 20th frame for the test set analogous to [16, 40]. Table 1 details the split of the resulting training/test sets.

*LLVIP*. In 2021, LLVIP was released as an RGB-IR paired pedestrian dataset for low-light vision. Image pairs are strictly aligned in time and space. Moreover, the LLVIP dataset was collected from different locations on the street between 6 and 10 o'clock in the evening. Table 1 details the split of the training and test sets.

| Dataset | Split | Images | Persons | Resolution |
|---|---|---|---|---|
| KAIST | Train | 25,086 | 26,642 | 640 × 512 |
| | Test-All | 2,252 | 2,757 | |
| | Test-Day | 1,455 | 2,003 | |
| | Test-Night | 797 | 754 | |
| LLVIP | Train | 12,025 | 34,135 | 1280 × 1024 |
| | Test | 3,463 | 8,302 | |

Table 1. KAIST and LLVIP multispectral pedestrian datasets.

### 4.2. Implementation Notes

YOLOv5 is our base model with the CSPDarkNet53 backbone. The training phase takes 200 epochs using two Nvidia A40 GPUs. We adopt the stochastic gradient descent (SGD) optimizer with an initial learning rate of 1e-2, a momentum of 0.937, and a weight decay of 0.0005. We use a batch size of 8 and an input image size of $640 \times 640$. We utilize a pre-trained YOLOv5 on COCO dataset for weight initialization, and both mosaic and random flipping for data augmentation.

We use the standard object detection evaluation metric: Average precision (AP) introduced by MS-COCO evaluation protocol[1]. We report AP at all default Intersection over Union (IoU) thresholds on the LLVIP dataset. However, due to the noisy annotations of KAIST, we report AP results using the least strict default threshold (IoU=.50).

For comparison, we carefully follow the instructions to train other fusion approaches on the same data. Please note that training MBNet [42] on LLVIP dataset was not feasible as the training set does not include daytime images. Also, since we train all models on the same data, we choose to unify the evaluation metric across both used datasets.

## 5. Experimental Results and Discussion

We have conducted several experiments to verify the effectiveness of our proposed fusion mechanism. In this section, we first examine our design choices and the contribution of each module through multiple ablation studies both quantitatively and qualitatively. Second, we compare our DaFF with the other state-of-the-art fusion mechanisms.

### 5.1. Ablation Studies

To study the impact of each attention module in DaFF, we conducted ablation experiments in this part. Quantitatively, we evaluate the detection performance by adding transformer-only, CWA-only, and then both to obtain the impact of their integration. Qualitatively, we demonstrate the interpretability of each module both separately and jointly by generating the visual explanations using Grad-CAM [24] and visualize the attention maps associated with the detection results.

| Dataset | Transformer | CWA | AP $^{\text{IoU=.50}}$ |
|---|---|---|---|
| KAIST | ✓ | | 59.2 |
| | | ✓ | 59.1 |
| | ✓ | ✓ | **61.9** (↑ **2.7**) |
| LLVIP | ✓ | | 97.2 |
| | | ✓ | 97.1 |
| | ✓ | ✓ | **97.8** (↑ **0.6**) |

Table 2. The table presents ablation experiments of each attention module in DaFF on KAIST (All test set) and LLVIP datasets. Used performance metrics is average precision (AP %) at IoU =.50. The **bold black** denotes the best performance. The results ensure that the integration of both transformer and channel-wise attention modules achieves the highest detection performance.

[1] https://cocodataset.org/#detection-eval

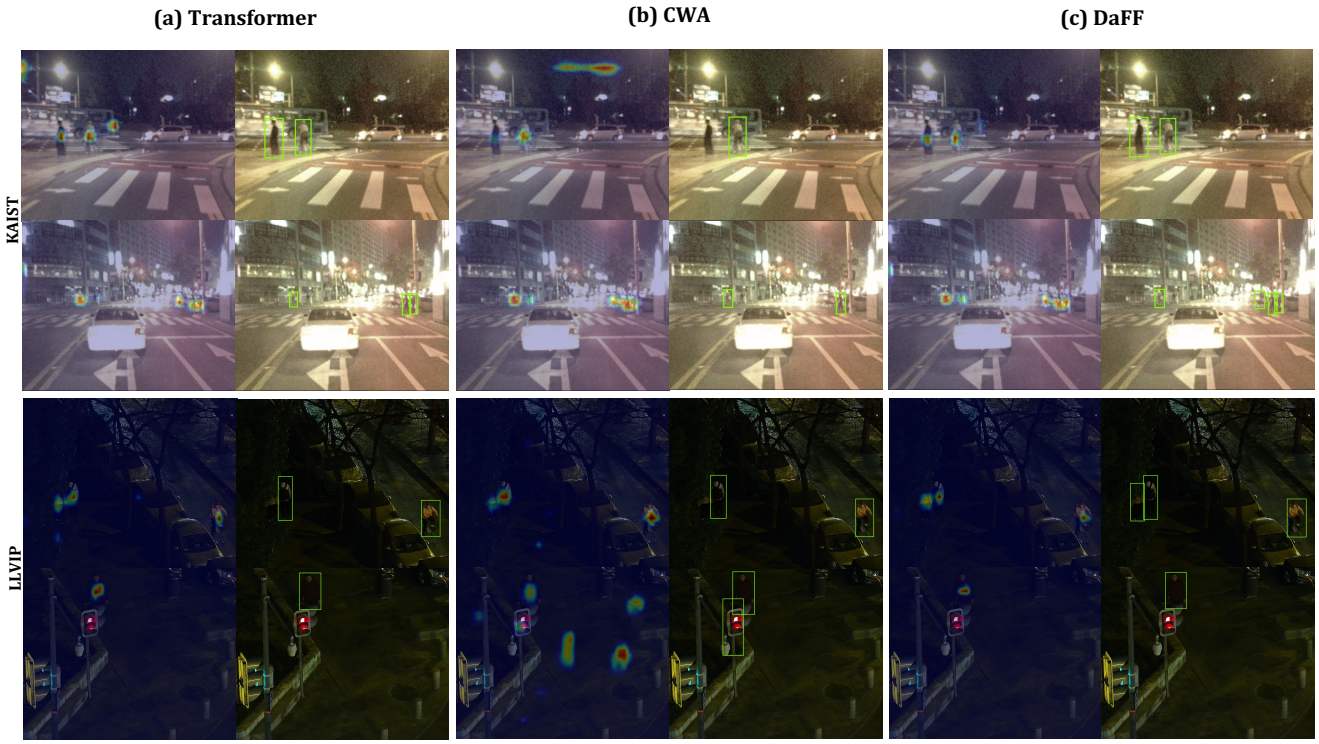


Figure 5. Visual explanations by Grad-CAM of attention modules in our proposed DaFF on KAIST and LLVIP datasets. Each example includes the attention map (left) and the detection result (right). Results reveal that the combination of transformer and CWA highlights the important regions of the image, where the pedestrian is located, more precisely compared to solo modules. Zoom in for more details.

As presented in Table 2, the results when both forms of attention are present, transformer and CWA, maximize the detection performance among all metrics on both KAIST and LLVIP. Alternatively stated, the duality of self-attentions at spatial and channel levels brings the best gain for multispectral pedestrian detection. Specifically, the AR is elevated by 1% and 0.2% compared to using a solo module on KAIST and LLVIP, respectively. More notably, the performance gain in terms of AP is 2.7% and 0.6% on KAIST and LLVIP, respectively.

Figure 5 depicts the visual interpretation by Grad-CAM to compare the effect of DaFF's attention modules in isolation and combination. The impact of the combination has led the multi-modal to focus on the target region of interest, pedestrians, with fewer distractions. As a result, the performance of DaFF brings the improvement of transformer and CWA in one framework and boosts the detection robustness.

## 5.2. Comparison with Other Approaches

We compare DaFF to other baselines: mono-modalities (RGB-only and IR-only) to emphasize the benefit of multi-modality, a basic multi-modality (addition) to justify incorporating attention modules, and other state-of-the-art fusion approaches to show the promise of our fusion method.

| Method | $AP^{IoU=.50}$ | | |
|---|---|---|---|
| | All | Day | Night |
| *mono-modality networks* | | | |
| IR | 50.8 | 47 | 60.4 |
| RGB | 51.4 | 54 | 44.1 |
| *multi-modality networks* | | | |
| Addition | 59.8 | 58.4 | 63.5 |
| MBNet [42] | 58.3 | 58.3 | 59.8 |
| DetFusion [27] | 28.9 | 30.9 | 23.8 |
| PIAFusion [28] | 45 | 46.2 | 43.8 |
| CFT [21] | 59.2 | 56.7 | **66** |
| ICAFusion [25] | **60.3** | **59** | 64.3 |
| DaFF (ours) | **61.9** (↑ 1.6) | **60.2** (↑ 1.2) | **66.8** (↑ 0.8) |

Table 3. The table presents a detection performance comparison of mono-modalities, state-of-the-art multi-modalities, and DaFF on KAIST dataset. Used performance metrics is average precision (AP %) at IoU = 50%. The **bold blue** denotes best performance and **bold black** denotes second best. Results indicate the superiority of DaFF over other approaches on KAIST dataset (considering all, night, and day test sets)

**On KAIST.** Quantitatively, Table 3 reports the detection performance of DaFF and other methods on three test sets (all, night, and day) of the KAIST dataset. It is observed


Figure 6. Qualitative comparison of detection performance, including the mono-modalities, and our multi-modality, DaFF, on KAIST and LLVIP datasets. The green boxes denote detected pedestrians, and the red triangles indicate failure cases, false positives, or false negatives. Results show that the performance of multi-modality is more powerful than the mono-modality in terms of multispectral pedestrian detection. Zoom in for more details.

that the IR-only results surpass the RGB-only ones on the night test set. Also, most multi-modalities substantially outperform mono-modalities, confirming that fusing features from RGB and IR images is beneficial for the pedestrian detection problem. However, it is notable that DetFusion, an object-aware fusion approach, and PIAFusion, an illumination-aware approach, give the lowest detection performance even compared to mono-modalities which affirms the importance of deciding how to fuse. Moreover, DaFF improves the AP by 1.6%, 1.2%, and 0.8% on all, day, and night test sets, respectively. Qualitatively, Figure 6 and 7 show a total of six successful examples of KAIST validating the effectiveness of DaFF in detecting pedestrians against mono-modalities and other multi-modalities.

**On LLVIP.** Quantitatively, Table 4 presents the detection performance of DaFF and other methods on the LLVIP dataset. Notably, the IR-only results are higher than the RGB-only ones by at least 6%. Also, performance of the addition multi-modality is less than IR mono-modality; we attribute that to the limited capacity of addition fusion in adaptively learning informative features. Also, DetFusion and PIAFusion degrade the performance compared to most cases in mono-modalities and all cases in multi-modalities. Moreover, DaFF achieves a gain of 0.6%, 1% and 0.8% at all default IoU, respectively. DaFF achieves the new state-of-the-art performance on LLVIP dataset[2]. Qualitatively,

[2] https://paperswithcode.com/sota/pedestrian-detection-on-llvip

| Method | AP IoU=.50 | AP IoU=.75 | AP IoU=.50:.05:.95 |
|---|---|---|---|
| *mono-modality networks* | | | |
| IR | 96.6 | 73.8 | 64.6 |
| RGB | 90.7 | 51.7 | 50.7 |
| *multi-modality networks* | | | |
| Addition | 96.3 | 72 | 62.5 |
| DetFusion [27] | 89.2 | 56.3 | 52.7 |
| PIAFusion [28] | 88.6 | 54.5 | 51.9 |
| CFT [21] | **97.2** | **74.7** | **64.8** |
| ICAFusion [25] | 96.3 | 71.7 | 62.3 |
| DaFF (ours) | **97.8** (↑ 0.6) | **75.7** (↑ 1.0) | **65.6** (↑ 0.8) |

Table 4. The table presents a detection performance comparison of mono-modalities, state-of-the-art multi-modalities, and DaFF on LLVIP dataset. Used performance metrics is average precision (AP %) at different IoU. The **bold blue** denotes best performance and **bold black** denotes second best. Results indicate the superiority of DaFF on LLVIP dataset.

Figure 6 and 7 illustrate successful examples of LLVIP validating the effectiveness of DaFF in detecting pedestrians against mono-modalities and other multi-modalities.

Overall, the detection performance on the LLVIP dataset is higher than the KAIST dataset. Indeed, many factors can contribute to this performance difference, including different image resolution, multispectral image misalignment problems, and the problematic annotation. Furthermore, the transformer-based fusion mechanisms, namely CFT and ICAFusion, are always the second best compared to the best

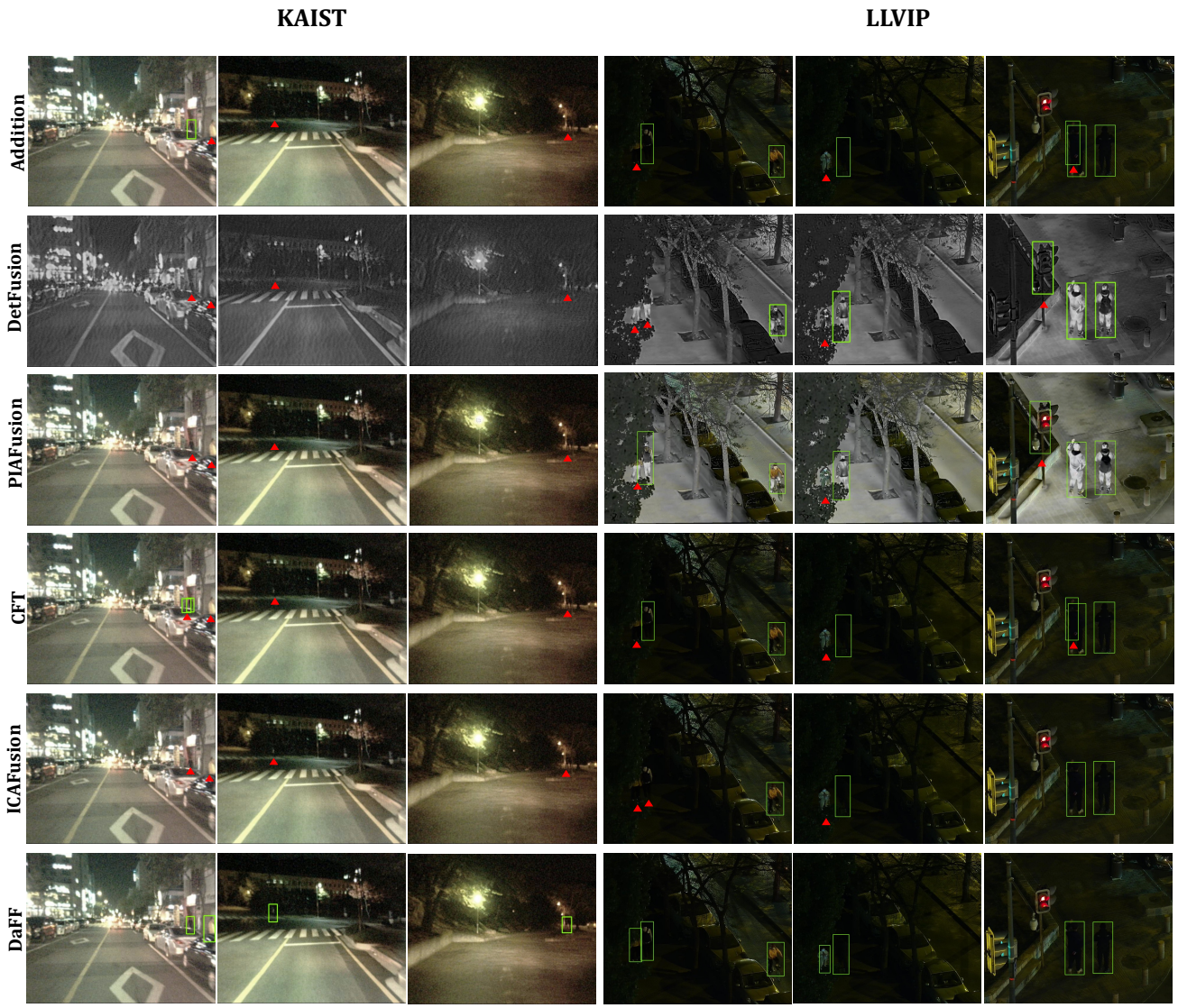


Figure 7. Qualitative comparison of detection performance between the DaFF and the other state-of-the-art fusion approaches on KAIST and LLVIP datasets. The green boxes denote detected pedestrians, and the red triangles indicate failure cases, false positives, or false negatives. Results show that DaFF outperforms other approaches. Zoom in for more details.

DaFF. This observation provides an empirical evidence of the power of the transformer in multi-modal learning.

## 6. Conclusion

In this paper, we introduce DaFF, a new feature fusion framework for multispectral pedestrian detection. The proposed DaFF leverages the integration of transformer, spatial attention, and channel-wise attention to improve the feature representation and fusion quality. DaFF provides a comprehensive attention-based framework that assists in combining RGB and IR features and enhances the performance degradation of nighttime pedestrian detection. The evaluation results reveal that the proposed method outperforms mono-modalities and state-of-the-art multi-modalities on KAIST and LLVIP datasets. We believe that fusing the complementary properties of RGB and IR modalities is a practical solution to close the gap between daytime and nighttime pedestrian detection performance.

## References


[1] Jeonghyun Baek, Sungjun Hong, Jisu Kim, and Euntai Kim. Efficient pedestrian detection at nighttime using a thermal camera. *Sensors*, 17(8):1850, 2017. 1

[2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 3

[3] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 4

[4] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8126–8135, 2021. 2

[5] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, pages 139–158. Springer, 2022. 2

[6] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661, 2022. 1

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[9] 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. 4

[10] 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. 5

[11] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12605–12614, 2020. 1

[12] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021. 1, 3, 5

[13] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 49–56, 2017. 5

[14] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015. 2

[15] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 3

[16] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference (BMVC)*, 2018. 5

[17] 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. 2, 5

[18] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. 1

[19] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519, 2019. 4

[20] 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*. BMVA Press, 2016. 2, 3, 5

[21] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021. 2, 3, 6, 7

[22] Shi Qiu, Saeed Anwar, and Nick Barnes. Pu-transformer: Point cloud upsampling transformer. In *Proceedings of the Asian Conference on Computer Vision*, pages 2475–2493, 2022. 2

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[24] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 5

[25] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, page 109913, 2023. 2, 3, 6, 7

[26] Mustafa Shukor, Guillaume Couairon, Asya Grechka, and Matthieu Cord. Transformer decoders with multimodal regularization for cross-modal food retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4567–4578, 2022. 1

[27] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4003–4011, 2022. 2, 6, 7

[28] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022. 2, 6, 7

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[30] Sunny Verma, Chen Wang, Liming Zhu, and Wei Liu. Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2019. 1

[31] Jörg Wagner, Volker Fischer, Michael Herman, Sven Behnke, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*, pages 509–514, 2016. 2

[32] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020. 4

[33] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020. 3

[34] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer vision*, pages 16269–16279, 2021. 2

[35] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 3

[36] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019. 3

[37] 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. 2

[38] 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. 2

[39] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 1

[40] 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. 5

[41] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 14–24. Springer, 2021. 3

[42] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 787–803. Springer, 2020. 2, 5, 6