Multimodal Object Detection by Channel Switching and Spatial Attention

Yue Cao
University of British Columbia
Kelowna, BC, Canada
caoyuecc@mail.ubc.ca

Junchi Bin
University of British Columbia
Kelowna, BC, Canada
junchi.bin@ubc.ca

Jozsef Hamari
TerraSense Analytics
Kelowna, BC, Canada
jozsef.hamari@terrasense.ca

Erik Blasch
MOVEJ Analytics
Fairborn, OH, USA
erik.blasch@gmail.com

Zheng Liu
University of British Columbia
Kelowna, BC, Canada
zheng.liu@ubc.ca

Abstract

Multimodal object detection has attracted great attention in recent years since the information specific to different modalities can complement each other and effectively improve the accuracy and stability of the detection model. However, compared to processing the inputs from a single modality, fusing information from multiple modalities can significantly increase the computational complexity of the model, thus impairing its efficiency. Therefore the multimodal fusion module needs to be carefully designed to enhance the performance of the detection model while keeping the computational consumption low. In this paper, we propose a novel lightweight fusion module that can efficiently fuse the inputs from different modalities using channel switching and spatial attention (CSSA). The effectiveness and generalizability of the module are tested using two public multimodal datasets LLVIP and FLIR, both of which comprise paired infrared (IR) and visible (RGB) images. The experiments demonstrate that the proposed CSSA module can substantially improve the accuracy of multimodal object detection without consuming excessive computing resources.

1. Introduction

Object detection, an integral branch of computer vision, is widely used in real-world applications. However, unimodal object detection, which is restricted by environmental factors, is sometimes insufficient for all realistic scenarios [18, 33]. For example, the quality of RGB images will be severely compromised in low light conditions, thus affecting the detection accuracy. In tasks that require high accuracy and robustness, such as autonomous driving and traffic monitoring, fusing signals from multiple modalities has become a typical way to improve the performance of the model [2, 7, 22, 28]. One common combination is to fuse RGB images with IR images as they are complementary. RGB cameras can obtain more details of an object when the light is sufficient, but provide little help in dim light conditions. IR images, on the other hand, can ensure that the contour of the object can be obtained in poorly lit or obscured situations, but information such as the texture and color of the object is absent. Figure 1 shows the paired RGB and IR samples from the FLIR dataset [30].

Figure 1. Daytime (top row) and nighttime (bottom row) samples from the FLIR dataset [30]. Objects in the bounding box indicate that the current modality can obtain more details and can be used as a complement to another modality.

A widely employed fusion method in recent research is mid-fusion, also known as feature fusion, where two back-

bone networks are used to extract feature maps from the input modalities separately, and then the feature maps are fused using a fusion model [4, 8, 10, 12]. Mid-fusion allows the detection model to extract detailed information from each input, thus achieving a better performance. However, the additional backbone network also adds more parameters to the model. To discover the underlying associations between different inputs, many researchers used complex fusion modules such as illumination-aware [8], or self-attention [4] to adequately exploit the information from both modalities. These additional modules lead to higher computational complexity, hence limiting the deployment of the model.

In this paper, a lightweight multimodal fusion module that uses channel switching and spatial attention (CSSA) is proposed to address the problems mentioned above. The proposed module can significantly improve the performance of multimodal object detection through channel switching and spatial attention without compromising efficiency. Specifically, channel switching replaces the feature map of each modality that has less impact on the detection result with the corresponding feature maps of another modality. The channel switching process allows each modality to retain its unique features while effectively fusing the features of other modalities. To enhance the spatial attention of the model without introducing additional parameters, max and average pooling are used to assess the importance of each location in the feature map from the channel dimension. Our experiment on the FLIR [30] and LLVIP [11] datasets demonstrates that compared to the recently proposed multimodal detection models, CSSA can significantly improve detection performance while consuming fewer computational resources. The contributions of this paper are summarized as follows:

- To the best of our knowledge, we are the first to introduce channel switching into multimodal object detection and demonstrate its effectiveness. In addition, we combine channel switching with spatial attention to enable the detection model to analyze the input modalities from both channel- and spatial-levels and thus achieve state-of-the-art performance.
- We propose a parameter-free spatial attention module that can efficiently assign weights to different inputs without increasing the complexity of the detection model, which can then be employed for time-critical tasks.
- We conduct extensive experiments on two public multimodal datasets and demonstrate the generalizability of the proposed model.

The rest of the paper is organized as follows: Section 2 reviews work related to multimodal object detection and the fusion strategies in multimodal object detection. Section 3 describes the details of the proposed model. The design of the experiments and the ablation study are presented in section 4. Finally, the conclusion is provided in section 5.

2. Related Work

In multimodal object detection tasks, the fusion strategy is a key factor affecting the performance of detection, as it determines the overall structure of the multimodal detector. In this section, we first present the work related to multimodal object detection, and then we dive into fusion strategies for multimodal detection.

2.1. Multimodal Object Detection

Benefiting from the growing availability of multimodal data, multimodal object detection has become an active research topic in computer vision. The FLIR [1] and LLVIP [11] datasets are two important benchmarks for multimodal object detection tasks, facilitating research on IR and RGB image fusion. Since the objective of multimodal detection is the same as unimodal detection, many related studies are based on the traditional RGB detection models such as RetinaNet [17], and YOLO [21]. Heng et al. [10] uses RetinaNet as the detection framework, where an extra backbone network is added as a feature extractor for the IR modality, and three convolution blocks are applied to input modalities from the spatial perspective to achieve inter- and intra-attention. For [20], the multimodal detector is based on YOLOv5, using a cross-modality attentive feature fusion module to identify the correlation between the input modalities. Additionally, [8, 14] utilize illumination-aware modules that allow the detection model to adjust weights to different input modalities based on light conditions, further improving the model's performance.

2.2. Fusion Strategies in Multimodal Object Detection

Multimodal object detection can be divided into three categories in light of the fusion strategy, including early-fusion, late-fusion, and mid-fusion [5, 13, 19, 24, 29].

Early-fusion (pixel fusion) is the most intuitive fusion approach, where IR and RGB images are concatenated to generate a 4-channel image, which is then fed into a regular object detection architecture such as [16, 17, 21]. Previous studies [12, 24] have revealed that fusing two inputs from an early stage forgoes features specific to each input, thus reducing detection accuracy.

In late-fusion (decision fusion), the inputs of the two modalities are fed separately into two unimodal object detection models to generate bounding boxes, and the predicted bounding boxes are fused using statistical methods [3, 13, 31]. For example, [3] uses Probabilistic Ensembling, which allows the model to cope with unaligned data. As

only the bounding boxes are fused, this approach is efficient but requires a highly precise unimodal object detection model.

Lastly, mid-fusion (feature fusion) adopts the two-backbone structure to facilitate the separate processing of different inputs and then the extracted feature maps are fused by fusion modules. Most research on multimodal object detection focuses on mid-fusion [10, 12, 20, 23], as this strategy provides greater flexibility in the design of fusion modules specific to inputs, thus allowing the detection model to explore deeper correlations between input modalities. However, these additional modules significantly increase the complexity of the detection model, resulting in high memory usage and latency. The authors of UA-CMDet [23] report in their paper that the time required by UA-CMDet to process a single input is 370ms, which is an unacceptable inference time for many time-critical tasks. To deal with this problem, GAFF [10] uses ResNet18 as the backbone and only focuses on spatial attention in the fusion module. The lightweight backbone, coupled with the absence of channel attention, not only cuts down the size of the model but also compromises the quality of the fusion. To overcome the aforementioned problems, we propose CSSA, which can take into account both channel and spatial level attention while ensuring low computational cost.

3. Methodology

The multimodal object detection model employed in this paper is adapted based on Faster R-CNN [16]. The detailed structure of the proposed module can be found in Figure 2.

3.1. Framework Overview

Faster R-CNN is adopted as our object detection framework, as it is a two-stage detection framework that can achieve high accuracy. As shown in Figure 2 (a), the model takes IR and RGB images as input and two ResNet 50s [9] are used as the backbone network, each of which contains four stages. Four CSSA modules are used to fuse the feature maps generated from each stage, and each CSSA module contains two sub-modules: channel switching and spatial attention. During channel switching, the weight of each channel from the input feature maps is evaluated by the Efficient Channel Attention (ECA) layer [25], and the channel with insignificant information for the final prediction is replaced with the corresponding channel from another modality. After channel switching, the spatial attention module calculates the significance of each location in the feature map using two channel-wise pooling operations and produces a fused feature map by summation operations (see Section 3.2). Finally, the fused feature maps are then fed into the Feature pyramid network (FPN) [16] and the detection head to generate the bounding boxes.

3.2. Proposed Fusion Module

3.2.1 Channel Switching

Channel attention can enrich the multimodal fusion module with feature-level information, allowing the module to learn the features shared between modalities while retaining features exclusive to the modality. To perform channel attention, channel switching is applied in our module, as it is efficient and effective for feature interaction across modalities [26, 32]. The first step of channel switching is to assign weights to the feature maps of each modality from the channel dimension. To ensure the efficiency of the module, we choose the ECA block [25], which consists of a global average pooling (GAP), a 1D convolution, and a sigmoid function. ECA can perceive local cross-channel interaction efficiently and can be described as:

$$\omega_m = \sigma\left(f\left(\mathrm{GAP}\left(X_m\right)\right)\right) \tag{1}$$

where

$$\mathrm{GAP}(X) = \frac{1}{HW}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1} X_{ij} \tag{2}$$

where $X_m \in \mathbb{R}^{H\times W\times C}$ is the feature maps obtained from modality m (a height $\times$ width $\times$ channel tensor), and $f(X)$ is a 1D convolution layer. The σ symbol denotes the sigmoid activation function. ECA [25] first uses GAP to downscale the input feature map to a $1\times1\times C$ vector and then a 1D convolution is used to obtain cross-channel interaction. Finally, the weight is calculated by a sigmoid function. The weights $\omega_m \in \mathbb{R}^{1\times1\times C}$ acquired from the ECA block are subsequently used for channel switching. The switching process can be represented as:

$$\begin{cases} X_{m,c} & if\ \omega_{m,c} \geqslant k \\ X_{m',c} & if\ \omega_{m,c} < k \end{cases} \tag{3}$$

where m' represents the modality from another input; c denotes c-th channel; k is a predefined threshold. When the weight of the c-th channel is below the value of the threshold k, the model replaces it with the corresponding channel of another modality.

3.2.2 Spatial Attention

The objective of spatial attention is to highlight the locations in a feature map containing core information, which can be used as a complement to channel switching. To achieve this goal, we utilize two parameter-free operations: channel-wise average pooling (CAP) and channel-wise max pooling (CMP). These two operations can effectively condense the information in the feature map without increasing the complexity of the module. CSSA first concatenates the two feature maps obtained from the channel switching block to form $X_{cat} \in \mathbb{R}^{H\times W\times 2C}$, which is then fed

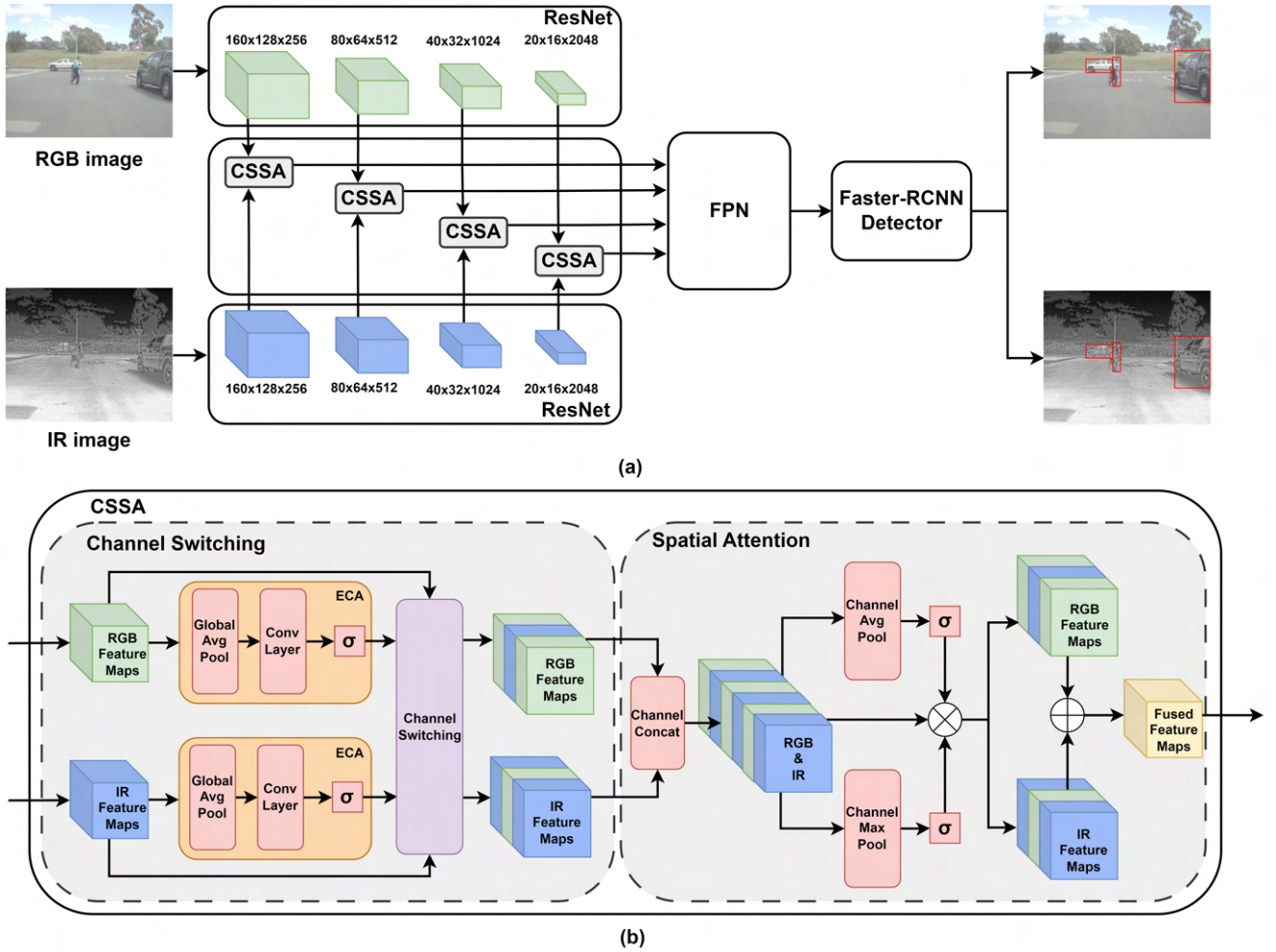

Figure 2. Overview of the proposed model. (a) shows the overall architecture of the detection model. (b) illustrates the detailed structure of CSSA. σ symbol denotes the sigmoid function, $\otimes$ represents element-wise multiplication and $\oplus$ means summation operation.

into CAP and CMP to acquire the attention map ω_{avg}, $\omega_{max} \in \mathbb{R}^{H\times W\times 1}$. Next, a weighted feature map X^w_{cat} is produced by performing element-wise multiplication between X_{cat}, ω_{avg}, and ω_{max}. The procedure can be precisely described as:

$$X^w_{cat} = X_{cat} \otimes \text{CAP}(X_{cat}) \otimes \text{CMP}(X_{cat}) \tag{4}$$

where

$$\text{CAP}(X) = \frac{1}{c}\sum_{c=0}^{c-1} X^c_{ij} \tag{5}$$

$$\text{CMP}(X) = \max(x^1_{ij}, ..., x^c_{ij}) \tag{6}$$

where $\otimes$ denotes element-wise multiplication. Lastly, the process that produces the final fusion result can be formulated as:

$$X_{fused} = \frac{X^w_{IR} + X^w_{RGB}}{2} \tag{7}$$

where

$$X^w_{IR}, X^w_{RGB} = \text{Split}(X^w_{cat}) \tag{8}$$

Here, the weighted feature maps $X^w_{IR}, X^w_{RGB} \in \mathbb{R}^{H\times W\times C}$ have the same dimensions as the input feature maps.

4. Experiments

Experiments are conducted on FLIR [30] and LLVIP [11] datasets to compare the overall performance of the proposed model with the baselines in terms of precision and efficiency. In addition, we have investigated the impact of different sub-blocks of the CSSA module on prediction results.

Method	Modality	FLIR			LLVIP			Average		
		AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75	mAP
Faster R-CNN	IR	73.4	34.2	37.9	92.6	48.8	50.7	83.0	41.5	44.3
Faster R-CNN	RGB	65.0	22.8	30.2	88.8	45.7	47.5	76.9	34.3	38.9
Halfway Fusion	RGB+IR	71.5	31.1	35.8	91.4	60.1	55.1	81.5	45.6	45.5
GAFF	RGB+IR	74.6	31.3	37.4	94.0	60.2	55.8	84.3	45.8	46.6
ProbEn	RGB+IR	75.5	31.8	37.9	93.4	50.2	51.5	84.5	41.0	44.7
CSAA (ours)	RGB+IR	**79.2**	**37.4**	**41.3**	**94.3**	**66.6**	**59.2**	**86.8**	**52.0**	**50.3**

Table 1. The evaluation results on two public datasets measured by AP in percentage.

4.1. Experimental Setup

4.1.1 Datasets

The FLIR dataset [1] is one of the most widely used datasets for multimodal object detection tasks. It contains RGB and IR image pairs collected from the perspective of automobile drivers, and these data contain both day and night scenarios. In this experiment, we use the FLIR dataset proposed by [30], as the RGB and IR images are heavily unaligned in the original version [1]. This version contains 5,142 well-aligned RGB-IR data pairs, of which 4,129 are used for training and 1,013 for testing. The aligned-FLIR involves four types of objects, including 8,987 people, 20,608 cars, 2,566 bicycles, and 95 dogs. We removed all dog labels as they are not adequate for training, so the objects left are "people", "cars", and "bicycles".

Besides FLIR [30], the LLVIP dataset [11], a recently released multimodal object detection dataset, is also used in the experiments. LLVIP contains RGB-IR image pairs captured by surveillance cameras in 26 different locations, with most of the data collected in dimly lit conditions. The dataset includes a total of 15,488 semi-manually aligned data pairs, 12,025 pairs of which are used for training and 3,463 for testing. "Pedestrian" is the only object category in the dataset.

4.1.2 Implementation details

Our CSSA detection model is adapted based on the Faster R-CNN model [16] from the Detectron2 library [27] and trained on a single NVIDIA Tesla V100 GPU. The backbones applied in the experiment are two ResNet 50s [9] pre-trained on ImageNet [6]. To train the model, we adopt random flipping and resize the data from both datasets to 640 $\times$ 512. The optimizer employed is AdamW with a learning rate of 2.5e-4, and the model is trained for 10 epochs with a batch size of 16. The hyper-parameter k mentioned in Equation 3 is set to 2e-3.

We employ three competing multimodal object detection models as the baseline, including Halfway fusion [12], GAFF [10], and ProbEn [3], and apply them on both FLIR [30] and LLVIP [11] datasets in our comparative study. Moreover, Faster R-CNN [16] is also seen as one of the baselines to prove the effectiveness of the fusion. For a fair comparison, the experimental setting for all the baselines is identical to ours.

4.1.3 Evaluation Metrics

To quantitatively compare the proposed model with the baselines, we utilize the Average Precision (AP) computed by the COCO evaluation metric [15], in which the Intersection Over Union (IOU) is measured between the ground truth and predicted bounding box, and a prediction is only considered as true positive when IOU $\geqslant$ a threshold. The COCO evaluation metric [15] calculates the AP50 and AP75 of the model with IOU threshold $\geqslant$ 0.5 and 0.75, and the mean Average Precision (mAP) is obtained using the IOU value ranges from 0.5 to 0.95 with a step of 0.05. Furthermore, the complexity of the models is evaluated by counting the inference time for a single image and the memory consumption when the model is loaded.

4.2. Comparative Studies

4.2.1 Quantitative Results

The performance of the proposed model and the baselines mentioned above are evaluated on the FLIR [30] and LLVIP [11] datasets. The results are shown in Table 1. Compared to the unimodal object detection models, our CSSA model outperforms Faster R-CNN (RGB) [16] by 14.2% and Faster R-CNN (IR) by 5.8% on AP50. The result demonstrates that CSSA can effectively fuse the information from both modalities to improve detection results. For multimodal object detection models, our model outperforms Halfway fusion [12] and GAFF [10], by 7.7% and 4.6% on AP50, respectively, as Halfway fusion does not assign attention to the input modalities and GAFF only considers spatial-level attention. Our model involves both feature- and spatial-level attention, thus greatly improving

the performance of detection. Lastly, CSSA also exceeds the latest model ProbEn [3] on AP50, AP75, and mAP. In our experiments, the AP of ProbEn is lower than what the ProbEn authors reported. A possible reason for the difference is that the authors of ProbEN use the original FLIR dataset [1], and the backbone they applied is ResNet 101 [9]; however, we replaced the backbone of all the baselines with ResNet 50 for a fair comparison with our model.

Different from FLIR, which captures images from the perspective of drivers, LLVIP [11] collects data using traffic cameras. Therefore, the experiment of LLVIP confirms that our model can be applied to different realistic scenarios. The evaluation results are reported in Table 1. Compared to the unimodal object detection models, the proposed model outperforms Faster R-CNN (IR) [16] by 8.5% and Faster R-CNN (RGB) by 11.7% on mAP. Our CSSA model also outperforms Halfway fusion [12] by 4.1%, GAFF [10] by 3.4%, and ProbEN by 7.7% on mAP, respectively.

In addition, the average AP scores of the two datasets are also calculated. As shown in Table 1, the proposed model still outperforms all the baselines, with an average AP50 score of 86.8%, and mAP score of 50.3%. These results further demonstrate the strong generalizability of the CSSA model and its ability to achieve state-of-art performance on both datasets.

4.2.2 Qualitative Results

On top of a quantitative evaluation, we also perform a qualitative evaluation on the FLIR dataset [30]. Figure 3 shows the detection results of the proposed model and the baselines. The results prove the complementarity between the IR and RGB images. For example, in Figure 3 (a), the IR detector fails to capture the bicycle because of the subtle difference in temperature between the object and the environment; however, the RGB image captures details of texture and color of the bicycle, thus enabling the Faster R-CNN (RGB) [16] to correctly detect the object. On the other hand, in Figure 3 (c), the RGB detector, in stark contrast to IR, misses most pedestrians.

For multimodal detectors, ProbEn [3] uses a late-fusion strategy, which allows the model to capture the object detected by either IR or RGB detectors and to achieve the lowest miss rate in this evaluation. Nevertheless, ProbEn needs to handle more overlapping detections. Experimental results show that ProbEn sometimes accidentally retains overlapping detections, leading to more false positive outputs, and therefore a higher false positive rate than that of our model. On the other hand, CSSA, GAFF, and Halfway fusion [10, 12] apply the mid-fusion strategy, which can make full use of the information from both input modalities. Thanks to the CSSA module, our detection model can extract information from both feature and spatial perspectives, resulting in the lowest miss rate and false-positive rate among the three models.

4.2.3 Comparison in Computational Efficiency

The analysis of memory consumption and inference time of each model is illustrated in Figure 4. Since Faster R-CNN (RGB) and Faster R-CNN (IR) use the same architecture, they together are referred to as Faster R-CNN. All models in the experiments are implemented based on Detectron2 [27], and the evaluation is performed on an NVIDIA 3080ti GPU.

As Faster R-CNN is employed for unimodal object detection, it only requires 972MB when the model is loaded. For the multimodal object detection models, Halfway fusion [12] and GAFF [10] require 3553MB and 4474MB respectively, as they both use the mid-fusion strategy, where the additional backbone and the complex fusion module consume extra memory. ProbEn [3] adopts Probabilistic Ensembling as the fusion method, which is a parameter-free process. However, ProbEn requires two unimodal object detectors to be loaded simultaneously, so the memory consumption is doubled. In contrast, the proposed model requires only 1575MB, outperforming all the multimodal object detectors. The design of the dual backbone is also applied to our CSSA model, but to address the problem of excessive memory cost, we adopt a super lightweight fusion module where only the one-dimensional convolution layers can incur additional memory load.

According to the results in Figure 4 (b), Faster R-CNN is the most efficient model, requiring only 23ms to complete the inference for a single input. Halfway fusion [12] and GAFF [10], as the models with the highest complexity, require 42ms and 61ms to complete the prediction, respectively. Thanks to the simplicity of the fusion module, ProbEn achieves the fastest speed among the multimodal detection models; it only takes 2ms to compute the fused bounding boxes and 25ms for the entire prediction process. Finally, the inference time required by the CSSA model is 31ms, which is an acceptable result compared to that of ProbEn [3] and Faster R-CNN [16] as the proposed model has higher accuracy and less memory usage.

4.3. Ablation study

In this section, we conduct an ablation study to verify our model design. CSSA consists of two sub-blocks, channel switching and spatial attention, and we investigate the performance of the detection model by applying the two sub-blocks separately as a fusion module and report the results in Table 2. The results indicate that both sub-blocks, compared to unimodal object detectors, are effective in increasing prediction accuracy. Channel switching adequately fuses features of the two modalities while retaining features specific to each modality, and spatial atten-

Figure 3. The detection results of the six models mentioned in the experiment. The cars, pedestrians, and bicycles detected by the model are represented by blue, yellow, and green bounding boxes, respectively. The red bounding boxes represent the error case (false positive and false negative).

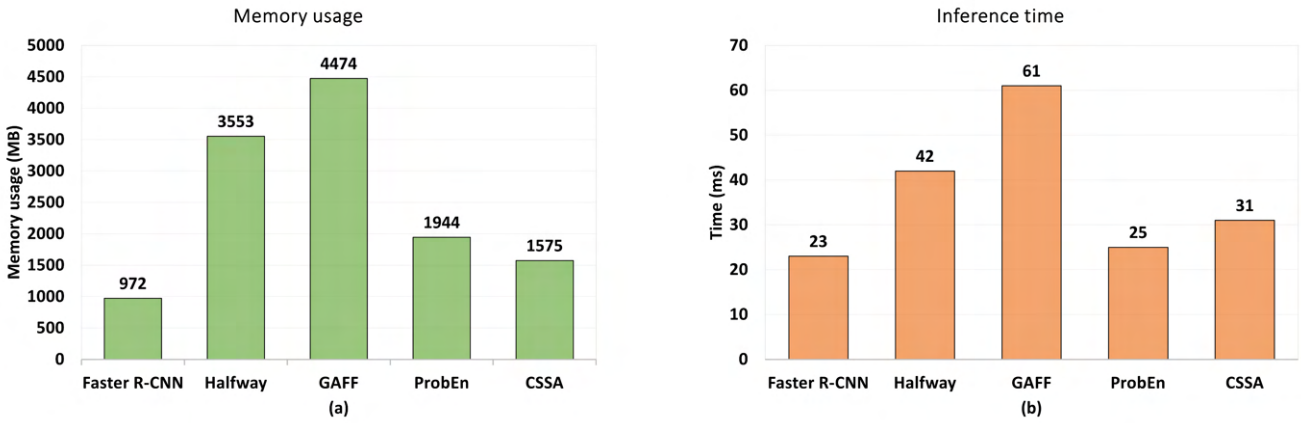

Figure 4. The evaluation results of different methods on memory usage (a) and inference time(b).

Method	AP50	AP75	mAP
Channel Switching	75.8	32.1	37.8
Spatial Attention	74.2	31.4	36.5
CSSA (ours)	**79.2**	**37.4**	**41.3**

Table 2. The results when only channel switching or spatial attention is applied.

tion captures the correlation between input modalities from a space perspective. Therefore, we conclude that these two modules can complement each other to further improve the performance of the model.

5. Conclusion and Future Work

In this paper, a lightweight multimodal fusion operator (CSSA) is proposed and applied to the multimodal object detection task. Our study proves that CSSA can effectively capture the information from both RGB and IR modalities compared to other methods. We also show that both the channel switching and spatial attention blocks in the proposed module can significantly improve detection accuracy and that their combination can further improve predictions as both feature level and spatial level attention are considered.

Additionally, the lightweight design enables the CSSA model to process 33 frames per second, which meets the requirements of most real-time object detection applications.

In the future, the generalizability of CSSA can be further explored by applying it to different detection frameworks. In addition, further research could be conducted to explore the feasibility of setting the channel switching threshold as a trainable parameter rather than requiring manual adjustment.

References

[1] Free Teledyne FLIR thermal dataset for algorithm training, 2018. https://www.flir.ca/oem/adas/adas-dataset-form. 2, 5, 6

[2] Thiemo Alldieck, Chris H. Bahnsen, and Thomas B. Moeslund. Context-aware fusion of rgb and thermal imagery for traffic monitoring. *Sensors*, 16(11), 2016. 1

[3] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 139–158. Springer Nature Switzerland, 2022. 2, 5, 6

[4] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2022. 2

[5] Hangil Choi, Seungryong Kim, Kihong Park, and Kwanghoon Sohn. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 621–626, 2016. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. 5

[7] Chen Fu, Christoph Mertz, and John M. Dolan. Lidar and monocular camera fusion: On-road depth completion for autonomous driving. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 273–278, 2019. 1

[8] 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. 2

[9] 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)*, pages 770–778, 2016. 3, 5, 6

[10] Zhang Heng, Fromont Elisa, Lefevre Sébastien, and Avignon Bruno. Guided attentive feature fusion for multispectral pedestrian detection. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 72–80, 2021. 2, 3, 5, 6

[11] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3489–3497, 2021. 2, 4, 5, 6

[12] Shu Wang Jingjing Liu, Shaoting Zhang and Dimitris Metaxas. Multispectral deep neural networks for pedestrian detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 73.1–73.13, September 2016. 2, 3, 5, 6

[13] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference (BMVC)*, September 2018. 2

[14] 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

[15] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 5

[16] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 3, 5, 6

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 2

[18] Shuo Liu and Zheng Liu. Multi-channel cnn-based object detection for enhanced situation awareness. *arXiv preprint arXiv:1712.00075*, 2017. 1

[19] Kieu My, Bagdanov Andrew D., Bertini Marco, and del Bimbo Alberto. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In *Computer Vision – ECCV 2020*, pages 546–562. Springer International Publishing, 2020. 2

[20] Fang Qingyun and Wang Zhaokui. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 130:108786, 2022. 2, 3

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 2

[22] Ivana Shopovska, Ljubomir Jovanov, and Wilfried Philips. Deep visible and thermal image fusion for enhanced pedestrian visibility. *Sensors*, 19(17), 2019. 1

[23] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. 3

[24] Jörg Wagner, Volker Fischer, Michael Herman, Sven Behnke, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*, volume 587, pages 509–514, 2016. 2

[25] 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 (CVPR)*, June 2020. 3

[26] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4835–4845. Curran Associates, Inc., 2020. 3

[27] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 5, 6

[28] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M. López. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):537–547, 2022. 1

[29] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4236–4244, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 2

[30] 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, 2020. 1, 2, 4, 5, 6

[31] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5126–5136, 2019. 2

[32] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[33] Yufeng Zheng, Erik Blasch, and Zheng Liu. *Multispectral image fusion and colorization*, volume 481. SPIE press Bellingham, Washington, 2018. 1