

Rethinking Multi-modal Object Detection from the Perspective of Mono-Modality Feature Learning

Tianyi Zhao^{1*}, Boyang Liu^{1*}, Yanglei Gao¹, Yiming Sun², Maoxun Yuan^{1†}, Xingxing Wei^{1†}
¹Institute of Artificial Intelligence, State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, ²Southeast University

ty_zhao@buaa.edu.cn, by.liu2004@gmail.com, gaoyl04@buaa.edu.cn
 sunyiming@seu.edu.cn, {yuanmaoxun, xxwei}@buaa.edu.cn

Abstract

Multi-Modal Object Detection (MMOD), due to its stronger adaptability to various complex environments, has been widely applied in various applications. Extensive research is dedicated to the RGB-IR object detection, primarily focusing on how to integrate complementary features from RGB-IR modalities. However, they neglect the mono-modality insufficient learning problem, which arises from decreased feature extraction capability in multi-modal joint learning. This leads to a prevalent but unreasonable phenomenon—Fusion Degradation, which hinders the performance improvement of the MMOD model. Motivated by this, in this paper, we introduce linear probing evaluation to the multi-modal detectors and rethink the multi-modal object detection task from the mono-modality learning perspective. Therefore, we construct a novel framework called M^2D-LIF , which consists of the Mono-Modality Distillation (M^2D) method and the Local Illumination-aware Fusion (LIF) module. The M^2D-LIF framework facilitates the sufficient learning of mono-modality during multi-modal joint training and explores a lightweight yet effective feature fusion manner to achieve superior object detection performance. Extensive experiments conducted on three MMOD datasets demonstrate that our M^2D-LIF effectively mitigates the Fusion Degradation phenomenon and outperforms the previous SOTA detectors. The codes are available at <https://github.com/Zhao-Tian-yi/M2D-LIF>.

1. Introduction

Recently, multi-modal object detection (MMOD) technology has been widely used in various safety-critical applications such as around-the-clock pedestrian detection for urban surveillance [10], object detection for autonomous driving [21] etc. Since relying solely on visible images [30, 40]

*: Equal contribution. †: Corresponding Author.

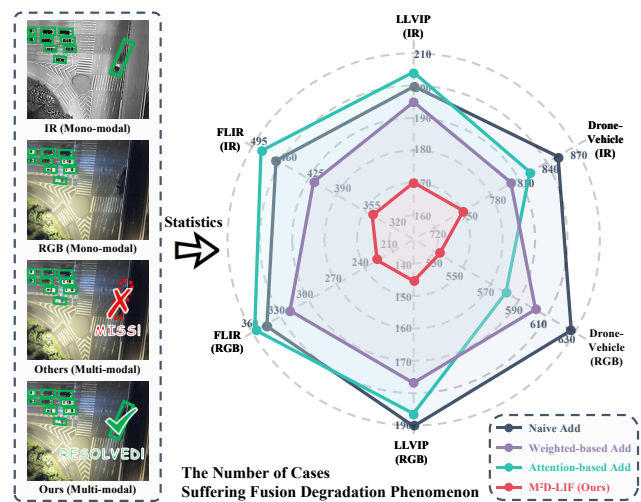


Figure 1. Fusion Degradation phenomenon. (Left) An instance of this phenomenon. The object highlighted by the red box is detected by a mono-modal method but missed by the multi-modal method. (Right) The Fusion Degradation phenomenon statistics of three methods and our method across different datasets.

will render objects invisible under limited illumination, utilizing both visible (RGB) images and infrared (IR) images has been widely used as one of the effective solutions to achieve full-time object detection tasks. Specifically, RGB modality can provide the texture of objects in daylight, while infrared modality can provide the outline of objects under poor lighting conditions. By exploring the complementary information between the two modalities, RGB-IR object detection [15, 42] can aggregate the advantages of each modality to achieve robust visual perception.

To design advanced fusion strategies for the MMOD task, previous works [13, 17, 29] reveal that the feature-level fusion methods outperform both image-level and decision-level fusion methods. Building upon this finding, recent RGB-IR object detection methods mainly focus on designing complex feature fusion structures to address vari-

ous challenges. For example, CSSA [2] and ICAFusion [25] introduce different attention mechanisms to better explore multi-modal complementarity. CALNet [7] proposes the selected cross-modal fusion module to address the semantic conflict issue between different modalities. AR-CNN [43, 44] and TSFADet [35, 37] focus on solving the misalignment problem and attempt to improve object detection performance by aligning RGB and IR modality features. Although these methods have achieved encouraging object detection performance, they primarily focus on integrating complementary features from RGB-IR modalities while neglecting the mono-modality insufficient learning problem. This results in an unreasonable phenomenon that some objects can be detected by the mono-modal detector but missed by the corresponding multi-modal detector. We name this phenomenon as Fusion Degradation as illustrated in Figure 1. Besides, we statistics this phenomenon on three commonly used MMOD datasets and find that it is prevalent in different types of multi-modal detectors. Motivated by this, we rethink the multi-modal object detection task from the mono-modality learning perspective.

In this paper, we construct **M²D-LIF**, an end-to-end framework for multi-modal object detection, to address the above problem. Different from the existing MMOD methods that design complex feature fusion modules, M²D-LIF facilitates the sufficient learning of mono-modality during multi-modal joint training and explore a lightweight yet effective feature fusion manner to achieve superior object detection performance. Inspired by knowledge distillation [8], we propose an **Mono-Modality Distillation (M²D)** method, which introduces the additional encoder pretrained on the mono-modality to distill the multi-modal encoders. To bridge the capability gap between the multi- and mono-modalities, we design the inner-modality and cross-modality distillation loss to jointly optimize the M²D-LIF during training. In this way, we can ensure sufficient learning of the mono-modal encoder in the multi-modality joint training. Besides, according to our observations in Section 3.1, we propose a **Local Illumination-aware Fusion (LIF)** module, which can dynamically set different weights for different illumination regions. The module ensures an explicitly complementary fusion to cooperate with M²D method, thereby enhancing both the accuracy and efficacy of the multi-modal object detection. Figure 1 shows that our framework can effectively solve the above problem. Our contributions in this paper are highlighted as follows:

- We introduce linear probing evaluation to the multi-modal detectors and identify the insufficient learning of mono-modality during training. To the best of our knowledge, it is the first time to rethink multi-modal object detection from a mono-modality learning perspective.
- We present M²D-LIF, a pioneering method to improve mono-modality learning capabilities during multi-modal

joint training, innovatively exploring a lightweight yet effective feature fusion framework for the MMOD task.

- Extensive experiments on the three MMOD datasets demonstrate that our M²D-LIF outperforms the previous state-of-the-art detectors and can be used as an effective feature fusion method in the MMOD task.

2. Related Works

2.1. RGB-Infrared Object Detection

Due to the deep exploration of complementary features through CNNs, feature-level fusion has become a widely used approach for RGB-Infrared object detection. In early studies, the weighted-based fusion methods have been used as a simple and effective way to fuse different features. Li *et al.* [14] proposed the first illumination-aware Faster R-CNN, which introduced illumination conditions into the RGB-Infrared object detection task. At the same time, Guan *et al.* [4] presented a multispectral pedestrian detection framework based on illumination-aware pedestrian detection and semantic segmentation. With the introduction of the attention mechanism, more and more attention-based fusion methods have been proposed to achieve complementary feature fusion. Fang *et al.* [22] proposed the cross-modality attentive feature fusion (CMAFF) module, leveraging common-modality and differential-modality attentions. Concurrently, Cao *et al.* [2] introduced the CSSA fusion module, which employs channel switching and spatial attention for feature fusion. Besides, CMX [41] was proposed to cross-modal feature rectification and feature fusion with intertwining cross-attention. Furthermore, Shen *et al.* [25] proposed a dual cross-attention transformers to model global feature interaction and capture complementary information between two modalities.

However, the above methods mainly focus on designing complex and parameter-intensive modules to achieve the so-called complementary fusion, while ignoring the insufficient representation of each modality caused by the multi-modal learning. Different from these approaches, in this paper, we conduct the RGB-IR object detection task from mono-modality feature learning perspective and propose a novel M²D-LIF framework that only introduces a lightweight feature fusion module to achieve superior object detection performance.

2.2. Knowledge Distillation for Object Detection

Knowledge Distillation (KD) [8] can transfer knowledge from a high-capacity teacher model to a compact student model, thereby reducing the complexity of the model while ensuring model performance. For the object detection task, FRS [47] and PGD [32] leveraged foreground masks to enhance the ability of the student model to capture critical object features. CWD [26] normalized activation maps of each

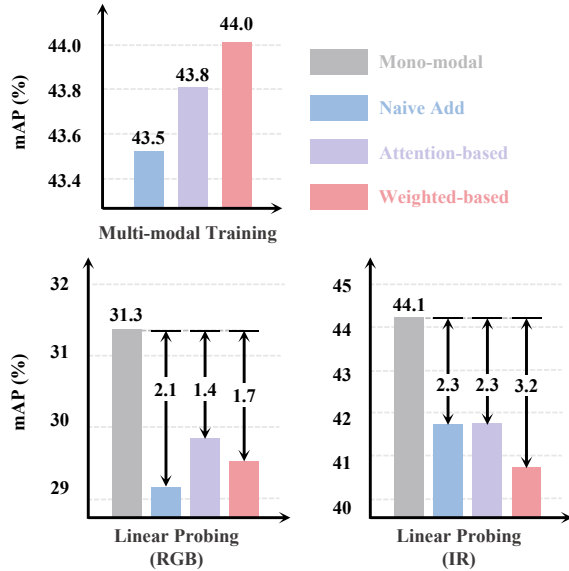


Figure 2. Linear probing evaluation on the FLIR dataset. Three types of feature fusion methods are selected for comparison, such as naive addition (Halfway Fusion [18]), weighted-based (IWM [9]), and attention-based (CMX [41]).

channel to obtain soft probability maps and minimized the KL divergence between these maps, enabling the student to focus on the most salient regions of each channel. Besides, PKD [1] imitated features with Pearson correlation coefficient to focus on the relational information from the teacher and relax constraints on the magnitude of the features. Recently, CrossKD [28] introduced cross-head knowledge distillation by transferring the intermediate features from the student detection head to the teacher providing a diversified perspective. As for the multi-modal object detection task, KD is usually used to transfer the knowledge from multi-modality features to mono-modality features. For example, distillation from Radar-Lidar to Radar [31], BEV-Lidar to BEV [6], and RGB-Infrared to RGB [20]. In our paper, unlike the above methods, we design a novel knowledge distillation method called Mono-Modality Distillation (M^2D) to improve the feature representation of each backbone network in multi-modality joint training.

3. Methodology

3.1. Linear Probing Evaluation

In the current MMOD framework, each modality is encoded by its corresponding backbone network, and then a fusion module is utilized to obtain the fused features for downstream perception tasks. Although achieving superior performance, we claim that this way will lead to insufficient learning of each modality and thus fail to achieve optimal detection performance. To validate this point, we employ the linear probing evaluation. Specifically, we first train the mono-modal object detectors and the multi-modal object

detectors, respectively. For multi-modal object detectors, we select three popular feature fusion methods for comparison, such as Halfway Fusion [18] (naive addition), IWM [9] (weighted-based), and CMX [41] (attention-based). Note that, all detectors utilize the CSPDarknet53 [11] as the backbone network. After that, each backbone from both the mono- and multi-modality detectors is frozen and connected with the new detection head for training and testing. As shown in Figure 2, the evaluation results reveal that:

(1) **The backbone networks from multi-modality joint training are insufficient learning.** After linear probing evaluation, we can observe that all backbone networks from multi-modality joint training are worse than those from mono-modality training. This indicates that due to the existence of the fusion module, the learning ability of each backbone network is limited during multi-modal training.

(2) **The weighted-based method can serve as a competitive fusion way to improve performance.** Compared with other fusion methods, the detection performance of the weighted-based backbone network decrease significantly. This indicates that the weighted-based fusion can achieve higher performance on the weaker backbone networks, which proves its effectiveness in the feature fusion.

3.2. M^2D -LIF Framework

From Section 3.1, we have observed that the recent MMOD methods still suffer from the insufficient learning of mono-modality features. To solve this issue, referring to the knowledge distillation technology, we consider utilizing the additional encoder pretrained on the mono-modality as the teacher model to distill the multi-modality backbone network during training. Thus, under the premise that the multi-modal encoder learning sufficiently, we design a novel weighted-based feature fusion method to further improve the performance. The overall framework (M^2D -LIF) is illustrated in Figure 3, which mainly consists of ① **Mono-Modality Distillation** and ② **Local Illumination-aware Fusion**. Note that the IR branch is omitted in Figure 3 since it has the same network structure as the RGB branch.

① **Mono-Modality Distillation (M^2D).** The M^2D aims to enhance the feature extraction capability of the multi-modal encoders, laying a solid foundation for improving object detection performance. To bridge the capability gap between the multi- and mono-modal encoders, we first employ a pretrained teacher model of the same modality to distill the multi-modal backbone network using the inner-modality distillation loss \mathcal{L}_{IM} . To further enhance object-relevant feature extraction, we propose a cross-modality distillation loss \mathcal{L}_{CM} , which leverages cross-modality salient object location priors to guide feature distillation.

Specifically, for the paired RGB image I_V and IR image I_I , we first input them into the teacher and student backbone

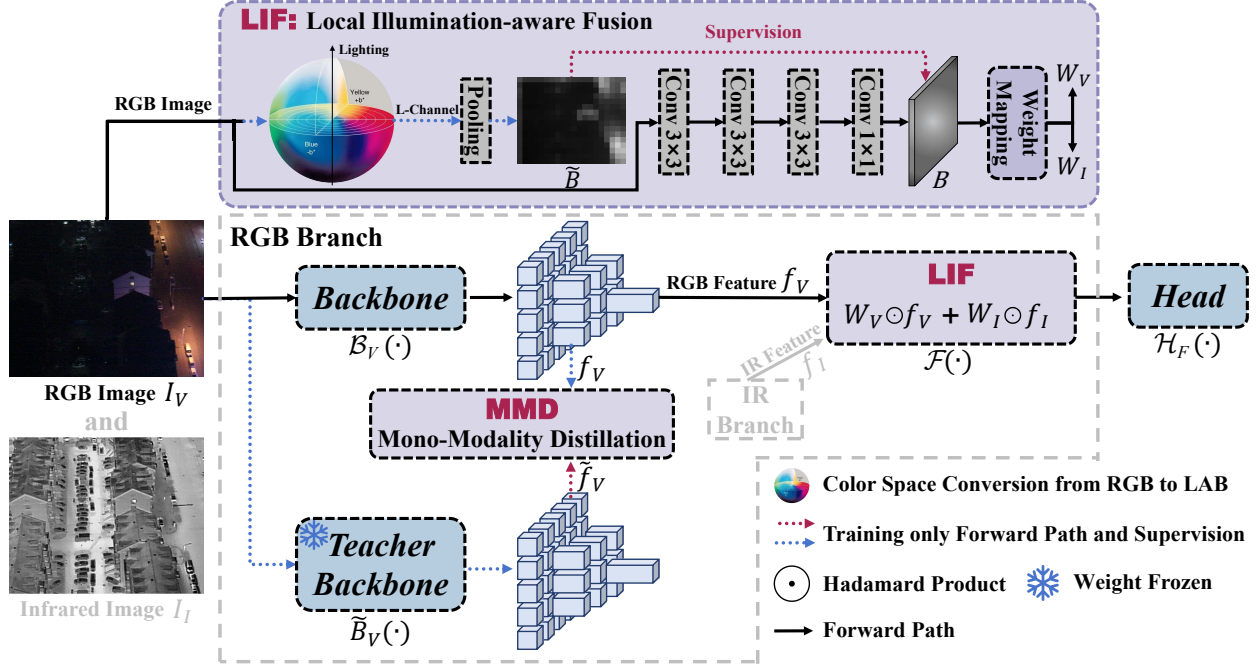


Figure 3. Overview of our M²D-LIF framework. The proposed M²D is for enhancing mono-modal feature extraction capability of the backbone in the multi-modal object detection task. The proposed LIF method can evaluate the mono-modal quality based on RGB image’s local illumination conditions. Noted that since our framework is designed from the mono-modality perspective, the design of the infrared branch and the RGB branch is symmetrical and has the same network structure. Therefore, only the image processing flow of the RGB modality is emphasized in the figure, and the infrared branch is omitted.

networks simultaneously, where \mathcal{B}_V and \mathcal{B}_I are the student backbone networks for multi-modal object detection, while $\tilde{\mathcal{B}}_V$ and $\tilde{\mathcal{B}}_I$ are the mono-modal teacher backbone networks. Therefore, the output features extracted by each backbone network can be expressed as:

$$\begin{aligned} \text{Student: } f_V &= \mathcal{B}_V(I_V), & f_I &= \mathcal{B}_I(I_I), \\ \text{Teacher: } \tilde{f}_V &= \tilde{\mathcal{B}}_V(I_V), & \tilde{f}_I &= \tilde{\mathcal{B}}_I(I_I), \end{aligned} \quad (1)$$

where f_I and f_V are the outputs of the student backbones. The \tilde{f}_V and \tilde{f}_I are the outputs of the teacher backbones. Based on these features, we design the Mono-modality Distillation loss to optimize the multi-modal backbone network. The Mono-modality Distillation loss consists of two components: the inner-modality feature distillation loss \mathcal{L}_{IM} and the cross-modality feature distillation loss \mathcal{L}_{CM} .

The loss function \mathcal{L}_{IM} represents the inner-modality feature distillation loss, which guides the backbone of the multi-modal object detection model to learn intermediate layer responses from a teacher backbone of the same modality. By minimizing \mathcal{L}_{IM} , the multi-modal backbone is encouraged to align with the feature responses of the teacher model. The loss is defined as follows:

$$\mathcal{L}_{IM} = D(f_V, \tilde{f}_V) + D(f_I, \tilde{f}_I), \quad (2)$$

where $D(\cdot, \cdot)$ denotes the specific distillation method.

As for the loss function \mathcal{L}_{CM} , it represents the cross-modality feature distillation loss, which integrates cross-modality target location priors. Specifically, in MMOD tasks, different modalities generally contain information about the same objects. We first employ an attention mechanism to extract the salient object feature attention map, which serves as the location prior. The attention map is then used as a mask to guide object-relevant feature learning. To maintain a lightweight design, we adopt the parameter-free attention method SimAM [33]. The attention map $\tilde{\mathcal{M}}$ can be calculated as:

$$\tilde{\mathcal{M}} = \text{Sigmoid}\left(\frac{(\tilde{f} - \tilde{\mu})^2 + 2\tilde{\sigma}^2 + 2\lambda}{4(\tilde{\sigma}^2 + \lambda)}\right), \quad (3)$$

where $\tilde{\mu}$ represents the mean of \tilde{f} across spatial dimensions, $\tilde{\sigma}^2$ represents the variance of \tilde{f} , and λ is a small positive constant added for numerical stability. The cross-modality feature distillation loss \mathcal{L}_{CM} is formulated as follows:

$$\mathcal{L}_{CM} = D(\tilde{\mathcal{M}}_V \odot f_I, \tilde{\mathcal{M}}_V \odot \tilde{f}_V) + D(\tilde{\mathcal{M}}_I \odot f_V, \tilde{\mathcal{M}}_I \odot \tilde{f}_I), \quad (4)$$

where $\tilde{\mathcal{M}}_V$ and $\tilde{\mathcal{M}}_I$ are the attention maps of different modalities. The overall loss function of M²D is defined as the sum of the inner- and cross-modality loss:

$$\mathcal{L}_{M^2D} = \mathcal{L}_{IM} + \mathcal{L}_{CM}. \quad (5)$$

② **Local Illumination-aware Fusion (LIF)**. After ensuring sufficient learning of mono-modality features, we design a weighted-based fusion method called Local Illumination-aware Fusion (LIF) to explicitly achieve complementary fusion. Different from previous weighted-based methods that only provide one weight for the entire RGB image, the LIF module provides a weight map through the brightness prediction, which can dynamically set different weights for different illumination region features. Specifically, as illustrated in Figure 3, the LIF module is constructed with several convolutions and one activation layer, which can be formulated as:

$$B = ConvBlock(I_V), \quad (6)$$

where B denotes the predicted brightness map. We transform the RGB image to the LAB color space and extract the L channel as the ground-truth \tilde{B} to supervise the brightness prediction, which is formulated as follows:

$$\mathcal{L}_{LI} = \|B, \tilde{B}\|_2. \quad (7)$$

Based on the predicted brightness map B , we design the following weight generation mechanism to adaptively adjust the weights matrix of different modality features, which is calculated as follows:

$$\begin{cases} W_V = \beta \times \min\left(\frac{B - \alpha}{2\alpha}, \frac{1}{2}\right) + \frac{1}{2}, \\ W_I = 1 - W_V, \end{cases} \quad (8)$$

where W_V and W_I represent the weight of the RGB and infrared modalities, respectively. The hyperparameter α is the threshold that determines the importance of RGB features, and β is the amplitude of the W_V . Thus, the boundary of W_V and W_I are:

$$\frac{1 - \beta}{2} \leq W_V, W_I \leq \frac{1 + \beta}{2}. \quad (9)$$

Finally, the LIF module uses the weight maps W_V and W_I to perform element-wise weighted fusion for the multi-modal features. Therefore, the final fused feature can be represented as:

$$f_F^i = \mathcal{F}(f_V, f_I) = W_V^i \odot f_V^i + W_I^i \odot f_I^i. \quad (10)$$

③ **Training and Inference**. For the training stage, we utilize \mathcal{L}_{M^2D} and \mathcal{L}_{LI} to optimize our M²D-LIF framework. Therefore, for the MMOD task, the overall loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{M^2D} \mathcal{L}_{M^2D} + \lambda_{LI} \mathcal{L}_{LI}, \quad (11)$$

where λ_{M^2D} and λ_{LI} are the hyperparameters that control the balance between each loss.

As for the inference, we only utilize the multi-modality backbone network with our proposed LIF module to perform the multi-modal object detection task.

Table 1. Comparison of detection performance (mAP, in%) and computational cost (Params, FLOPs) between the LIF and other SOTA fusion methods on the FLIR (F), DroneVehicle (D), and LLVIP (L) datasets. The symbol * indicates training with our M²D method. The Best results are highlighted by **bold**.

Fusion Method	Fusion Type	F	D	L	Params (M)	Flops (G)
Naive add	-	43.5	66.9	68.2	36.47	116.7
CMX [41]	Attention	43.8	68.0	67.3	+22.3	+33.7
*CMX [41]		44.1	68.9	69.5		
IWM [9]	Weighted	44.0	68.9	68.8	+1.1	+20.1
*IWM [9]		44.9	70.4	69.3		
LIF (Ours)	Weighted	44.9	68.3	67.9	+0.06	+6.5
*LIF (Ours)		46.1	70.6	70.8		

3.3. Why Using LIF Module for Feature Fusion

Actually, various fusion methods can be utilized with our proposed M²D methods to further improve object detection performance, as shown in Table 1. However, we introduce the LIF module into our framework to cooperate with the M²D method for the following two reasons:

(1) **robustness-and-reasonable**. As analyzed in Section 3.1, the weighted-based method can serve as a competitive fusion way to improve performance. As a weighted-based fusion module, LIF is more robust to the insufficient learning of mono-modality encoders, thus achieving superior performance. Furthermore, unlike current weighted-based methods, the LIF module provides reasonably fine-grained quality assessment for the mono-modality images and explicitly achieves the complementary fusion.

(2) **lightweight-yet-effective**. As shown in Table 1, although attention-based methods can adaptively calculate the required features between modalities, they typically rely on high computational complexity. Unlike other methods, our LIF introduces only an additional 0.1M parameters and 6.4 GFLOPs in computational cost to achieve the superior performance, which is more lightweight and effective.

4. Experiments

4.1. Datasets and Evaluation Metrics

1) **DroneVehicle**: This dataset [27] includes images captured by drones in urban areas under different lighting conditions. It has annotations with oriented bounding boxes for five categories: ‘car’, ‘truck’, ‘bus’, ‘van’, and ‘freight car’. There are 28,439 pairs of RGB and IR images, with 17,990 for training, 1,469 for validation, and 8,980 for testing.

2) **FLIR-aligned**: This dataset [38] contains paired RGB and IR images in day and night scenes. It includes 5,142 aligned RGB-IR pairs, with 4,129 for training and 1,013 for testing, focusing on ‘person’, ‘car’, and ‘bicycle’ cate-

Table 2. Ablation on our M²D-LIF framework with performance measured by mAP (%). The best results are highlighted in **bold**. Note: Params/FLOPs tested on FLIR during reference.

M ² D	LIF	FLIR	DroneVehicle	LLVIP	Params	FLOPs
		43.5	66.9	68.2	36.47M	116.7G
✓		45.1	69.2	70.7		
	✓	45.0	68.5	69.4	36.53M	123.2G
✓	✓	46.1	70.6	70.8		

Table 3. Ablation on different ways of the weight map generation in the LIF module with performance measured by mAP (%). The best results are highlighted in **bold**. Sup. represents supervision.

Fusion weight	FLIR	DroneVehicle	LLVIP
L-channel	44.2	67.4	70.2
w/o L-channel Sup.	45.0	69.8	70.3
w/ L-channel Sup.	46.1	70.6	70.8

gories. The ‘dog’ category was removed due to its rarity.

3) **LLVIP**: This dataset [10] is designed for low-light environments, featuring 15,488 strictly aligned RGB-IR image pairs, mostly in very dark scenes. It is divided into 12,025 pairs for training and 3,463 pairs for testing.

4) **Mean Average Precision (mAP)**: mAP is a standard metric for object detection that evaluates performance based on classification accuracy and Intersection over Union (IoU). Specifically, mAP₅₀ denotes the mAP across all classes at a fixed IoU threshold of 0.50. The mAP is the average of mAP values calculated over a range of IoU thresholds from 0.50 to 0.95 with a step of 0.05.

4.2. Implementation Details

All experiments are conducted on NVIDIA GeForce RTX 4090 GPUs. We employ CSPDarknet53 as the backbone network and modified it to process dual-modal inputs. The network is optimized using SGD with a momentum of 0.937 and a weight decay of 5×10^{-4} . We initialize the learning rate at 1×10^{-2} and gradually decrease it to 1×10^{-4} throughout the training process. In the testing phase, we set the confidence threshold to 0.25 to filter detection results. If there is no special state, the ablation study results on the DroneVehicle dataset are conducted on the validation set.

4.3. Ablation Study

1) **Study on Each Component**: To evaluate the effectiveness of the M²D-LIF framework, we conducted ablation studies on the M²D and LIF modules. As shown in Table 2, we incrementally applied these components to the baseline model to evaluate their individual contributions. Without either of these two modules, the model achieves mAP scores of 44.4%, 66.9%, and 68.2% on the FLIR, DroneVehicle, and LLVIP datasets, respectively. The integration of

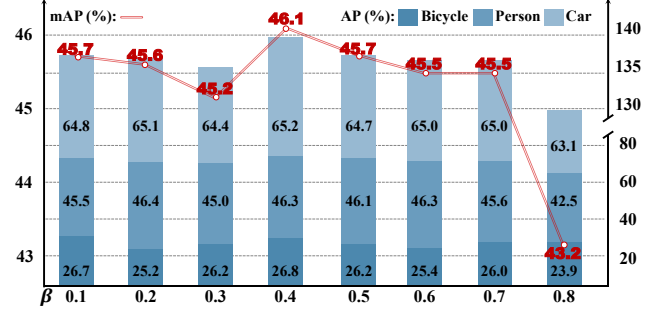


Figure 4. Ablation on the hyper-parameter β in the Equation 8.

the M²D training improves model performance by 1.6% in FLIR, 2.3% in DroneVehicle, and 2.5% in LLVIP and LIF fusion module improves by 1.5%, 1.6%, and 1.2%, respectively. The best results are achieved when combining both the M²D and LIF modules, yielding significant mAP increases of 2.6%, 3.7%, and 2.6% over the baseline while requiring only a marginal 0.1% parameter increase and 5.6% additional computational overhead. These improvements demonstrate the significant impact of our complete M²D-LIF framework.

2) **Study on the LIF**: To demonstrate that our proposed LIF module surpasses simply sensing illumination, we conducted a comparative experiment against the method directly using the normalized L-Channel from LAB color space as fusion weight. As shown in Table 3, our LIF module outperforms this approach by 1.9%, 3.2%, and 0.6% on the FLIR, DroneVehicle, and LLVIP datasets, respectively. Simply using the normalized L-Channel as the fusion weight cannot enable dynamic learning of the fusion weight, which is not conducive to the complementary fusion of multiple modalities. In the FLIR dataset, it even underperforms compared to the baseline method. These results further confirm that our LIF module can not only utilize the illumination context but also effectively leverage the complementary relationship between different modalities.

3) **Study on the Hyper-parameter β** : To investigate the impact of the β value in the Equation 8 on M²D-LIF framework performance, we conducted ablation studies across a pre-defined set of β values. As shown in Figure 4, The bar chart represents the AP of different classes, and the line chart represents the mAP. When beta takes the value of 0.4, our method achieves the best result. If the value of beta exceeds 0.8, it may lead to the modality imbalance, thus causing a serious decline in performance. More ablation studies are detailed in the supplementary material.

4.4. Visualization

1) **Visualization of the Fusion Degradation phenomenon**. To demonstrate that our method can effectively mitigate the Fusion Degradation phenomenon, as illustrated in Figure 5, we visualized the detection results from the mono-modal

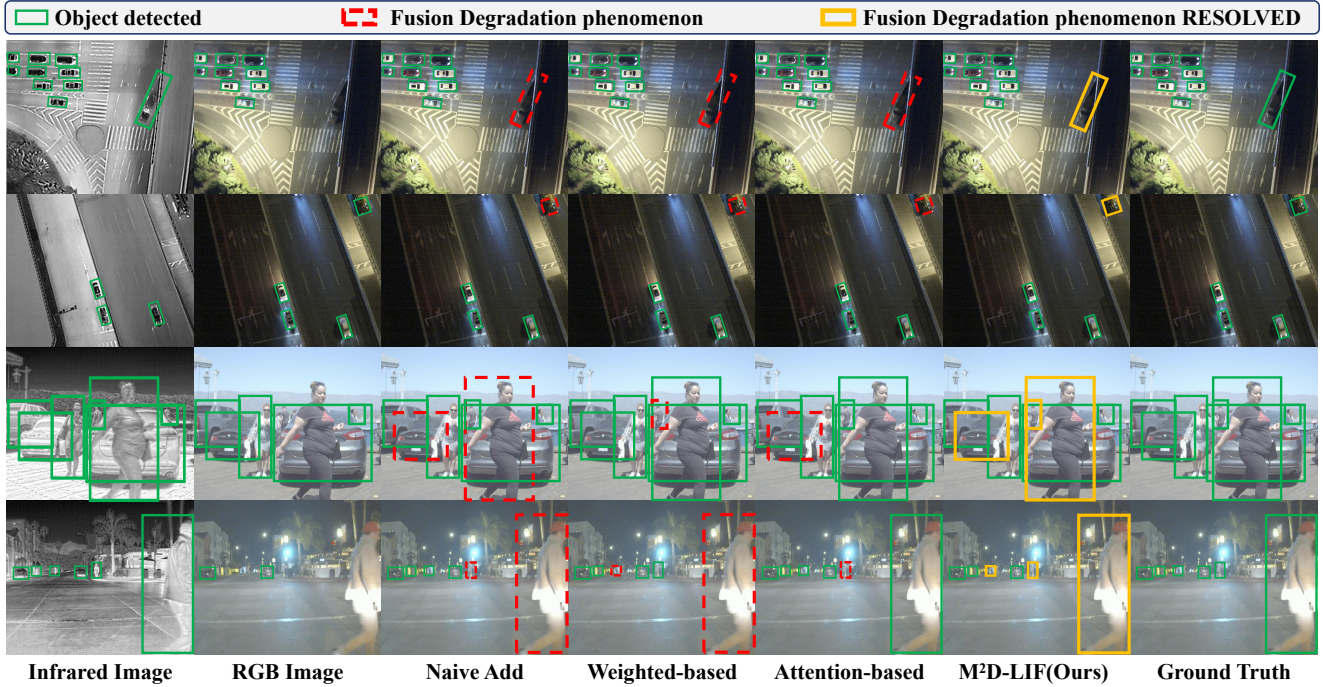


Figure 5. Visualization of the Fusion Degradation phenomenon across the DroneVehicle (first and second rows) and FLIR (third and fourth rows) datasets. Columns from left to right represent infrared modality, RGB modality, naive add method, weighted-based method, attention-based method, our M²D-LIF, and the ground truth. The **Green boxes** indicate correctly detected objects, while the **red dashed boxes** represent the Fusion Degradation phenomenon, where objects are detected by mono-modal models but missed by the multi-modal object detection model. The **Orange boxes** emphasize how our proposed method effectively resolves this issue. Zoom in for details.

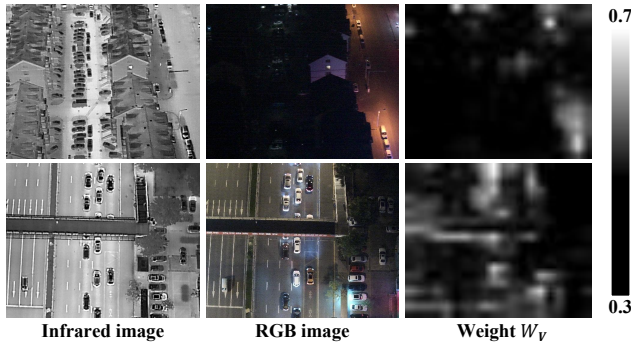


Figure 6. Visualization of the Weight Map W_V in Equation 8. From left to right are the infrared image, the RGB image, the weight map, and the grayscale color corresponding to the value.

method, naive addition (Halfway Fusion [18]), weighted-based (IWM [9]), attention-based (CMX [41]) method and our proposed M²D-LIF framework on the DroneVehicle (first and second rows) and FLIR dataset (third and fourth rows). These examples highlight the widespread occurrence of the Fusion Degradation phenomenon. In contrast, our proposed M²D-LIF framework effectively addresses this issue, ensuring that objects detectable by mono-modal methods remain detectable when using the multi-modal approach, thereby enhancing detection performance.

2) **Visualization of the LIF Weight Map:** To evaluate whether our LIF module effectively perceives illumination and dynamically assigns different weights to regions with varying lighting conditions, we visualize the weight map W_V in Equation 8. As shown in Figure 6, object regions with high-quality illumination in RGB images are assigned with higher weights, thereby explicitly achieving complementary feature fusion. More visualization results are detailed in the supplementary material.

4.5. Comparison With State-of-the-Art Methods

1) **On DroneVehicle:** We evaluated the detection performance and parameter count of our method against four mono-modal oriented bounding box (OBB) detection methods, as well as eight SOTA multi-modal OBB detection methods on the DroneVehicle test and validation set. Table 4 presents the comparative results. Notably, our M²D-LIF framework achieves the highest mAP₅₀ and mAP of 81.4% and 68.1% on the test set. Additionally, it recorded the highest mAP₅₀ of 85.2% on the validation set. Furthermore, our method attained the highest mAP₅₀ across all five categories on the validation set and three first-place and two second-place rankings on the test set. Our method has not only achieved state-of-the-art (SOTA) performance, but also the parameter count of our M²D-LIF framework is optimal

Table 4. Comparison of the performance measured by mAP₅₀, mAP on the DroneVehicle dataset. The best results are highlighted in **red** and the second-place results are highlighted in **blue**. Noted that it uses the ‘OBB’ detectors.

Method	Modality	DroneVehicle test							DroneVehicle val					Params	
		Car	Tru	Fre	Bus	Van	mAP ₅₀	mAP	Car	Tru	Fre	Bus	Van		mAP ₅₀
RetinaNet [16]	RGB	67.5	28.2	13.7	62.1	19.3	38.1	23.4	78.8	39.9	19.5	67.3	24.9	46.1	36.4M
Faster R-CNN [24]		67.9	38.6	26.3	67.0	23.2	44.6	28.4	79.0	49.0	37.2	77.0	37.0	55.9	41.8M
S ² A Net [5]		88.6	58.7	37.3	85.2	41.4	62.2	36.9	80.0	54.2	42.2	84.9	43.8	61.0	33.3M
YOLOv8m [11]		91.0	55.8	43.1	86.3	43.4	63.9	45.9	91.3	56.3	44.1	87.7	46.7	65.2	26.4M
RetinaNet [16]	IR	79.9	32.8	28.1	67.3	16.4	44.9	27.8	89.3	38.2	40.0	79.0	32.1	55.7	36.4M
Faster R-CNN [24]		88.6	42.5	35.2	77.9	28.5	54.6	31.1	89.4	53.5	48.3	87.0	42.6	64.2	41.8M
S ² A Net [5]		90.0	58.3	42.9	87.5	38.9	63.5	38.7	89.9	54.5	55.8	88.9	48.4	67.5	33.3M
YOLOv8m [11]		96.5	63.3	54.6	90.6	45.6	70.1	53.7	96.4	57.8	57.7	92.7	53.2	71.6	26.4M
TarDAL [19]	RGB+IR	89.5	68.3	56.1	89.4	59.3	72.6	43.3	-	-	-	-	-	-	-
UA-CMDet [27]		87.5	60.7	46.8	87.1	38.0	64.0	40.1	-	-	-	-	-	-	-
GLFNet [12]		90.3	72.7	53.6	88.0	52.6	71.4	42.9	-	-	-	-	-	-	-
TSFADet [35]		89.2	72.0	54.2	88.1	48.8	70.4	-	89.9	67.9	63.7	89.8	54.0	73.1	104.7M
CALNet [7]		90.3	76.2	63.0	89.1	58.5	75.4	-	90.3	73.7	68.7	89.7	59.7	76.4	-
C ² Former [34]		90.0	72.1	57.6	88.7	55.4	72.8	42.8	90.2	68.3	64.4	89.8	58.5	74.2	100.8M
CAGTDet [37]		-	-	-	-	-	-	-	90.8	69.7	66.3	90.5	55.6	74.6	140.3M
OAFA [3]		90.3	76.8	73.3	90.3	66.0	79.4	-	-	-	-	-	-	-	-
M ² D-LIF (Ours)		97.8	81.0	67.9	96.0	64.6	81.4	68.1	98.1	81.6	76.5	96.4	69.7	84.5	37.1M

Table 5. Comparison of the performance measured by mAP (%) on the FLIR and LLVIP datasets. The best results are highlighted in **red** and the second-place are highlighted in **blue**.

Method	Modality	FLIR	LLVIP	Params.
RetinaNet [16]	RGB	21.9	42.8	35.9M
Faster R-CNN [24]		28.9	45.1	41.3M
DDQ-DETR [45]		30.9	46.7	-
YOLOv8m [11]		27.8	51.7	25.9M
RetinaNet [16]	IR	31.5	55.1	35.9M
Faster R-CNN [24]		37.6	54.5	41.3M
DDQ-DETR [45]		37.1	58.6	-
YOLOv8m [11]		36.5	58.9	25.9M
Halfway Fus. [18]	RGB+IR	35.8	55.1	-
GAFF [39]		37.3	55.8	31.4M
CFT [23]		40.2	63.9	206.0M
CSAA [2]		41.3	59.2	-
ICAFusion [25]		41.4	64.3	120.2M
RSDet [46]		43.8	61.3	386.0M
UniRGB-IR [36]		44.3	63.2	147.0M
M ² D-LIF (Ours)		46.1	70.8	36.5M

with 37.1M. It has only increased by 0.1M compared to the baseline model in the test phase, which demonstrates its superiority over other multi-modal methods.

2) **On FLIR and LLVIP:** Furthermore, we conducted comprehensive comparisons of our method against state-of-the-art object detection methods on the FLIR and LLVIP

datasets, involving four mono-modal and seven multi-modal methods, as presented in Table 5. Our M²D-LIF framework achieves 46.1% and 70.8% of mAP on the FLIR and LLVIP datasets, respectively, while only employing 36.5M parameters, the second lowest among multi-modal methods. The above results underscore the effectiveness of our M²D-LIF framework in handling diverse datasets. The superior performance on all three datasets highlights the adaptability of the proposed framework. Our method not only competes with but often surpasses existing state-of-the-art multi-modal methods while achieving higher computational efficiency with a lower parameters.

5. Conclusions

In this paper, we first observed the fusion degradation phenomenon. Through linear probing evaluation, we further identified mono-modality insufficient learning as its underlying cause of this phenomenon. To address this issue, we rethought the multi-modal object detection task from the mono-modality learning perspective and constructed an end-to-end M²D-LIF framework. The Mono-Modality Distillation (M²D) method was designed to enhance the feature extraction capability of the multi-modal encoders. The Local Illumination-aware Fusion (LIF) module was designed to dynamically set different weights for the regions with different illumination conditions. Extensive experimental results demonstrated that M²D-LIF achieved superior performance. We believe that our framework holds potential for expansion to other multi-modal tasks.

Acknowledgement

This work was supported by the Fundamental Research Funds for the Central Universities.

References

- [1] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *Advances in Neural Information Processing Systems*, 35: 15394–15406, 2022. 3
- [2] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 403–411, 2023. 2, 8
- [3] Chen Chen, Jiahao Qi, Xingyue Liu, Kangcheng Bin, Ruigang Fu, Xikun Hu, and Ping Zhong. Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26836–26845, 2024. 8
- [4] 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
- [5] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE transactions on geoscience and remote sensing*, 60:1–11, 2021. 8
- [6] Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang. Mapdistill: Boosting efficient camera-based hd map construction via camera-lidar fusion model distillation. *arXiv preprint arXiv:2407.11682*, 2024. 3
- [7] Xiao He, Chang Tang, Xin Zou, and Wei Zhang. Multispectral object detection via cross-modal conflict-aware learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1465–1474, 2023. 2, 8
- [8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2
- [9] Ke Hu, Yudong He, Yuan Li, Jiayu Zhao, Song Chen, and Yi Kang. Ei 2 det: Edge-guided illumination-aware interactive learning for visible-infrared object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 3, 5, 7
- [10] 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, 6
- [11] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 3, 8
- [12] Xudong Kang, Hui Yin, and Puhong Duan. Global–local feature fusion network for visible–infrared vehicle detection. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024. 8
- [13] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference (BMVC)*, 2018. 1
- [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] Ke Li, Di Wang, Zhangyuan Hu, Shaofeng Li, Weiping Ni, Lin Zhao, and Quan Wang. Fd2-net: Frequency-driven feature decomposition network for infrared-visible object detection. *arXiv preprint arXiv:2412.09258*, 2024. 1
- [16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 8
- [17] J Liu, S Zhang, S Wang, and DN Metaxas. Multispectral deep neural networks for pedestrian detection. *arxiv* 2016. *arXiv preprint arXiv:1611.02644*, 2016. 1
- [18] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016. 3, 7, 8
- [19] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5811, 2022. 8
- [20] Tianshan Liu, Kin-Man Lam, Rui Zhao, and Guoping Qiu. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 315–329, 2021. 3
- [21] Michael Person, Mathew Jensen, Anthony O Smith, and Hector Gutierrez. Multimodal fusion object detection system for autonomous vehicles. *Journal of Dynamic Systems, Measurement, and Control*, 141(7):071017, 2019. 1
- [22] Fang Qingyun and Wang Zhaokui. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 130:108786, 2022. 2
- [23] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021. 8
- [24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(06):1137–1149, 2017. 8
- [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*, 145:109913, 2024. 2, 8
- [26] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5311–5320, 2021. 2
- [27] 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. 5, 8
- [28] Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. Crosskd: Cross-head knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16520–16530, 2024. 3
- [29] Alexander Wolpert, Michael Teutsch, M Saquib Sarfraz, and Rainer Stiefelhagen. Anchor-free small-scale multispectral pedestrian detection. *arXiv preprint arXiv:2008.08418*, 2020. 1
- [30] Chunlong Xia, Xinliang Wang, Feng Lv, Xin Hao, and Yifeng Shi. Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5493–5502, 2024. 1
- [31] Ruoyu Xu, Zhiyu Xiang, Chenwei Zhang, Hanzhi Zhong, Xijun Zhao, Ruina Dang, Peng Xu, Tianyu Pu, and Eryun Liu. Sckd: Semi-supervised cross-modality knowledge distillation for 4d radar object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 3
- [32] Chenhongyi Yang, Mateusz Ochal, Amos Storkey, and Elliot J Crowley. Prediction-guided distillation for dense object detection. In *European Conference on Computer Vision*, pages 123–138. Springer, 2022. 2
- [33] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International conference on machine learning*, pages 11863–11874. PMLR, 2021. 4
- [34] Maoxun Yuan and Xingxing Wei. C 2 former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 8
- [35] Maoxun Yuan, Yinyan Wang, and Xingxing Wei. Translation, scale and rotation: cross-modal alignment meets rgb-infrared vehicle detection. In *European Conference on Computer Vision*, pages 509–525. Springer, 2022. 2, 8
- [36] Maoxun Yuan, Bo Cui, Tianyi Zhao, and Xingxing Wei. Unirgb-ir: A unified framework for visible-infrared downstream tasks via adapter tuning. *arXiv preprint arXiv:2404.17360*, 2024. 8
- [37] Maoxun Yuan, Xiaorong Shi, Nan Wang, Yinyan Wang, and Xingxing Wei. Improving rgb-infrared object detection with cascade alignment-guided transformer. *Information Fusion*, page 102246, 2024. 2, 8
- [38] 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. 5
- [39] 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. 8
- [40] Han Zhang, Yunchao Gu, Xinliang Wang, Junjun Pan, and Minghui Wang. Lane detection transformer based on multi-frame horizontal and vertical attention and visual transformer module. In *European Conference on Computer Vision*, pages 1–16. Springer, 2022. 1
- [41] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiqing Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12): 14679–14694, 2023. 2, 3, 5, 7
- [42] Jiaqing Zhang, Mingxiang Cao, Weiying Xie, Jie Lei, Wenbo Huang, Yunsong Li, Xue Yang, et al. E2e-mfd: Towards end-to-end synchronous multimodal fusion detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1
- [43] 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. 2
- [44] Lu Zhang, Zhiyong Liu, Xiangyu Zhu, Zhan Song, Xu Yang, Zhen Lei, and Hong Qiao. Weakly aligned feature fusion for multimodal object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2
- [45] Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7329–7338, 2023. 8
- [46] Tianyi Zhao, Maoxun Yuan, Feng Jiang, Nan Wang, and Xingxing Wei. Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion. *arXiv preprint arXiv:2401.10731*, 2024. 8
- [47] Du Zhixing, Rui Zhang, Ming Chang, Shaoli Liu, Tianshi Chen, Yunji Chen, et al. Distilling object detectors with feature richness. *Advances in Neural Information Processing Systems*, 34:5213–5224, 2021. 2