M2S-KD: A Multi-to-Single Modal Distillation Framework for Visible-Infrared Object Detection

Abstract. While Visible-Infrared (RGB-IR) object detection achieves remarkable performance by fusing complementary information, its deployment is hindered by registration requirements and the high computational cost of multi-stream architectures. To overcome these limitations, we propose a novel **M**ulti-**t**o-**S**ingle **M**odal **K**nowledge **D**istillation (M2S-KD) framework. This framework leverages a powerful multi-modal teacher to guide a cost-effective single-modal student model, transferring rich cross-modal knowledge without incurring additional inference overhead. We systematically investigate diverse distillation strategies, integrating channel-wise saliency alignment and structural alignment at the feature level with cross-head prediction alignment at the logit level. Through extensive experiments across diverse distillation configurations on two RGB-IR benchmark datasets, DroneVehicle and LLVIP, we reveal that the core of cross-modal distillation lies in providing attention guidance, enabling the student to focus on discriminative regions. M2S-KD consistently and significantly boosts the performance of single-modal models across all settings, without introducing any extra inference cost.

Keywords: RGB-IR object detection· Knowledge distillation· Hierarchical knowledge transfer.

1 Introduction

Object detection in adverse conditions such as low illumination and complex backgrounds is crucial for autonomous systems like drones and surveillance platforms [9]. To achieve robustness, the field has increasingly turned to multi-modal object detection (MMOD), particularly combining visible (RGB) and infrared (IR) images [20,29]. This pairing is natural due to their complementary characteristics: the RGB modal provides rich textural and semantic detail in well-lit conditions, while the IR modal mainly captures thermal radiation, offering reliable object outlines in darkness or through obscurants [27].

Existing methods exploit this complementarity through intricate feature fusion within multi-stream architectures. Models such as OAFA [4], CAGTDET [32], and C2Former [33] are designed to interactively combine cross-modal features, extracting modal-specific information, which has led to superior performance compared to single-modal detectors [36,4,1]. However, this performance gain from RGB-IR object detection comes with costs that hinder real-world deployment. First, multi-stream architectures substantially increase computational demands, requiring approximately 2× more parameters (Params) and 1.5× floating-point operations (FLOPs) than single-modal baselines [36,1,33]. Second, most

fusion paradigms critically depend on registered image pairs [30,33]. In practice, inherent sensor differences create a persistent modal gap [13]. Imperfect registered multi-modal data leads to misaligned feature fusion, ultimately degrading rather than enhancing detection accuracy [4,21].

To address above bottlenecks, we investigate a question: Can a single-modal network match the performance of a multi-modal network without requiring multi input? In response, we introduce M2S-KD, a novel Multi-to-Single Modal Knowledge Distillation framework. To the best of our knowledge, we are the first to propose the method of transferring knowledge from multi-modal to single-modal in the field of RGB-IR object detection. M2S-KD leverages a state-of-the-art (SOTA) MMOD model as a teacher to distill its rich cross-modal knowledge into a student network that operates on a single modal (IR-only). This enables the student network to learn rich cross-modal knowledge from the teacher network.

Our framework incorporates three key distillation strategies that across both feature and logit levels. At the feature level, we employ **Channel-Wise Knowledge Distillation (CWD)** [22] to guide channel-wise saliency, and **Pearson Correlation Coefficient Knowledge Distillation (PKD)** [2] to align inter-feature structural correlations. At the logit level, we adopt **Cross-Head Knowledge Distillation (CrossKD)** [24] to mimic teacher's final prediction. Extensive experiments on two challenging RGB-IR detection benchmarks show that M2S-KD significantly boosts the performance of single-modal baselines. Remarkably, the distilled student model achieves competitive accuracy with multi-modal teacher, validating the feasibility of cross-modal knowledge distillation for a single-modal detection. Our core contributions are summarized as follows:

- We propose M2S-KD, the first general knowledge distillation framework tailored for RGB-IR object detection. It effectively bridges the performance gap between single-modal and multi-modal detectors while maintaining inference efficiency.
- We construct a hierarchical knowledge transfer mechanism that systematically integrates saliency, structural, and prediction alignment. Our extensive experiments reveal that saliency alignment is particularly robust for cross-modal distillation, offering key insights for future work.
- We conduct comprehensive experiments and ablations on two RGB-IR benchmarks, DroneVehicle [23] and LLVIP [9]. The results demonstrates the efficacy of our framework, yielding significant performance gains of **3.7%** mAP and **2.2%** mAP over the single-modal baseline, respectively. Moreover, the distilled student model outperforms most prior multi-modal methods.

2 Related Works

2.1 RGB-IR Object Detection

Multi-stream architectures, typically comprising two or more branches, have become the dominant paradigm in RGB-IR object detection, with numerous studies

demonstrating their effectiveness. Early works, such as Li et al. [12], incorporated an illumination-aware weighting mechanism within a two-stream network to adaptively fuse and select multispectral features. Addressing the challenge of spatial misalignment, Chen et al. [4] proposed OAFA, which maps RGB and IR images into a shared feature space. It utilizes a bias-correction branch with deformable convolutions to adaptively predict feature offsets and rectify spatial deviations. Similarly, Shen et al. [21] proposed ICAFusion, leveraging a two-stream cross-attention network to capture cross-modal complementary information and implicitly correct displacements. Wang et al. [26] introduced the MGMF module, marking the first application of the Mamba architecture in RGB-IR object detection to extract discriminative representations with linear complexity.

Despite these advancements, multi-stream architectures inevitably incur significant computational overhead and rely heavily on registered data. In contrast, our M2S-KD framework extracts cross-modal knowledge from a teacher model, enabling a single-stream network to achieve detection performance comparable to multi-stream models without the burden of registration or extra computation.

2.2 Knowledge Distillation

Knowledge Distillation (KD) [8] facilitates knowledge transfer from a high-capacity teacher to a compact student, reducing model complexity while maintaining performance. Existing distillation methods can be broadly categorized into feature-based and logit-based distillation [17]. Feature-based Distillation focuses on aligning intermediate representations. FitNet [19] pioneered this approach using hint layers to match activations, though its strict element-wise alignment often leads to overfitting. ReviewKD [5] utilizes cross-layer feature aggregation to guide student learning. FGD [28] employs ground-truth masks to balance knowledge transfer between foreground and background. MSCD [25] enables scale-adaptive representations through multi-branch cross-scale feature distillation. More relevant to our work are methods that relax strict numerical constraints: CWD [22] focuses on channel-wise saliency by normalizing activation maps to minimize the divergence of soft probability distributions; PKD [2] aligns features via Pearson Correlation Coefficients, relaxing constraints on feature magnitude to focus on spatial structural relationships.

Logit-based Distillation operates on the final prediction layers. MLKD [10] proposes a multi-level framework to distill instance, batch, and class information. CrossKD [24] extracts intermediate features from the student's detection head and feeds them into the teacher's reused head. KD has also been explored in cross-modal scenarios. For instance, Zhao et al. [36] proposed M2D-LIF, which addresses feature degradation in RGB-IR fusion by using two single-modal teachers to guide a multi-modal student. Inspired by these works, we explore the inverse paradigm: transferring knowledge from a multi-modal teacher to a single-modal student.

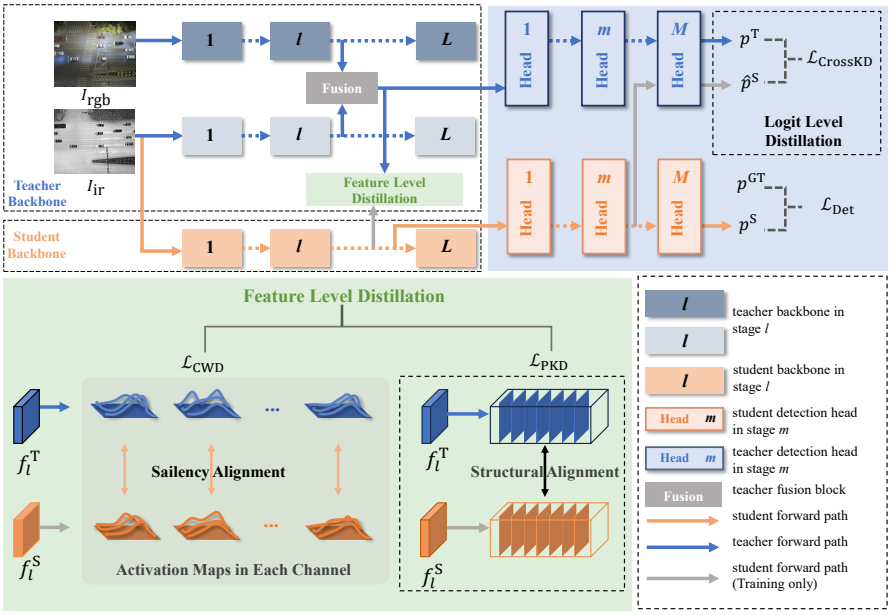

Fig. 1. Overall framework of the proposed M2S-KD. Given a teacher-student pair, M2S-KD computes feature-level distillation losses at stage l of the backbone. For logit-level distillation, CrossKD calculates the distillation loss between cross-head predictions and the original teacher outputs. Note that since distillation involves multiple backbone and head stages, for simplicity we only illustrate the computation at stage l in the diagram. Dashed arrows indicate omitted backbone or head stages at other levels, and the FPN process of the detection network is also simplified for clarity.

3 Method

In this section, we present our proposed M2S-KD framework. As depicted in Fig. 1, the overall architecture consists of a multi-modal teacher and a single-modal student. The teacher network accepts both RGB and IR images, whereas the student network is designed to take only IR input. To effectively transfer cross-modal knowledge, our framework supports distillation at both feature level and logit level. During training, the student's backbone features at multiple scales are aligned with the teacher's corresponding fused features via two strategies: CWD and PKD. At the logit level, intermediate features from the student's detection head are forwarded into the reused teacher detection head, and the resulting outputs are supervised via a CrossKD loss. During inference, the teacher model and the distillation loss are removed, leaving only the single-modal student to perform detection using IR images as input.

We adopt the SOTA M2D-LIF [36] as our teacher model $\mathcal{T}$, which consists of two YOLOv8m [11] networks with dedicated RGB and IR branches respectively, integrated through the teacher model fusion block. For the student model S, we employ a YOLOv8m [11] network that only using IR as the input.

Given the paired RGB image I_{rgb} and IR image I_{ir}, the input images are processed by the teacher and student backbones. Both networks comprise a hierarchical structure of convolutional stages. Specifically, the multi-modal teacher backbone can be formally represented as $\{\mathrm{B}_l^{\mathrm{T,rgb}}\}_{l=1}^{L}$ for the RGB branch and $\{\mathrm{B}_l^{\mathrm{T,ir}}\}_{l=1}^{L}$ for the IR branch, while the student backbone is denoted as $\{\mathrm{B}_l^{\mathrm{S}}\}_{l=1}^{L}$, where l indexes the stage depth (with larger l indicating deeper semantic levels). For simplicity, we assume there are L total stages in the backbone (in YOLOv8m, $L = 5$, though the feature pyramid network primarily utilizes stages 3–5 for detection). Correspondingly, we denote the output features as $f_l^{\mathrm{T,rgb}}$, $f_l^{\mathrm{T,ir}}$, and f_l^{S} for the RGB branch, IR branch, and student backbone respectively. The teacher's fusion module at each stage l is represented as $\mathrm{F}_l^{\mathrm{T}}$, which integrates the multi-modal features to produce the fused representation:

$$f_l^{\mathrm{T}} = \mathrm{F}_l^{\mathrm{T}}\big(f_l^{\mathrm{T,rgb}},\, f_l^{\mathrm{T,ir}}\big) \tag{1}$$

3.1 Feature-Level Alignment

Strict feature alignment often leads to overfitting [2,5,22], resulting in suboptimal performance. We adopt a soft alignment strategy to facilitate more robust cross-modal knowledge transfer. This is achieved through two approaches: **Saliency Alignment** and **Structural Alignment**.

Saliency Alignment. To guide the student to focus on semantically salient foreground regions irrespective of textural differences, we employ CWD [22]. CWD relaxes strict numerical matching by converting activation maps into spatial probability distributions. For the c-th channel at level l, we compute a spatial probability map $\phi(\cdot)$ via a channel-wise Softmax normalization:

$$\phi(f_{l,c,i}) = \frac{\exp(f_{l,c,i}/\tau)}{\sum_{j=1}^{H_l W_l} \exp(f_{l,c,j}/\tau)} \tag{2}$$

where $i \in \{1, \ldots, H_l W_l\}$ indexes the flattened spatial locations, and τ is a temperature hyperparameter. The distillation loss is defined as the Kullback-Leibler (KL) divergence between the teacher's and student's channel-wise distributions, summed over all levels and channels:

$$\mathcal{L}_{\text{CWD}} = \sum_{l=1}^{L} \frac{\tau^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{H_l W_l} \phi(f_{l,c,i}^{\mathrm{T}}) \log\left(\frac{\phi(f_{l,c,i}^{\mathrm{T}})}{\phi(f_{l,c,i}^{\mathrm{S}})}\right) \tag{3}$$

This compels the student to replicate the teacher's spatial attention, effectively filtering out modal-specific background noise.

Structural Alignment To enable the student to capture the relational topology of objects, we employ PKD based on Pearson Correlation. PKD aligns the relative relationships within features rather than their absolute values. For each feature pyramid level l and channel c, we compute the Pearson Correlation Coefficient between student and teacher features as:

$$r_l = \sum_{c=1}^{C} \frac{\sum_{k=1}^{BH_l W_l} (f_{l,c,k}^{\mathrm{S}} - \mu_{l,c}^{\mathrm{S}})(f_{l,c,k}^{\mathrm{T}} - \mu_{l,c}^{\mathrm{T}})}{\sqrt{\sum_{k=1}^{BH_l W_l} (f_{l,c,k}^{\mathrm{S}} - \mu_{l,c}^{\mathrm{S}})^2} \sqrt{\sum_{k=1}^{BH_l W_l} (f_{l,c,k}^{\mathrm{T}} - \mu_{l,c}^{\mathrm{T}})^2}} \tag{4}$$

where $\mu_{l,c}^{\mathrm{S}}$ and $\mu_{l,c}^{\mathrm{T}}$ are the channel-wise means computed across the entire batch and spatial dimensions for level l and channel c. The distillation loss minimizes the correlation distance:

$$\mathcal{L}_{\mathrm{PKD}} = \sum_{l=1}^{L} (1 - r_l) \tag{5}$$

This formulation focuses on preserving structural relationships between features while being invariant to their absolute magnitudes, which is crucial for cross-modal knowledge transfer.

3.2 Logit-Level Alignment

For the detection head, both the teacher and student networks consist of convolutional blocks organized into different stages. The detection heads of the teacher and student network are denoted as $\mathrm{H}_m^{\mathrm{T}}$ and $\mathrm{H}_m^{\mathrm{S}}$ respectively, where m indexes the corresponding stage of convolutional blocks. For simplicity, we suppose there are totally M stages in the detection head (e.g., M= 3 in YOLOv8m). The output feature of $\mathrm{H}_m^{\mathrm{S}}$ is represented as f_m^{S}. p^{T} denotes the teacher network's output prediction. By feeding f_m^{S} into the remaining layers of the teacher's detection head, we obtain the cross prediction result $\hat{p}^{\mathrm{S}}$:

$$\hat{p}^{\mathrm{S}} = \mathrm{H}_{m+1:M}^{\mathrm{T}}(f_m^{\mathrm{S}}) \tag{6}$$

The total CrossKD loss can be expressed as follows:

$$\mathcal{L}_{\mathrm{CrossKD}} = \mathcal{L}_{\mathrm{CrossKD}}^{\mathrm{cls}}(\hat{p}_{\mathrm{cls}}^{\mathrm{S}}, p_{\mathrm{cls}}^{\mathrm{T}}) + \mathcal{L}_{\mathrm{CrossKD}}^{\mathrm{reg}}(\hat{p}_{\mathrm{reg}}^{\mathrm{S}}, p_{\mathrm{reg}}^{\mathrm{T}}) + \mathbb{I} \cdot \mathcal{L}_{\mathrm{CrossKD}}^{\mathrm{ang}}(\hat{p}_{\mathrm{ang}}^{\mathrm{S}}, p_{\mathrm{ang}}^{\mathrm{T}}) \tag{7}$$

where $\mathcal{L}_{\mathrm{CrossKD}}^{\mathrm{cls}}$ and $\mathcal{L}_{\mathrm{CrossKD}}^{\mathrm{reg}}$ represent the classification and regression distillation losses. They measure the discrepancy between the class predictions $\hat{p}_{\mathrm{cls}}^{\mathrm{S}}$, $p_{\mathrm{cls}}^{\mathrm{T}}$ and the regulation predictions $\hat{p}_{\mathrm{reg}}^{\mathrm{S}}$, $p_{\mathrm{reg}}^{\mathrm{T}}$, respectively. The additional angular loss term $\mathcal{L}_{\mathrm{CrossKD}}^{\mathrm{ang}}$ captures the alignment of orientation predictions between the student and teacher networks. $\mathbb{I}$ is an indicator function that equals 1 for oriented object detection tasks and 0 for horizontal object detection tasks. This mechanism forces the student's features to produce outputs compatible with the teacher's refined prediction boundaries.

3.3 Overall Optimization Objective

The total training objective for the single-modal student integrates the original detection loss with the distillation losses:

$$\mathcal{L}_{\mathrm{Total}} = \mathcal{L}_{\mathrm{Det}} + \lambda_{\mathrm{CWD}}\mathcal{L}_{\mathrm{CWD}} + \lambda_{\mathrm{PKD}}\mathcal{L}_{\mathrm{PKD}} + \lambda_{\mathrm{Cross}}\mathcal{L}_{\mathrm{Cross}}, \tag{8}$$

where $\mathcal{L}_{\mathrm{Det}}$ denotes the loss of the student predictions p^{S} and the ground truth p^{GT}. The hyperparameters $\lambda_{\mathrm{CWD}}, \lambda_{\mathrm{PKD}}, \lambda_{\mathrm{Cross}}$ balance the contributions of the respective distillation modules. Through extensive experiments in the ablation study, we validated the effectiveness of different distillation combinations, with all corresponding weights set to 1.0.

4 Experiments

4.1 Datasets and Evaluation Metrics

M2S-KD is evaluated on two large-scale RGB-IR object detection benchmarks, DroneVehicle [23] and LLVIP [9], using the mean Average Precision (mAP) and mAP_{50} as the evaluation metrics.

DroneVehicle. We evaluate our method on the DroneVehicle dataset [23], a large-scale multimodal benchmark for oriented vehicle detection in unmanned aerial vehicle imagery. It contains 28,439 registered RGB-IR image pairs with 953,087 annotated vehicles across diverse altitudes, viewing angles, and illumination conditions. Oriented Bounding Boxes (OBB) are provided for five categories: car, truck, bus, van, and freight car.

LLVIP. To validate the generalization of our framework, we also employ the LLVIP dataset [9], designed for pedestrian detection in low-light conditions. It consists of 15,488 strictly registered RGB-IR pairs, predominantly captured in dark scenes, with Horizontal Bounding Box (HBB) annotations. The dataset is split into 12,025 training and 3,463 testing pairs.

Evaluation Metrics. We adopt the standard mAP_{50} and mAP for evaluation. Specifically, mAP_{50} denotes the average precision at an Intersection over Union (IoU) threshold of 0.50, while mAP represents the average across IoU thresholds from 0.50 to 0.95 with a step of 0.05.

Table 1. Models results on the DroneVehicle using the mAP_{50} and mAP metrics for detection accuracy. Params, FLOPs and FPS are reported to assess computational efficiency. All models are configured for oriented object detection. The best results are highlighted in **red** and the second-place are highlighted in **blue**.

Method	Modal	test							val	Params.	Flops	FPS
		Car	Tru	Fre	Bus	Van	mAP_{50}	mAP	mAP_{50}			
TarDAL [16]	RGB+IR	89.5	68.3	56.1	89.4	59.3	72.6	43.3	-	-	-	-
UA-CMDet [23]	RGB+IR	87.5	60.7	46.8	87.1	38.0	64.0	40.1	-	-	-	-
TSFADet [30]	RGB+IR	89.2	72.0	54.2	88.1	48.8	70.4	-	73.1	104.7M	109.8G	18.6
CALNet [7]	RGB+IR	90.3	76.2	63.0	89.1	58.5	75.4	-	76.4	-	-	-
C^2Former [33]	RGB+IR	90.0	72.1	57.6	88.7	55.4	72.8	42.8	74.2	100.8M	**89.9G**	-
CAGTDet [32]	RGB+IR	-	-	-	-	-	-	-	74.6	-	-	-
OAFA [4]	RGB+IR	90.3	76.8	**73.3**	90.3	**66.0**	79.4	-	-	-	-	33.1
DDCINET [1]	RGB+IR	91.0	**78.9**	66.1	90.7	**65.5**	78.4	-	-	121.7M	124.8G	17.2
M2D-LIF (teacher) [36]	RGB+IR	**97.8**	**81.0**	**67.9**	**96.0**	64.6	**81.4**	**68.1**	**84.5**	**37.0M**	124.6G	**73.5**
RetinaNet [14]	RGB	67.5	28.2	13.7	62.1	19.3	38.1	23.4	46.1	36.4M	-	-
Faster RCNN [18]	RGB	67.9	38.6	26.3	67.0	23.2	44.6	28.4	55.9	41.8M	-	-
S^2A Net [6]	RGB	88.6	58.7	37.3	85.2	41.4	62.2	36.9	61.0	33.3M	-	-
YOLOv8m [11]	RGB	95.4	76.1	57.2	94.4	61.0	76.8	62.6	79.2	26.4M	80.8G	105.2
RetinaNet [14]	IR	79.9	32.8	28.1	67.3	16.4	44.9	27.8	55.7	36.4M	-	-
Faster R-CNN [18]	IR	88.6	42.5	35.2	77.9	28.5	54.6	31.1	64.2	41.8M	-	-
S^2A Net [6]	IR	90.0	58.3	42.9	87.5	38.9	63.5	38.7	67.5	33.3M	-	-
YOLOv8m (baseline) [11]	IR	**97.5**	75.1	61.6	93.0	57.7	76.9	63.9	80.1	**26.4M**	**80.8G**	**106.4**
M2S-KD (ours)	IR	**97.2**	**77.4**	**67.2**	**93.5**	**62.1**	**79.5**	**67.6**	**82.3**	**26.4M**	**80.8G**	**105.2**

4.2 Implementation Details

All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU. Our implementation is based on PyTorch 2.1.0 and the Ultralytics framework [11].

We use **CSPDarknet53** as the backbone for both teacher and student models, with the teacher's data pipeline adapted for multi-stream input.

Training Strategy. To ensure a robust teacher, we first pretrain two unimodal YOLOv8m networks (RGB and IR) for 36 epochs. These are used to initialize the M2D-LIF [36] teacher network, which is then trained for 100 epochs. Following the standard protocol in RGB-IR detection, only IR annotations are used for final evaluation. For our **M2S-KD** framework, the pretrained M2D-LIF serves as the frozen teacher to distill knowledge into a single-modal YOLOv8m student (IR-only), which is trained for 100 epochs.

Optimization. We use Stochastic Gradient Descent (SGD) with a momentum of 0.937 and weight decay of 5×10^{-4}. The initial learning rate is set to 1×10^{-2} and decayed linearly to 1×10^{-1}. During inference, we apply a confidence threshold of 0.25 to filter predictions.

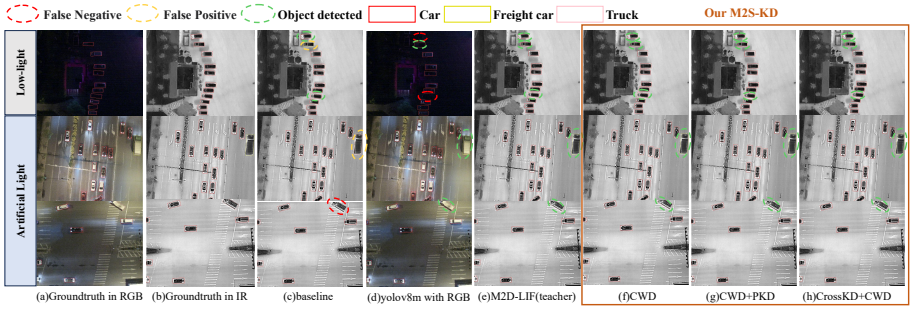

Fig. 2. Illustrative detection results from the DroneVehicle test set. The confidence threshold is set to 0.25. Red, yellow, and pink rectangles denote car, freight car, and truck objects, respectively. Correctly detected objects are marked with green dashed circles, while false negatives and false positives are indicated by red and yellow dashed circles. our M2S-KD is outlined with an orange rectangle.

4.3 Comparison with SOTA Methods

Results on DroneVehicle Dataset We compare our method with the SOTA RGB-IR detector M2D-LIF (Teacher), the baseline YOLOv8m with IR input, and a wide range of existing detectors. Single-modal methods: including one-stage detectors (RetinaNet [14], S2ANet] [6]) and two-stage detectors (Faster R-CNN [18]). Multi-modal fusion methods: UA-CMDet [23], Halfway Fusion [15], TSFADet [30], C2Former [33], DDCINET [1], and OAFA [4]. For a fair comparison, we report results on both validation and test sets using each method's original settings.

Quantitative Analysis. As shown in Table. 1, the YOLOv8m baseline demonstrates strong performance, achieving 76.8% mAP_{50} on RGB and 76.9% on IR, outperforming several early fusion methods. The dual-stream M2D-LIF [36] teacher achieves the highest performance. Remarkably, our M2S-KD attains the second-best overall performance. Without RGB input during inference, it outperforms all comparison methods except the teacher, achieving **3.7%** mAP and

Table 2. Models results on the LLVIP dataset using the mAP metrics for detection accuracy. Params, FLOPs and FPS are reported to assess computational efficiency. All models are configured for OBB detection. The best results are highlighted in **red** and the second-place are highlighted in **blue.**

Method	Modal	mAP	Params	Flops	FPS
Halfway Fus. [15]	RGB+IR	55.1	-	-	-
GAFF [34]		55.8	**31.4M**	-	-
CSAA [3]		59.2	-	-	-
ICAFusion [21]		64.3	120.16M	-	-
RSDet [37]		61.3	386.0M	-	-
UniRGB-IR [31]		63.2	147.0M	-	-
M^{2}D-LIF (teacher)		**70.8**	36.6M	**122.4G**	**62.5**
RetinaNet [14]	RGB	42.8	35.9M	-	-
Faster R-CNN [18]		45.1	41.3M	-	-
DDQ-DETR [35]		46.7	-	-	-
YOLOv8m [11]		65.4	25.9M	78.7G	82.3
RetinaNet [14]	IR	55.1	35.9M	-	-
Faster R-CNN [18]		54.5	41.3M	-	-
DDQ-DETR [35]		58.6	-	-	-
YOLOv8m (baseline) [11]		68.5	**25.9M**	**78.7G**	**89.7**
M^{2}S-KD (Ours)		**70.7**	**25.9M**	**78.7G**	**85.4**

2.6% mAP$_{50}$ improvement over the IR baseline. Class-wise analysis shows significant gains in challenging categories (e.g., Van and Freight-car), while maintaining high accuracy in easier ones (e.g., Car and Bus)

Efficiency Analysis. During inference, our M2S-KD student retains the single stream to the baseline YOLOv8m. As shown in Table. 1, it maintains significantly lower Parameters and FLOPs, and higher FPS than all multimodal methods. Compared to the teacher, M2S-KD reduces parameters by **10M** and FLOPs by **43G**, while increasing speed by **31 FPS**. Crucially, our method relies solely on IR input, eliminating the dependency on registered multimodal data.

Qualitative Analysis. Fig. 2 presents visual detection results from different models and various distillation configurations of M2S-KD. In low-light scenes, YOLOv8m (RGB) fails to detect two cars, while the IR-only baseline produces false positives due to the lack of textural details. In the night scenes with artificial light, YOLOv8m (RGB) demonstrates a certain capability to recognize challenging samples, whereas the baseline suffers from both false negative and false positive. By effectively fusing complementary RGB and IR features, the teacher model delivers superior performance across all scenarios. Notably, our M2S-KD framework yields detection results consistent with those of the teacher, demonstrating that the proposed distillation successfully enables the student to acquire the teacher's advanced cross-modal representations.

Results on LLVIP Dataset Table. 2 presents a comprehensive comparison on LLVIP against four single-modal and seven multi-modal methods. Our M2S-KD framework achieves **70.7%**mAP, merely 0.1% lower than the teacher model. This negligible performance drop comes with substantial efficiency gains: a reduction of **11M** parameters and **43G** FLOPs, and a **22 FPS** speed increase.

Table 3. Ablation study on DroneVehicle dataset. P3, P4, and P5 denote the application of feature-level distillation losses on the output features from the l= 3, 4, and 5 levels of the backbone, respectively. Meanwhile, stage1 and stage2 refer to the configurations in CrossKD where the outputs from the student's detection head at stages m= 1 and m= 2 are fed into the reused teacher detection head. The best result in each group is shown in **bold**, and the strategies selected for visualization are highlighted in **red**.

PKD			**CWD**			**CrossKD**		**mAP**$_{50}$
P3	P4	P5	P3	P4	P5	stage1	stage2	
								76.9
		✓						79.2
	✓	✓						**79.3**
✓	✓	✓						78.8
					✓			79.2
				✓	✓			**79.5**
			✓	✓	✓			79.1
						✓		79.3
							✓	78.9
		✓				✓		79.2
	✓	✓				✓		79.2
✓	✓	✓				✓		**79.3**
		✓					✓	79.1
	✓	✓					✓	78.9
✓	✓	✓					✓	**79.3**
					✓	✓		**79.4**
				✓	✓	✓		**79.4**
			✓	✓	✓	✓		79.3
					✓		✓	79.3
				✓	✓		✓	79.3
			✓	✓	✓		✓	79.3
		✓		✓				79.4
	✓				✓			79.4
		✓	✓	✓				79.4
✓	✓				✓			**79.5**

This result underscores the superior accuracy-efficiency trade-off and strong generalization capability of our framework across diverse datasets.

4.4 Ablation Studies

We conduct ablation studies to verify the contribution of each distillation component within M2S-KD. As summarized in Table 3, we explore various combinations. Even with all loss weights set to 1.0 (no intricate tuning), all distillation combinations improve upon the baseline (minimum gain: +1.9% mAP$_{50}$). Most combinations achieve mAP$_{50}$ over **79.3%**, validating the proposed strategies. Combining all three losses did not yield further gains, likely due to gradient conflicts between different objectives and the information bottleneck of the single-modal IR input. Therefore, we recommend dual-combinations or a single strong module (like CWD) for a simplicity-performance balance.

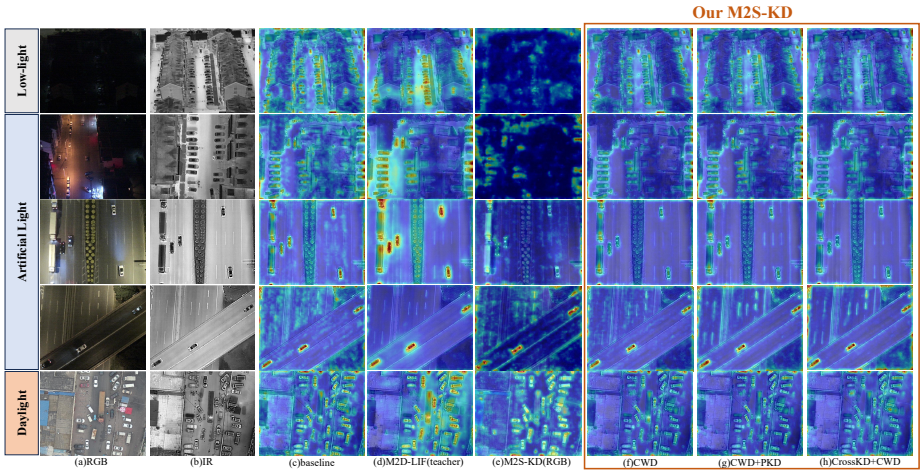

Fig. 3. Visual comparison of heatmaps across methods. The plus sign indicates combinations of distillation strategies. (a)–(b) are original RGB and IR images. Visualizations in (c) and (e)–(h) are based on backbone feature maps, with the teacher model showing fused features at corresponding layers. (e) presents heatmaps from M2S-KD guided single-modal RGB student, while (f)–(h) show those from M2S-KD guided single-modal IR student.

4.5 Visualization: Unveiling Cross-Modal Knowledge Transfer

To elucidate what knowledge is transferred, we visualize gradient-weighted class activation mapping for backbone of the baseline, teacher, and M2S-KD student, as shown at Fig. 3.

Enhanced Foreground Focus. M2S-KD consistently concentrates attention on foreground objects with sharper contours, whereas the baseline exhibits scattered and blurry activations. This pattern closely mirrors the teacher, confirming its role as a soft attention guide.

Learning Distribution over Intensity. While the student learns correct attention locations, its activation intensities are generally lower than the teacher's. This suggests the student distinguishes the foreground-background distinction rather than overfitting to the teacher's absolute feature values.

Modal Suitability Analysis. We also applied distillation to a single-modal RGB student. In the low-light scene shown in Fig. 3(e), RGB inputs contain minimal valid information. Forcing the RGB student to mimic IR-derived features results in attention dispersion and unstable learning. This confirms infrared as the optimal target modal for our distillation framework, given its consistent information content across lighting conditions.

5 Conclusion

In this work, we address two critical limitations of multimodal object detection: depending on registered data and requiring high computational cost. To this

end, we propose **M2S-KD**, a novel Multi-to-Single Modal Knowledge Distillation framework. M2S-KD transfers knowledge from a multi-modal teacher to a single-modal student through a hierarchical knowledge distillation mechanism. We systematically investigate distillation strategies at both feature and logit levels. Extensive experiments on two RGB-IR benchmarks demonstrate that our framework robustly boosts single-modal performance. M2S-KD uses only single IR images as input, avoiding modal gap. Ablation and visualization studies reveal that the core of effective cross-modal transfer lies in providing **attention guidance**, which helps the student refine foreground-background discrimination and hard sample recognition.

References

1. Bao, W., Huang, M., Hu, J., Xiang, X.: Dual-dynamic cross-modal interaction network for multimodal remote sensing object detection. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–13 (2025)
2. Cao, W., Zhang, Y., Gao, J., Cheng, A., Cheng, K., Cheng, J.: Pkd: General distillation framework for object detectors via pearson correlation coefficient. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) NeurIPS 2022. vol. 35, pp. 15394–15406. Curran Associates, Inc., Red Hook (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/631ad9ae3174bf4d6c0f6fdca77335a4-Paper-Conference.pdf
3. Cao, Y., Bin, J., Hamari, J., Blasch, E., Liu, Z.: Multimodal object detection by channel switching and spatial attention. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 403–411. IEEE, Vancouver, BC, Canada (2023)
4. Chen, C., Qi, J., Liu, X., Bin, K., Fu, R., Hu, X., Zhong, P.: Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26826–26835. Publisher: IEEE, Seattle, WA, USA (2024)
5. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5006–5015. IEEE, Nashville, TN, USA (2021)
6. Han, J., Ding, J., Li, J., Xia, G.S.: Align deep features for oriented object detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–11 (2022)
7. He, X., Tang, C., Zou, X., Zhang, W.: Multispectral object detection via cross-modal conflict-aware learning. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1465–1474. Association for Computing Machinery, New York, NY, USA (2023)
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015), https://arxiv.org/abs/1503.02531
9. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 3489–3497. IEEE, Montreal, BC, Canada (2021)
10. Jin, Y., Wang, J., Lin, D.: Multi-level logit distillation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24276–24285. IEEE, Vancouver, BC, Canada (2023)

11. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO, https://github.com/ultralytics/ultralytics
12. Li, C., Song, D., Tong, R., Tang, M.: Illumination-aware faster r-cnn for robust multispectral pedestrian detection. Pattern Recognition **85**, 161–171 (2019)
13. Li, H., Zhao, J., Li, J., Yu, Z., Lu, G.: Feature dynamic alignment and refinement for infrared–visible image fusion: Translation robust fusion. Information Fusion **95**, 26–41 (2023)
14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2999–3007. IEEE, Venice, Italy (2017)
15. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: (2016), http://arxiv.org/abs/1611.02644
16. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5792–5801. IEEE, New Orleans, LA, USA (2022)
17. Miles, R., Yucel, M.K., Manganelli, B., Saà-Garriga, A.: Mobilevos: Real-time video object segmentation contrastive learning meets knowledge distillation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10480–10490. IEEE, Vancouver, BC, Canada (2023)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017)
19. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets (2015), https://arxiv.org/abs/1412.6550
20. Shakhatreh, H., Sawalmeh, A.H., Al-Fuqaha, A., Dou, Z., Almaita, E., Khalil, I., Othman, N.S., Khreishah, A., Guizani, M.: Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges. IEEE Access **7**, 48572–48634 (2019)
21. Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H., Yang, W.: Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognition **145**, 109913 (2024)
22. Shu, C., Liu, Y., Gao, J., Yan, Z., Shen, C.: Channel-wise knowledge distillation for dense prediction. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5291–5300. IEEE, Montreal, QC, Canada (2021)
23. Sun, Y., Cao, B., Zhu, P., Hu, Q.: 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)
24. Wang, J., Chen, Y., Zheng, Z., Li, X., Cheng, M.M., Hou, Q.: Crosskd: Cross-head knowledge distillation for object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16520–16530. IEEE, Seattle, WA, USA (2024)
25. Wang, K., Wang, Z., Li, Z., Teng, X., Li, Y.: Multi-scale cross distillation for object detection in aerial images. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 452–471. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72967-6_25
26. Wang, S., Wang, C., Shi, C., Liu, Y., Lu, M.: Mask-guided mamba fusion for drone-based visible-infrared vehicle detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–12 (2024)

27. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4236–4244. IEEE, Honolulu, HI, USA (2017)
28. Yang, Z., Li, Z., Jiang, X., Gong, Y., Yuan, Z., Zhao, D., Yuan, C.: Focal and global knowledge distillation for detectors. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4633–4642. IEEE, New Orleans, LA, USA (2022)
29. Yao, H., Qin, R., Chen, X.: Unmanned aerial vehicle for remote sensing applications—a review. Remote Sensing **11**(12), 1–22 (2019)
30. Yuan, M., Wang, Y., Wei, X.: Translation, Scale and Rotation: Cross-Modal Alignment Meets RGB-Infrared Vehicle Detection. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. vol. 13671, pp. 509–525. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20077-9_30
31. Yuan, M., Cui, B., Zhao, T., Wang, J., Fu, S., Yang, X., Wei, X.: Unirgb-ir: A unified framework for visible-infrared semantic tasks via adapter tuning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 2409–2418. Association for Computing Machinery, New York, NY, USA (2025)
32. Yuan, M., Shi, X., Wang, N., Wang, Y., Wei, X.: Improving rgb-infrared object detection with cascade alignment-guided transformer. Information Fusion **105**, 1–11 (2024)
33. Yuan, M., Wei, X.: C2former: Calibrated and complementary transformer for rgb-infrared object detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–12 (2024)
34. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Guided attentive feature fusion for multispectral pedestrian detection. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 72–80. IEEE, Waikoloa, HI, USA (2021)
35. Zhang, S., Wang, X., Wang, J., Pang, J., Lyu, C., Zhang, W., Luo, P., Chen, K.: Dense distinct query for end-to-end object detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7329–7338. IEEE, Vancouver, BC, Canada (2023)
36. Zhao, T., Liu, B., Gao, Y., Sun, Y., Yuan, M., Wei, X.: Rethinking multi-modal object detection from the perspective of mono-modality feature learning (2025), https://arxiv.org/abs/2503.11780
37. Zhao, T., Yuan, M., Jiang, F., Wang, N., Wei, X.: Removal then selection: A coarse-to-fine fusion perspective for rgb-infrared object detection (2024), https://arxiv.org/abs/2401.10731