Multimodal Decomposed Distillation with Instance Alignment and Uncertainty Compensation for Thermal Object Detection

Yanfeng Liu
National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
Wuhan, Hubei, China
liuyanfeng99@whu.edu.cn

Lefei Zhang*
National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
Wuhan, Hubei, China
zhanglefei@whu.edu.cn

Abstract

RGB-Thermal images leverage complementary optical and thermal modalities to identify objects. While achieving superior performance, the reliance on multimodal fusion inherently limits inference efficiency and adaptability to harsh RGB-failure environments. In this work, we propose a multimodal decomposed distillation framework to develop robust thermal-only detectors by transferring knowledge from multimodal teachers. Unlike conventional one-to-one distillation, we decouple the tasks of simultaneously mimicking RGB-T teacher representations and preserving thermal-specific student feature integrity into dual branches to avoid intrinsic semantic conflicts. Specifically, we present channel-adaptive prompt learning for cross-modal decomposition and a frequency-guided dynamic module for decomposed knowledge integration. The dual-branch architecture employs asymmetric training objectives to ensure effective cross-modal knowledge transfer while preserving the integrity of thermal information. Furthermore, to exploit finer-grained instance knowledge across both feature and prediction levels, we introduce a customized instance alignment distillation to enhance the local discriminability in feature pyramids, and propose an uncertainty-aware logit distillation to compensate for ambiguous predictions in detection heads. Experiments on three datasets validate the effectiveness of our framework in boosting thermal-based detectors. Code is released at https://github.com/lyf0801/DecomKD.

CCS Concepts

• **Computing methodologies** → **Object detection**.

Keywords

Mutimodal Decomposed Distillation, Thermal Object Detection

ACM Reference Format:
Yanfeng Liu and Lefei Zhang. 2025. Multimodal Decomposed Distillation with Instance Alignment and Uncertainty Compensation for Thermal Object Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755841

*Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755841

1 Introduction

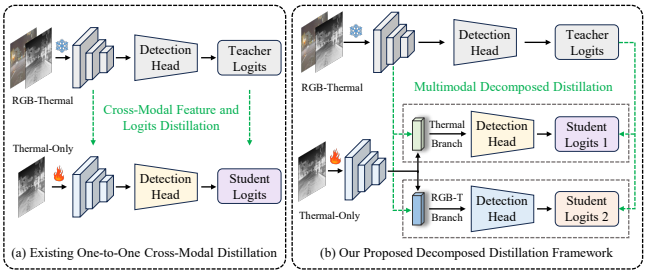

Figure 1: Illustration of the mainstream cross-modal object detection paradigms and our proposed multimodal decomposed distillation framework for thermal object detection.

RGB-Thermal (RGB-T) object detection aims to locate and identify the potential objects by leveraging optical and thermal image pairs. As a fundamental vision task [29], it offers broad utility for autonomous driving [9], remote sensing [30], and multimedia perception [2]. Most existing RGB-T detectors predominantly adopt dual-branch architectures with modality-specific encoders coupled with elaborate multimodal feature fusion modules to capture cross-modal complementary clues for object detection [14]. Despite their demonstrated accuracy, such designs suffer from inherent limitations: (1) its computational overhead restricts real-time deployment in edge device induced by parallel encoder networks (e.g., dual ResNet-50 backbones), and (2) poor adaptability in harsh RGB-failure unimodal scenarios (e.g., adverse weather or nighttime).

To address the above issues, some researchers have explored cross-modal knowledge distillation (KD) for RGB-T scene understanding, leveraging the robust multimodal generalization of a teacher model to guide the optimization of a unimodal student model. For instance, Zhang et al. [54] distill both modality-common and modality-specific knowledge from a powerful RGB-T teacher tracker to enhance the performance of a compact thermal student tracker. Similarly, Hnewa et al. [15] propose a distillation framework in which a multimodal teacher (trained on both RGB and gated images) enhances the robustness of a RGB student detector for pedestrian detection under low-light and adverse weather scenes. However, RGB imagery heavily depends on imaging conditions, showing high sensitivity to ambient factors like lighting variations and color distortions. In contrast, thermal imaging remains robust under varying external conditions (e.g., illumination or extreme weather), enabling reliable pedestrian and vehicle detection at lower hardware costs. Inspired by cross-modal knowledge

distillation [17, 31, 50], we propose to build a thermal-only student detector guided by complementary knowledge priors distilled from a strong RGB-T teacher model, aiming to mitigate environmental dependencies while maintaining efficiency, and yield a universal solution for both traffic surveillance and remote sensing scenarios.

As illustrated in Figure 1(a), existing cross-modal KD algorithms for RGB-T object detection predominantly employ a one-to-one distillation paradigm [18, 21, 40] that transfers feature-level, relation-level, or prediction-level knowledge from a multimodal teacher model to a unimodal student model. While this approach is straightforward and computationally efficient, it suffers from a critical limitation due to the inherent misalignment between the multimodal teacher's and unimodal student's feature/prediction spaces, where indiscriminate knowledge transfer may disrupt the student's unimodal feature representation and induce semantic conflicts between the unimodal and multimodal semantic spaces. To address this issue, we propose an adaptive decoupled learning framework that decomposes the unimodal student's features into two distinct representation spaces, as shown in Figure 1(b), where one branch aligns with the teacher's multimodal joint representation through distillation while the other preserves thermal-specific semantics, enabling the student model to effectively integrate its unimodal contextual information with complementary multimodal guidance while maintaining the integrity of thermal features, thereby facilitating a more compact and robust thermal detector.

To achieve the aforementioned research objectives, we design a channel-adaptive prompt learning mechanism that dynamically routes the top-K salient channels of student features to the multimodal distillation branch while preserving the remaining channels as thermal-specific representations. The decomposed branches of the student model are respectively guided by both the RGB-T fused features and the unimodal thermal backbone features from the multimodal teacher, thereby mitigating semantic conflicts across different modal spaces. Beyond feature-level distillation for both decomposed branches, we further introduce two independent detection heads for prediction-level distillation as shown in Figure 1(b). To effectively leverage the dual-branch representation, we propose a frequency-guided dynamic integration module that adaptively combines cross-branch contextual knowledge for enhanced detection. During inference, the auxiliary detection heads of thermal student could be discarded, with only the joint detection head processing the dynamically integrated dual-branch features. Furthermore, we develop an instance feature alignment distillation strategy to transfer high-confidence object features from the teacher model. Additionally, we devise a centerness-guided uncertainty-aware logit distillation loss that penalizes low-confidence object predictions for anchor-free detectors. Extensive experiments on three RGB-T datasets demonstrate that the proposed method achieves superior performance in both AP and AP_{50} metrics compared with 9 state-of-the-art approaches, while maintaining some advantages across other AP-based metrics. Meanwhile, typical qualitative visualizations further validating the effectiveness of our framework.

The main contribution of this paper are summarized as follows.

- To cope with harsh RGB-failure unimodal conditions, we propose a multimodal decomposed distillation framework to develop a robust and compact thermal-only detector.
- To effectively decouple the multimodal joint representation and preserve thermal-specific information, we present a channel-adaptive prompt learning for cross-modal feature decomposition. To integrate the dual-branch decomposed context knowledge, we propose a frequency-guided dynamic integration module for cross-modal combination.
- We introduce an instance alignment method and a simple yet effective uncertainty-aware compensation strategy for customized cross-modal feature- and logit-level distillation.

2 Related Work

In this section, we briefly discuss the related studies about RGB-Thermal object detection, knowledge distillation for object detection, and cross-modal knowledge distillation as follows.

2.1 RGB-Thermal Object Detection

Benefiting from the complementary attributes of various imaging sensors, leveraging visible and thermal image pairs for multispectral pedestrian and vehicle detection makes increasing research interest. Recently, there are some RGB-T object detection datasets released publicly to foster the benchmark of this area, such as FLIR [49] and M3FD [26], where typical object categories include pedestrians, vehicles, and other objects in natural images from traffic scenarios. In addition, RGB-T object detection can also be applied in remote sensing images from aerial scenes [35].

Most existing research efforts aim to leverage complementary features of RGB-T image pairs. He et al. [14] simultaneously consider the cross-modal semantic conflicts and complementary information between visible and thermal images. Besides, some researchers exploit the cross-task correlation between RGB-T object detection and image fusion. For instance, Sun et al. [38] propose a detection-driven fusion network with object-aware content loss. Liu et al. [27] present target-aware adversarial learning and dual-perspective optimization formulation for image fusion and object detection. Zhao et al. [56] introduce a joint fusion and detection learning framework via meta-feature embedding. Additionally, due to the distinct imaging mechanisms between modalities, RGB-T image pairs often exhibit weak misalignment issues, which yields another research focus. For example, Yuan et al. [48] develop a two-branch feature alignment detector based on modality selection strategy. Chen et al. [4] investigate modality-invariant features for offset prediction and modal-specific features for cross-modal alignment. In addition to these CNN-based approaches, recent advancements have adapted DETR architectures to RGB-T object detection scenarios, demonstrating promising performance [10, 11].

2.2 Knowledge Distillation for Object Detection

Knowledge distillation has emerged as an effective solution for both model compression and performance enhancement, enabling the transfer of learned knowledge from computationally intensive teacher models to compact student networks while maintaining competitive accuracy [36]. Although initially developed for image classification tasks, this technique has been successfully extended to object detection tasks in recent years [24].

Early object detection distillation methods transferred knowledge either by mimicking the teacher's RoI features [22] or by

learning the teacher's intermediate hint features and soft logits [5]. Subsequently, researchers have developed more sophisticated mechanisms for feature distillation. For instance, Wang et al. [43] introduce selective distillation using foreground masks within RoI regions to prioritize critical features. Du et al. [8] propose a feature richness score to distill the most generalizable features for object detection. Furthermore, relational knowledge between different object instances and pixels has been actively explored [6]. For example, Yang et al. [47] propose global distillation to reconstruct cross-pixel relationships. Vibashan et al. [41] design an instance relation graph to guide contrastive representation learning in source-free domain adaptive object detection. Ni et al. [34] introduce a dual-relation distillation framework, incorporating both pixel-wise and instance-wise relation distillation. Recently, some novel techniques such as label-guided distillation [53], prediction-guided distillation [46], cross-head distillation [42], localization distillation [58], and structured knowledge distillation [52] have been introduced to object detection tasks.

2.3 Cross-Modal Knowledge Distillation

Cross-modal distillation leverages knowledge from either a dominant modality or joint multi-modal representations to boost the learning of weaker modalities, and has been widely applied in numerous visual understanding tasks [12]. For example, Dai et al. [7] propose a cross-modal distillation formulation which consists of sequential KD loss to facilitate RGB-based action recognition. Huo et al. [19] design a on-the-fly selection distillation approach to alleviate the modality imbalance and soft label misalignment issue. As for empirical research progress, Xue et al. [45] conduct a comprehensive understanding for cross-modal distillation, and propose modality Venn diagram and modality focusing hypothesis to analyze inter-modal relationships. Ma et al. [32] formalize the cross-modal gap as a knowledge misalignment problem, and introduce a meta-learning approach for modality alignment. With respect to cross-modal object detection, researchers have made diverse research efforts for various cross-modal scenarios, such as LiDAR-point clouds, BEV, event data, multiview images, and radar. For instance, Bang et al. [1] propose to transfer desirable characteristics of LiDAR features into radar representation. Zheng et al. [57] propose a multi-representation knowledge adaptation to transfers the knowledge to event-based object detection. Zhao et al. [55] design a cross-modal KD framework based on a strong LiDAR-Camera teacher to compensate for Camera-Radar detectors.

3 Methodology

This section first describes an overview of the proposed multimodal decomposed framework (DecomKD), then details several key components for cross-modal feature decomposition and adaptive integration, as well as the feature- and logit-level distillation.

3.1 Overview of the Proposed DecomKD

Figure 2 illustrates the overall workflow of our presented DecomKD. We introduce a well-trained multimodal detector as the teacher model to provide soft guidance for the thermal student model. For the RGB-T teacher detector, we adopt a widely-used paradigm that equips dual ResNet50 backbones fo modal-specific feature extraction and integrates cross-modal features through channel-wise concatenation. Subsequently, the teacher incorporates a feature pyramid network (FPN) to capture multiscale contextual information, followed by a detection head for joint object localization and classification. As for thermal student model, we introduce a ResNet50 backbone with a FPN for multi-stage context learning. Then, we design the channel-adaptive prompt decomposition (CAPD) to decouple the student features into two individual branches, i.e., one for thermal-only knowledge preservation and the other for cross-modal feature compensation. The above two branches both contain FPNs and detection heads for model training. To obtain both thermal-only and cross-modal features from the teacher model, we introduce a trainable FPN after the teacher's thermal backbone to align multi-level features between the teacher's thermal representations and the student's thermal-only decomposed branch, followed by our proposed instance-level feature alignment distillation (IFAD) that selectively performs cross-modal feature alignment and distillation only on the student's top-k predicted instances, effectively eliminating interference from invalid predictions and background regions while focusing distillation on foreground objects. Additionally, we introduce the frequency-guided dynamic integration (FGDI) to aggregate multi-scale features from the two decomposed branches and equip a joint detection head as the final output prediction.

During the training stage of student, in addition to the ground truth supervision, we incorporate feature-level distillation based on IFAD and prediction-level distillation with uncertainty-ware logit distillation (UALD). The overall loss function for training the student model can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{Det} + \mathcal{L}_{IFAD} + \mathcal{L}_{UALD}, \tag{1}$$

where the detection head of thermal-only branch and the joint head are both supervised by $\mathcal{L}_{Det}$ using ground truths, while the features from both thermal-only and RGB-T branches undergo distillation supervision through $\mathcal{L}_{IFAD}$ and $\mathcal{L}_{UALD}$, as illustrated in Figure 2.

During the inference phase, the multimodal teacher is removed, and the detection heads of dual decomposed branches of student can also be omitted. Only the CAPD, FGDI, and joint head are retained for inference, ensuring considerable computational efficiency.

3.2 Channel-Adaptive Prompt Decomposition

Conventional distillation approaches for object detection typically perform one-to-one feature and output distillation between teacher and student models, which works effectively when both models share identical modalities. However, in cross-modal scenarios, the multi-scale features from teacher and student models inherently contain divergent contextual representations and exhibit uncertain semantic conflicts. These conflicts may significantly compromise distillation effectiveness, presenting a critical issue that has been largely overlooked [16, 20, 44]. To address this challenge, we propose a novel feature decomposition method named CAPD, as shown in Figure 2(b). This approach effectively decouples student features into two independent branches while simultaneously learning the teacher's multi-modal semantic representations and preserving the integrity of the student's thermal-only information.

As illustrated in Figure 2(b), for a deep feature $X \in \mathbb{R}^{B\times C\times H\times W}$ of the student, we first apply global average pooling (GAP) to obtain

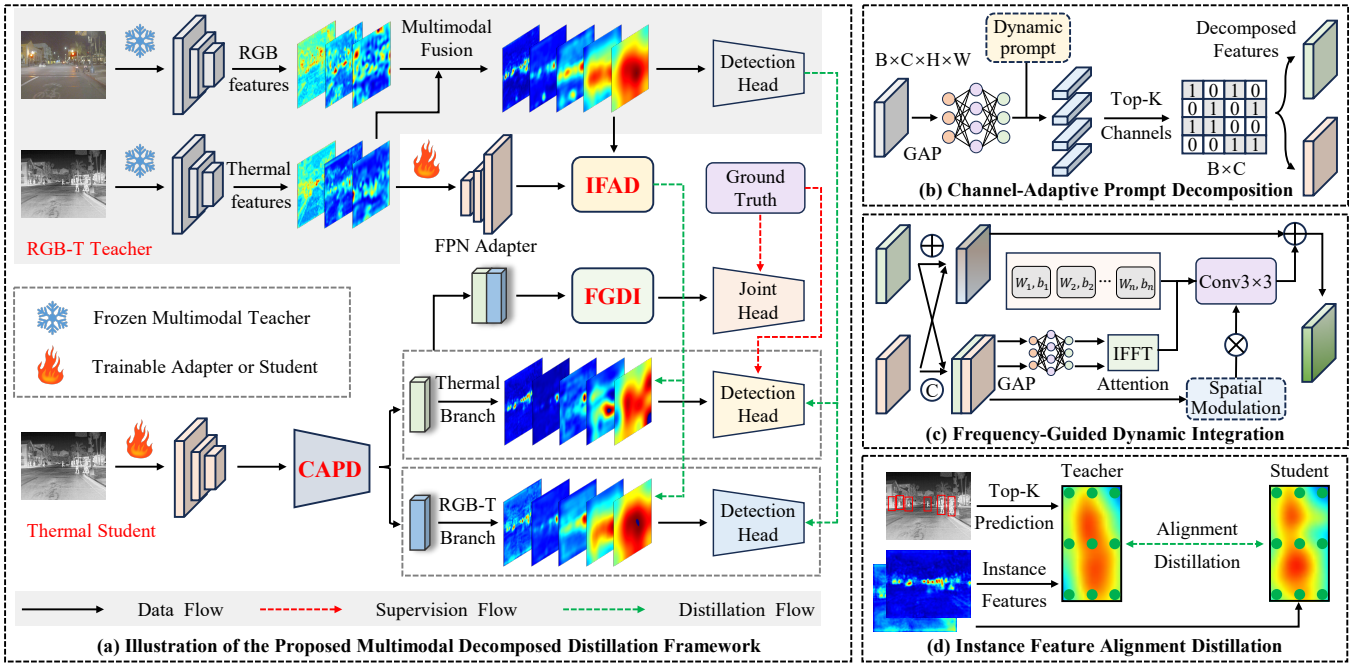

Figure 2: (a) Illustration of the proposed multimodal decomposed distillation framework. (b) Channel-Adaptive Prompt Decomposition (CAPD). (c) Frequency-Guided Dynamic Integration (FGDI). (d) Instance Feature Alignment Distillation (IFAD).

its channel-wise global representation as follows:

$$X_{avg} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} X[:,\ :,\ h, w] \in \mathbb{R}^{B\times C}, \tag{2}$$

where B indicates batch size of the input images, C is the number of channels, H and W denote height and width of the feature map.

After that, we employ two fully connected layers to perform non-linear learning on X_{avg}, capturing deep interdependencies among channels. Then, a dynamic prompt pool is utilized to map channel-wise feature richness scores, which define the routing weights for determining whether channel features belong to the thermal-only decomposed branch or RGB-T decomposed branch, i.e.,

$$S = softmax(\tau \cdot \frac{\phi(X_{avg})^{\mathrm{T}} \cdot P}{||\phi(X_{avg})|| \cdot ||P||}) \in \mathbb{R}^{B\times C}, \tag{3}$$

where S is the dynamic selection matrix, $\phi(\cdot)$ represents the fully connected layers, and τ is a learnable temperature that controls the sharpness of the routing distribution. The dynamic prompt pool $P \in \mathbb{R}^{d\times C}$ is jointly optimized with the network parameters during training to learn discriminative channel-wise representations.

Based on the dynamic selection matrix, we employ a top-k strategy to decompose the student feature: one group of channels is selected to preserve thermal-specific patterns in the student, while the remaining channels are allocated to learn cross-modal guidance features distilled from the teacher's RGB-T representations, i.e.,

$$\mathcal{T}_i = top\text{-}k(S[i,\ :]), \quad \mathcal{R}_i = \{1, ..., C\} - \mathcal{T}_i, \tag{4}$$

where $\mathcal{T}_i$ denotes the top-k channel indices for the i-th image within a batch, and $\mathcal{R}_i$ represents the remaining indices. This differentiable decomposition enables simultaneous optimization of both modality-specific feature preservation and cross-modal knowledge transfer during end-to-end training. The two decomposed features can be aggregated according to all top-k channel indices as follows:

$$F_1[i] = X[i, \mathcal{T}_i, :,\ :], \quad F_2[i] = X[i, \mathcal{R}_i, :,\ :] \in \mathbb{R}^{C/2\times H\times W}, \tag{5}$$

To ensure balanced decomposition, we set $k = C/2$, making both decomposed features F_1 and F_2 have equal channel dimensions of $C/2$. Then, we can stack the decomposed features from each sample to form the aggregated representations for the entire batch,

$$F_1 = \text{stack}(\{F_1[i]\}_{i=1}^{B}), \quad F_2 = \text{stack}(\{F_2[i]\}_{i=1}^{B}). \tag{6}$$

Furthermore, we incorporate point-wise convolutions and spatial attention mechanisms to model the internal feature correlations within each decomposed branch, i.e., $\tilde{F}_i = \sigma(C_{3\times3}(F_i)) \odot C_{1\times1}(F_i)$, where $\sigma(\cdot)$ indicates the sigmoid function, $\odot$ is the element-wise Hadamard product, and $C_{i\times i}(\cdot)$ denotes the $i \times i$ convolution. Finally, we introduce residual connections to maintain partial correlation between the thermal-only branch and cross-modal branch while accelerating the convergence of distillation training, i.e., $Y_1 = C_{3\times3}(\tilde{F}_1 + X), Y_2 = C_{3\times3}(\tilde{F}_2 + X)$, where Y_1 and Y_2 denote the final decomposed representations for thermal-specific preservation and cross-modal compensation.

As revealed in Figure 2(a), both decomposed branches of the student model are equipped with the IFAD distillation, and thus the total IFAD distillation loss could be formulated as

$$L_{IFAD} = L_{IFAD}(Y_1, FPN(T_{thermal})) + L_{IFAD}(Y_2, T_{rgb+t}), \tag{7}$$

where $T_{thermal}$ and T_{rgb+t} denote teacher's thermal backbone features and multimodal combined features, and $FPN(\cdot)$ stands for

the trainable FPN layer to align the student's decomposed features with the teacher's thermal backbone features.

3.3 Frequency-Guided Dynamic Integration

To effectively explore the spectral discrepancy and integrate multimodal representations with dual decomposed features of the student model, we propose FGDI that dynamically combines cross-modal knowledge through frequency-domain kernel modulation and attention-based aggregation. As shown in Figure 2(c), the module contains three key components: (1) frequency-disjoint weight modulation, (2) dynamic kernel aggregation, and (3) spatial-frequency adaptive fusion. The module processes two decomposed features $Y_1 \in \mathbb{R}^{B\times C\times H\times W}$ and $Y_2 \in \mathbb{R}^{B\times C\times H\times W}$ through parallel pathways to achieve complementary feature enhancement.

The first direct path generates a basic fusion through element-wise combination, while the second compression path processes the concatenated features to reduce dimensionality as follows:

$$F_{direct} = C_{3\times3}(Y_1 + Y_2), \quad F_{comp} = C_{1\times1}([Y_1, Y_2]), \tag{8}$$

where $[\cdot,\cdot]$ indicates the channel-wise concatenation.

The frequency-disjoint modulation processing begins with the decomposition of learnable weights into real and imaginary components across total K frequency bands, i.e.,

$$W_k = W_k^R + jW_k^I \quad \text{for} \quad k = 1, ..., K, \tag{9}$$

where $W_k^R,\ W_k^I \in \mathbb{C}^{C_{out}\times C_{in}\times K_h\times K_w}$ are learnable parameters of the real and imaginary weight tensors for the k-th frequency band, $j = \sqrt{-1}$ is the imaginary unit, C_{out} is the output channels, and C_{in} denotes the input channels. Each frequency band undergoes dynamic selection via a frequency band modulation as follows:

$$M_k^R,\ M_k^I = \tanh(C_{1\times1}(GAP(F_{comp}))) \in \mathbb{R}^{C_{out}\times1\times1}, \tag{10}$$

where $M_k^R, M_k^I \in \mathbb{R}^{C_{out}\times1\times1}$ are the frequency band modulation coefficients, $tanh(\cdot)$ denotes the Tanh activation function, and $GAP(\cdot)$ denotes the function of global average pooling as Equation (2).

These frequency-domain learnable parameters and frequency band coefficients are combined and transformed to spatial kernels through inverse Fourier transform with dynamic modulation, i.e.,

$$\mathcal{W}_k = \mathcal{F}^{-1}(M_k^R \odot W_k^R,\ M_k^I \odot W_k^I) \in \mathbb{R}^{C_{out}\times C_{in}\times K_h\times K_w}, \tag{11}$$

where $\mathcal{F}^{-1}(\cdot,\cdot)$ represents the inverse fast Fourier transform (IFFT).

FGDI employs an attention mechanism to dynamically aggregate kernels across frequency bands. The attention weights $\alpha = softmax(\phi(GAP(F_{comp}))) \in \mathbb{R}^{B\times K}$ are computed through channel-wise feature transformation in a nonlinear manner. These weights are used to generate the dynamic convolution kernel [23] through weighted sum-to-one strategy, i.e.,

$$\mathcal{W} = \sum_{k=1}^{K} \alpha_k \cdot \mathcal{W}_k \in \mathbb{R}^{B\times C_{out}\times C_{in}\times K_h\times K_w}, \tag{12}$$

where $\mathcal{W}$ is the weights matrix of a $K_h \times K_w$ convolution, e.g., 3×3 convolution. The dynamic convolution operation is implemented as a group convolution across the batch dimension for computational efficiency, which can be defined as:

$$F''_{comp} = ReLU(\mathcal{W} \cdot F'_{comp} + \mathcal{B}) \in \mathbb{R}^{B\times C_{out}\times H\times W}, \tag{13}$$

where F'_{comp} is the batch-collapsed version of F_{comp}, and $\mathcal{B} = \sum_{k=1}^{K} \alpha_k \cdot b_k$ is the combined bias matrix, where b_k is learning parameter. As shown in Figure 2(c), the final output incorporates spatial modulation and connection from the direct fusion path. The spatial-wise adaptive integration employs sigmoid-activated operation to project a spatial modulation matrix $SMM = \sigma(C_{3\times3}(F_{comp})) \in \mathbb{R}^{B\times C_{out}\times H\times W}$, and the final output can be formulated as

$$F_{out} = F''_{comp} \odot SMM + F_{direct}, \tag{14}$$

where F_{out} is the combined decomposed feature representation of the student model, which serves as input for joint head to produce the final prediction. This integrated design provides three key advantages. (1) The frequency-disjoint weights preserve distinct frequency components during feature integration. (2) The dynamic kernel aggregation adaptively emphasizes the most relevant frequency bands. (3) The spatial-frequency fusion maintains both localized details and global contextual relationships. Therefore, FGDI enables effective cross-modal feature integration with modality-specific characteristics through frequency-guided processing, thus facilitating cross-modal object detection tasks.

3.4 Instance Feature Alignment Distillation

Existing feature-level distillation methods widely focus on holistic similarity learning between teacher and student features, such as MSE or L1 loss. However, recent studies have revealed that such global feature distillation is suboptimal [47], particularly because different models exhibit varying perceptual distributions for foreground objects and background noises, thereby significantly compromising the effectiveness of feature distillation. To address this limitation, we propose a novel customized distillation approach called IFAD, which selectively concentrates on the most critical foreground instance features while effectively filtering out irrelevant background noise and invalid object predictions. The workflow of our proposed IFAD is illustrated in Figure 2(d).

The core idea of IFAD lies in its dynamic selection of high-confidence foreground regions for instance distillation. During the training stage, we select the region features of the top-k samples with the highest classification confidence from the student's joint head predictions for distillation. This approach enables the student model to progressively refine its feature learning for the most discriminative foreground objects, ultimately enhancing the capability of how to identify objects. However, we avoid simple foreground mask weighting, instead, we focus on aligning both the peripheral and central contextual information of the instances, thereby achieving more sparse and effective instance-level feature distillation. To achieve this goal, we introduce an efficient instance alignment method that computes only nine key points per instance: eight extreme boundary points and the center point as shown in Figure 2(d). These coordinates are processed through RoI Align [13] to extract their corresponding feature responses from both teacher and student feature maps, forming the distillation representation vectors for IFAD, the above process can be formulated as follows:

$$\mathcal{L}_{IFAD} = \frac{1}{N}\sum_{i=1}^{N}\sum_{p\in\mathcal{P}} \left\| \frac{\mathcal{A}(F_t^i(p))}{|\mathcal{A}(F_t^i(p))|_2} - \frac{\mathcal{A}(F_s^i(p))}{|\mathcal{A}(F_s^i(p))|_2} \right\|^2, \tag{15}$$

where N denotes the number of selected **top-k** instances with highest classification scores, $\mathcal{A}(\cdot)$ represents the RoI Align function, F_t^i and F_s^i indicate the i-th stage features of RGB-T teacher and thermal-only student, $\mathcal{P}$ is the sampled nine coordinates of an object instance, and $\|\cdot,\cdot\|^2$ denotes MSE loss for feature distillation.

3.5 Uncertainty-Aware Logit Distillation

Traditional object detection logit distillation typically employs L1 or cross-entropy loss for distilling the classification sub-task predictions between teacher and student models, and uses IoU, L1, or L2 loss for the regression sub-task [42]. However, these methods are inefficient for anchor-free detectors [39], which introduce an additional centerness prediction branch—a critical component that quantifies how close an object is to the center of a location, with values ranging from 0 to 1 (where -1 indicates an invalid box). We argue that effective distillation of centerness predictions is essential for anchor-free detectors. To address this, we propose an uncertainty-aware logit distillation method specifically designed to compensate for comprehensive prediction-level distillation from RGB-T teachers to thermal-only students in anchor-free detectors.

First of all, we regard the ground truth (GT) centerness score of each coordinate point as conditional information to guide the logit distillation process. For regions with invalid centerness scores (i.e., -1), we simply set the distillation weight to 1 to enforce knowledge transfer for invalid predictions. For valid regions (i.e., [0, 1]), we treat 0.5 as the symmetry axis, hypothesizing that predictions with centerness scores closer to 0.5 exhibit higher uncertainty and thus require penalization. To quantify this uncertainty, we propose a differentiable logarithmic function $\ln(2-|2c-1|)$, where c denotes the GT centerness score. To maintain balance with invalid regions (when $c=-1$), we set the distillation coefficients for valid regions as $\gamma = \ln(2-|2c-1|)+0.5$. Consequently, our proposed uncertainty-aware logit distillation can be expressed as

$$\mathcal{L}_{UALD} = \sum_{i=1}^{N} \gamma^i \cdot (\mathbf{KL}(p_t^i||p_s^i) + |b_t^i - b_s^i|_1 + |c_t^i - c_s^i|_1), \quad \text{where } \gamma^i = \begin{cases} 1, & c^i = -1, \\ \ln(2-|2c^i-1|)+0.5, & 0 \le c^i \le 1, \end{cases} \tag{16}$$

where $p_t^i, p_s^i \in \mathbb{R}^K$, $b_t^i, b_s^i \in \mathbb{R}^4$, and $c_t^i, c_s^i \in \mathbb{R}^1$ denote the i-th classification, regression, and centerness score predictions of the teacher and student, respectively. K defines the number of object categories, N indicates the total number of object predictions, $\mathbf{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence, and $|\cdot,\cdot|_1$ represents the L1 loss.

4 Experiments

4.1 Experimental Protocol

Dataset Description. We employ three RGB-T object detection datasets for experiments, i.e., FLIR [49], M3FD [26], and VEDAI [35]. Among them, the FLIR dataset contains 5,142 image pairs with the size of 512 × 640, including 4,129 images for training and 1,013 images for testing. It contains three object categories: bicycle, car, and person, with a total of 40,752 object instances. The M3FD dataset contains 4,200 image pairs covering six object categories, with a total of 34,408 object instances. Following the random split convention [51], we divide the dataset into 2,940 image pairs for training and 1,260 for testing. During training, all images were resized to a uniform resolution of 640 × 640 pixels. As for VEDAI dataset, it comprises 1,246 image pairs, all with a resolution of 1024 × 1024 pixels, covering eight distinct object categories. Following [28], we split 1,089 image pairs for training and 121 pairs for testing.

Evaluation Metrics. For a comprehensive comparison of all KD methods, we not only report basic metrics **AP** and $\mathbf{AP}_{50}$, but also introduce AP_{75}, AP_S, AP_M, and AP_L to evaluate detection accuracy under different IoU thresholds and for objects of varying scales.

Implementation Details. We use FCOS [39] and RetinaNet [25] as baseline methods. For the dual-modal detectors, two ResNet50 encoders loaded with pre-trained weights are employed to extract modal-specific features, and element-wise summation is used for feature fusion. All experiments are implemented on 4 NVIDIA RTX 4090 GPUs with PyTorch 2.1 toolbox. To ensure a fair comparison, we uniformly reproduce all methods. The AdamW optimizer is adopted for training with an initial learning rate of 2e-4. The number of epochs is 60, and the learning rate is reduced to 2e-5 in the last 30 epochs. For the FLIR and VEDAI datasets, the batch size is set to 8, while for VEDAI, it is 2. During training, we introduce random horizontal flipping, crop resizing, and rotation for augmentation. The number of top-k in IFAD is set to $min(30, N)$, where N is the number of predicted instances in each feature map of students.

4.2 Comparison with State-of-the-Art Methods

In this paper, we conduct extensive experiments on three RGB-T datasets to validate the performance of the proposed DecomKD.

FLIR dataset. As shown in Table 1, we report the AP_{50} for three categories (bicycle, car, and person) along with six AP-based metrics on the full dataset. Among all competitors, our proposed DecomKD achieves the best performance on seven metrics and secures top-3 results in both bicycle and AP_L. Notably, it outperforms the baseline student model by over **4%** in AP_{50}, significantly surpassing all comparative methods. This demonstrates the effectiveness of our proposed cross-modal decomposed framework. Some methods, such as FitNet, FGFI, SKD, and SRD, fail to achieve comprehensive improvements across all metrics. In contrast, our approach exhibits more consistent and superior performance across all indicators.

M3FD dataset. Similarly, our DecomKD also shows superior performance on M3FD, achieving best results on five metrics and showing further advantages over the state-of-the-art CrossKD [42] and LogitStdKD [37] in Table 2. Notably, DecomKD achieves approximately **4%** gain in AP_{50} and AP_S, along with a **6%** boost in AP_L. These results validate the effectiveness of the proposed CAPD and FGDI modules, as well as IFAD and UALD distillation methods.

VEDAI dataset. As reported in Table 3, our proposed DecomKD demonstrates dominant superiority among numerous state-of-the-art competitors. Specifically, it achieves a 4.5% AP_{50} improvement over the FCOS-based thermal student model and establishing itself as the sole method exceeding 47% AP. In summary, the above comprehensive experiments across three datasets validate DecomKD's strong generalization capability across diverse dataset scenarios.

4.3 Ablation Study

As shown in Table 4, we conduct ablation study on the FLIR dataset with FCOS as baseline detector. While introducing IFAD and UALD,

Table 1: Quantitative comparison on the FLIR [49]. The top three results are marked in red, green, and blue, respectively.

FCOS [39]	Publication	Bicycle	Car	Person	**AP**	**AP**$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
RGB-T Teacher	TPAMI 2022	57.74	84.61	77.58	35.62	73.31	28.59	24.28	41.54	50.55
Thermal Student	TPAMI 2022	47.90	77.80	70.49	31.27	65.40	24.71	21.57	36.67	43.26
+ FitNet [36]	ICLR 2016	47.60	79.67	73.15	31.45	66.81	25.26	22.53	36.58	38.95
+ FGFI [43]	CVPR 2019	48.23	82.12	72.26	32.41	67.54	26.69	22.97	37.69	42.88
+ FRS [8]	NeurIPS 2021	49.33	81.19	72.56	32.46	67.69	26.62	22.79	37.92	43.95
+ FGD [47]	CVPR 2022	50.36	78.37	71.00	31.32	66.58	24.54	21.85	36.79	40.16
+ PKD [3]	NeurIPS 2022	48.61	79.64	71.45	30.18	66.57	22.80	20.47	35.47	44.55
+ SKD [52]	TPAMI 2023	54.08	77.49	71.92	32.03	67.83	25.45	23.13	36.94	41.73
+ SRD [33]	AAAI 2024	52.04	80.34	71.51	31.76	67.97	24.40	22.13	37.34	41.80
+ CrossKD [42]	CVPR 2024	48.80	81.49	74.37	32.42	68.22	25.95	22.47	38.18	41.95
+ LogitStdKD [37]	CVPR 2024	51.47	80.56	72.74	31.87	68.26	25.43	22.16	37.71	40.04
+ Our Proposed	—	51.85	82.84	74.79	33.32	69.83	26.70	24.11	38.37	43.69

Table 2: Quantitative comparison on M3FD [26].

FCOS [39]	**AP**	**AP**$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
RGB-T Teacher	48.97	77.94	51.85	29.05	61.60	77.90
Thermal Student	41.40	68.99	43.02	21.54	53.48	73.20
+ FitNet [36]	42.46	70.22	43.17	22.68	54.11	72.38
+ FGFI [43]	42.73	71.20	44.37	23.81	53.50	74.14
+ FRS [8]	43.37	71.19	44.42	22.47	54.98	73.29
+ FGD [47]	41.61	69.65	42.66	21.87	52.75	73.02
+ PKD [3]	41.86	69.19	43.43	20.68	53.76	77.66
+ SKD [52]	42.99	70.90	44.86	22.87	54.80	72.97
+ SRD [33]	42.56	70.03	43.43	22.98	53.50	78.89
+ CrossKD [42]	42.53	70.22	44.88	21.94	54.22	74.42
+ LogitStdKD [37]	41.96	69.49	42.79	21.70	52.23	72.78
+ Our Proposed	43.89	73.01	45.04	25.22	54.27	79.28

Table 3: Anchor-free comparison on VEDAI [35].

FCOS [39]	**AP**	**AP**$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
RGB-T Teacher	49.20	80.04	55.54	41.17	51.96	60.10
Thermal Student	45.14	73.39	51.48	34.25	48.68	52.91
+ FitNet [36]	46.18	76.11	51.92	37.32	48.64	52.15
+ FGFI [43]	46.17	75.88	52.01	37.67	49.52	52.28
+ FRS [8]	45.87	76.67	49.13	41.97	47.01	42.06
+ FGD [47]	46.57	76.75	52.45	35.71	49.65	56.17
+ PKD [3]	46.26	76.55	50.72	35.43	49.71	55.76
+ SKD [52]	46.00	75.46	50.10	43.46	47.54	52.28
+ SRD [33]	46.77	76.35	52.28	39.43	49.60	48.31
+ CrossKD [42]	46.79	76.17	51.93	41.29	49.26	52.42
+ LogitStdKD [37]	46.70	76.99	50.15	43.16	48.56	42.33
+ Our Proposed	47.77	77.91	52.53	43.07	49.91	56.93

Table 4: Ablation study with FCOS on the FLIR dataset [49].

Methods	CAPD	FGDI	IFAD	UALD	AP	AP$_{50}$
RGB-T	–	–	–	–	35.62	73.31
Thermal	–	–	–	–	31.27	65.40
Thermal	✓				31.96	66.91
Thermal	✓	✓			32.60	68.17
Thermal	✓	✓	✓		32.86	69.34
Thermal	✓	✓	✓	✓	**33.32**	**69.83**

we employ FitNet [36] as the distillation method. Through observation, both CAPD and FGDI demonstrate improvements exceeding 1% in terms of AP$_{50}$. Furthermore, we design two finer-grained distillation approaches at the feature- and prediction-level, i.e., IFAD and UALD, which also prove their effectiveness in further enhancing the cross-modal distillation of DecomKD for thermal-only detection.

4.4 Extension Experiments

Table 5: Anchor-based comparison on VEDAI [35].

RetinaNet [25]	AP	AP$_{50}$	AP$_{75}$
RGB-T Teacher	52.50	85.18	59.36
Thermal Student	47.88	78.29	54.63
+ FitNet [36]	48.54	79.73	52.29
+ FGFI [43]	49.22	79.07	55.12
+ FRS [8]	49.11	79.47	57.36
+ PKD [3]	49.09	80.46	55.43
+ SKD [52]	48.41	78.91	54.55
+ CrossKD [42]	48.54	79.73	52.29
+ LogitStdKD [37]	49.95	81.05	57.32
+ Our Proposed	50.30	81.45	57.13

Extension to Anchor-Based Detector. To validate the model-agnostic capability of DecomKD, we conduct experiments on the VEDAI dataset using anchor-based RetinaNet [25] as baseline method, with quantitative results presented in Table 5. Notably, UALD in DecomKD is only applicable to anchor-free detectors and thus is excluded in anchor-based RetinaNet. Despite this limitation, the results still show that our proposed framework maintains decent performance in AP, AP$_{50}$, and AP$_{75}$ metrics. This evidences DecomKD's generalization across different detection architectures.

Extension to RGB-only students. To validate the modality-agnostic capability of DecomKD, we implement an RGB-only student with cross-modal knowledge distillation for the M3FD dataset [26] based on FCOS [39], with experimental results detailed in Table 6. The quantitative results demonstrate that DecomKD consistently surpasses state-of-the-art CrossKD and LogitStdKD, achieving a leading AP$_{50}$ of over 74% (nearly 6% performance gain). This comparative analysis confirms its modality-agnostic characteristics.

4.5 Visual Analysis

Here, we provide visual analysis about detection results and features. **More visualization analysis is presented in Appendix.**

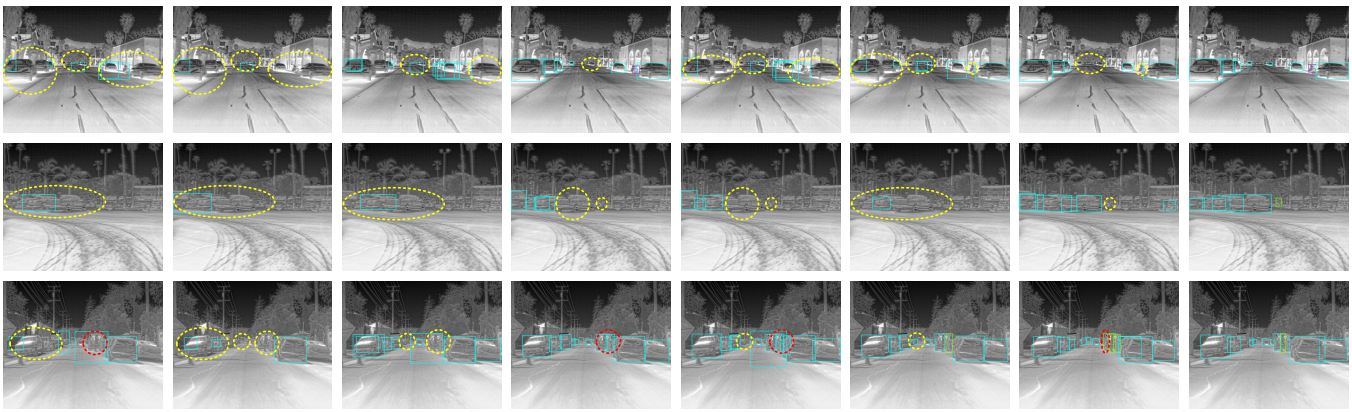

Figure 3: Detection visualization on the FLIR dataset. (a) Thermal student. (b) FRS. (c) SKD. (d) SRD. (e) CrossKD. (f) LogitStdKD. (g) Our DecomKD. (h) GT. Red dashed circles denote false detection, and yellow dashed circles indicate missing detection.

Table 6: Quantitative comparison on M3FD with RGB student.

FCOS [39]	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RGB-T Teacher	48.97	77.94	51.85	29.05	61.60	77.90
RGB Student	40.52	68.14	41.84	20.99	51.24	72.37
+ FGD [47]	43.63	72.18	45.40	23.55	55.84	75.58
+ PKD [3]	42.83	70.38	43.63	21.71	54.89	78.84
+ SKD [52]	43.83	72.52	44.98	23.49	55.42	76.71
+ SRD [33]	43.70	71.74	45.63	23.28	55.36	75.63
+ CrossKD [42]	44.10	72.64	45.50	24.88	54.82	71.69
+ LogitStdKD [37]	42.93	71.60	43.63	22.87	54.33	75.15
+ Our Proposed	**45.61**	**74.18**	**47.70**	**26.28**	**56.15**	**81.03**

Detection visualization. We present some typical samples on the FLIR dataset to show the predicted detection results for six competitive KD methods as shown in Figure 3. Comparative results reveal that while existing KD algorithms suffer significantly in scenarios with challenging foreground-background confusion, our DecomKD effectively enhances foreground object detection while minimizing both false positives and missing detection.

Feature visualization. To validate the rationality of dual-branch decomposition, we visualize the features of the student's dual branches, the aggregated features of FGDI, and the features of RGB-T teacher in Figure 4. The results clearly show that the dual-branch features in DecomKD successfully decompose foreground objects in shallow layers, enabling two branches to focus on nearby persons and distant cars, respectively. In intermediate layers, the decomposed branches exhibit similar spatial activation patterns, while in deep layers they capture distinct global contexts, reflecting their complementary nature. This is because one decomposed branch is designed to mimic the cross-modal representation from the RGB-T teacher, while the other preserves the integrity of thermal-specific features. Ultimately, the features combined by FGDI closely mimic those of the RGB-T teacher, thereby achieving more competitive performance. Additionally, the third and fourth rows of Figure 4(a) present heatmap from IFAD and uncertainty map from UALD. The visualizations reveal that IFAD and UALD specifically attend to the most accurately predicted foreground objects and the most confusing instances, thereby boosting cross-modal instance distillation.

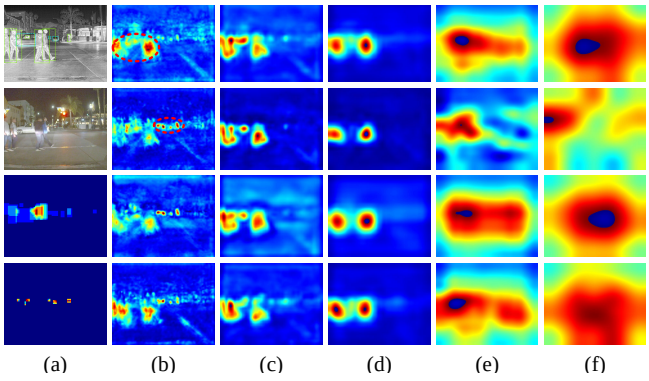

Figure 4: (a) From top to bottom rows represent GT, RGB image, heatmap of IFAD, and uncertainty map of foreground objects of UALD, respectively. (b)-(f) show five multiscale feature maps, from top to bottom rows correspond to the first decomposed branch, the second decomposed branch, the merged branch of FGDI, and the RGB-T teacher model.

5 Conclusion

In this paper, we present DecomKD to effectively transfer the knowledge from RGB-T teacher to thermal-only student for thermal object detection. To address the semantic conflict between simultaneously mimicking RGB-T teacher features and preserving thermal-specific student feature integrity, we propose CAPD to perform task-specific distillation learning for dual decomposed branches. Furthermore, we design FGDI to aggregate features from the two decomposed branches, followed by a joint head to output the student's final detection results. For more customized distillation of the most accurately predicted foreground objects and the most confusing instances, we develop IFAD and UALD to enhance cross-modal distillation at both feature-level and prediction-level. Extensive experiments across three datasets demonstrate the model-agnostic and modality-agnostic effectiveness of DecomKD, delivering consistent improvements for both anchor-free and anchor-based detectors, whether applied to thermal or RGB-based student models.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62431020, the Foundation for Innovative Research Groups of Hubei Province under Grant 2024AFA017, and the Fundamental Research Funds for the Central Universities under Grant 2042025kf0030.

References

[1] Geonho Bang, Kwangjin Choi, Jisong Kim, Dongsuk Kum, and Jun Won Choi. 2024. RadarDistill: Boosting Radar-based Object Detection Performance via Knowledge Distillation from LiDAR Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15491–15500.

[2] Francesco Bongini, Lorenzo Berlincioni, Marco Bertini, and Alberto Del Bimbo. 2021. Partially Fake it Till you Make It: Mixing Real and Fake Thermal Images for Improved Object Detection. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*. 5482–5490.

[3] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. 2022. PKD: General Distillation Framework for Object Detectors via Pearson Correlation Coefficient. In *Advances in Neural Information Processing Systems (NeurIPS)*. 15394–15406.

[4] Chen Chen, Jiahao Qi, Xingyue Liu, Kangcheng Bin, Ruigang Fu, Xikun Hu, and Ping Zhong. 2024. 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 (CVPR)*. 26836–26845.

[5] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. 2017. Learning Efficient Object Detection Models with Knowledge Distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[6] Shidi Chen, Lili Wei, Liqian Liang, and Congyan Lang. 2024. Joint Homophily and Heterophily Relational Knowledge Distillation for Efficient and Compact 3D Object Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*. 2127–2135.

[7] Rui Dai, Srijan Das, and François Bremond. 2021. Learning an Augmented RGB Representation With Cross-Modal Knowledge Distillation for Action Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13053–13064.

[8] Zhixing Du, Rui Zhang, Ming Chang, Xishan Zhang, Shaoli Liu, Tianshi Chen, and Yunji Chen. 2021. Distilling Object Detectors with Feature Richness. In *Advances in Neural Information Processing Systems (NeurIPS)*. 5213–5224.

[9] Wassim El Ahmar, Yahya Massoud, Dhanvin Kolhatkar, Hamzah AlGhamdi, Mohammad Alja'afreh, Riad Hammoud, and Robert Laganiere. 2023. Enhanced Thermal-RGB Fusion for Robust Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 365–374.

[10] Haolong Fu, Jin Yuan, Guojin Zhong, Xuan He, Jiacheng Lin, and Zhiyong Li. 2024. CF-Deformable DETR: An End-to-End Alignment-Free Model for Weakly Aligned Visible-Infrared Object Detection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 758–766.

[11] Junjie Guo, Chenqiang Gao, Fangcen Liu, Deyu Meng, and Xinbo Gao. 2024. DAMSDet: Dynamic Adaptive Multispectral Detection Transformer with Competitive Query Selection and Adaptive Feature Fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 464–481.

[12] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross Modal Distillation for Supervision Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2827–2836.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2020. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 2 (2020), 386–397.

[14] Xiao He, Chang Tang, Xin Zou, and Wei Zhang. 2023. Multispectral Object Detection via Cross-Modal Conflict-Aware Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*. 1465–1474.

[15] Mazin Hnewa, Alireza Rahimpour, Justin Miller, Devesh Upadhyay, and Hayder Radha. 2023. Cross Modality Knowledge Distillation for Robust Pedestrian Detection in Low Light and Adverse Weather Conditions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[16] Wuliang Huang, Yiqiang Chen, Xinlong Jiang, Chenlong Gao, Qian Chen, Teng Zhang, Bingjie Yan, Yifan Wang, and Jianrong Yang. 2024. Correlation-Driven Multi-Modality Graph Decomposition for Cross-Subject Emotion Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*. 2272–2281.

[17] Xun Huang, Hai Wu, Xin Li, Xiaoliang Fan, Chenglu Wen, and Cheng Wang. 2024. Sunshine to Rainstorm: Cross-Weather Knowledge Distillation for Robust 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2409–2416.

[18] Zhanchao Huang, Wei Li, and Ran Tao. 2023. Multimodal Knowledge Distillation for Arbitrary-Oriented Object Detection in Aerial Images. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[19] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. 2024. C2KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16006–16015.

[20] Pilhyeon Lee, Taeoh Kim, Minho Shim, Dongyoon Wee, and Hyeran Byun. 2023. Decomposed Cross-Modal Distillation for RGB-Based Temporal Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2373–2383.

[21] Ang Li, Shouxiang Ni, Yanan Chen, Jianxin Chen, Xin Wei, Liang Zhou, and Mohsen Guizani. 2023. Cross-Modal Object Detection via UAV. *IEEE Transactions on Vehicular Technology* 72, 8 (2023), 10894–10905.

[22] Quanquan Li, Shengying Jin, and Junjie Yan. 2017. Mimicking Very Efficient Network for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6356–6364.

[23] Yunsheng Li, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Ye Yu, Lu Yuan, Zicheng Liu, Mei Chen, and Nuno Vasconcelos. 2021. Revisiting Dynamic Convolution via Matrix Decomposition. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 1–11.

[24] Siyuan Liang, Aishan Liu, Jiawei Liang, Longkang Li, Yang Bai, and Xiaochun Cao. 2022. Imitated Detectors: Stealing Knowledge of Black-box Object Detectors. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*. 4839–4847.

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 2 (2020), 318–327.

[26] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, et al. 2022. 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 Conference on Computer Vision and Pattern Recognition (CVPR)*. 5802–5811.

[27] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. 2022. 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 (CVPR)*. 5802–5811.

[28] Yanfeng Liu, Wei Guo, Chaojun Yao, and Lefei Zhang. 2025. Dual-Perspective Alignment Learning for Multimodal Remote Sensing Object Detection. *IEEE Transactions on Geoscience and Remote Sensing* 63 (2025), 1–15.

[29] Yanfeng Liu, Qiang Li, Yuan Yuan, Qian Du, and Qi Wang. 2022. ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14.

[30] Yanfeng Liu, Qiang Li, Yuan Yuan, and Qi Wang. 2022. Single-Shot Balanced Detector for Geospatial Object Detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2529–2533.

[31] Andong Lu, Jiacong Zhao, Chenglong Li, Yun Xiao, and Bin Luo. 2024. Breaking Modality Gap in RGBT Tracking: Coupled Knowledge Distillation. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*. 9291–9300.

[32] Wenxuan Ma, Shuang Li, Lincan Cai, and Jingxuan Kang. 2024. Learning Modality Knowledge Alignment for Cross-Modality Transfer. In *Proceedings of the International Conference on Machine Learning (ICML)*. 1–13.

[33] Roy Miles and Krystian Mikolajczyk. 2024. Understanding the Role of the Projector in Knowledge Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 4233–4241.

[34] Zhen-Liang Ni, Fukui Yang, Shengzhao Wen, and Gang Zhang. 2023. Dual Relation Knowledge Distillation for Object Detection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 1276–1284.

[35] Sebastien Razakarivony and Frederic Jurie. 2016. Vehicle Detection in Aerial imagery : A Small Target Detection Benchmark. *Journal of Visual Communication and Image Representation* 34 (2016), 187–203.

[36] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for Thin Deep Nets. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 1–9.

[37] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. 2024. Logit Standardization in Knowledge Distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 15731–15740.

[38] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. 2022. DetFusion: A Detection-driven Infrared and Visible Image Fusion Network. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*. 4003–4011.

[39] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2022. FCOS: A Simple and Strong Anchor-Free Object Detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2022), 1922–1933.

[40] Xiaozhong Tong, Xiaojun Guo, Xiaoyong Sun, Runze Guo, Shaojing Su, and Zhen Zuo. 2025. CMDistill: Cross-Modal Distillation Framework for AAV Image Object Detection. *IEEE Journal of Selected Topics in Applied Earth Observations*

and Remote Sensing 18 (2025), 1395–1409.

[41] Vibashan VS, Poojan Oza, and Vishal M. Patel. 2023. Instance Relation Graph Guided Source-Free Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3520–3530.

[42] Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. 2024. CrossKD: Cross-Head Knowledge Distillation for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16520–16530.

[43] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. 2019. Distilling Object Detectors With Fine-Grained Feature Imitation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4933–4942.

[44] Haoran Xu, Peixi Peng, Guang Tan, Yuan Li, Xinhai Xu, and Yonghong Tian. 2024. DMR: Decomposed Multi-Modality Representations for Frames and Events Fusion in Visual Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26508–26518.

[45] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. 2023. The Modality Focusing Hypothesis: Towards Understanding Crossmodal Knowledge Distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[46] Chenhongyi Yang, Mateusz Ochal, Amos Storkey, and Elliot J. Crowley. 2022. Prediction-Guided Distillation for Dense Object Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 123–138.

[47] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. 2022. Focal and Global Knowledge Distillation for Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4643–4652.

[48] Maoxun Yuan, Yinyan Wang, and Xingxing Wei. 2022. Translation, Scale and Rotation: Cross-Modal Alignment Meets RGB-Infrared Vehicle Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 509–525.

[49] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. 2020. Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. 276–280.

[50] Haiming Zhang, Xu Yan, Dongfeng Bai, Jiantao Gao, Pan Wang, Bingbing Liu, Shuguang Cui, and Zhen Li. 2024. RadOcc: Learning Cross-Modality Occupancy Knowledge through Rendering Assisted Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 7060–7068.

[51] Jiaqing Zhang, Mingxiang Cao, Weiying Xie, Jie Lei, Daixun Li, Wenbo Huang, Yunsong Li, and Xue Yang. 2024. E2E-MFD: Towards End-to-End Synchronous Multimodal Fusion Detection. In *Advances in Neural Information Processing Systems (NeurIPS)*. 52296–52322.

[52] Linfeng Zhang and Kaisheng Ma. 2023. Structured Knowledge Distillation for Accurate and Efficient Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 12 (2023), 15706–15724.

[53] Peizhen Zhang, Zijian Kang, Tong Yang, Xiangyu Zhang, Nanning Zheng, and Jian Sun. 2022. LGD: Label-Guided Self-Distillation for Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 3309–3317.

[54] Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. 2023. Efficient RGB-T Tracking via Cross-Modality Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5404–5413.

[55] Lingjun Zhao, Jingyu Song, and Katherine A. Skinner. 2024. CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15470–15480.

[56] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. 2023. MetaFusion: Infrared and Visible Image Fusion via Meta-Feature Embedding From Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13955–13965.

[57] Xu Zheng and Lin Wang. 2024. EventDance: Unsupervised Source-free Cross-modal Adaptation for Event-based Object Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17448–17458.

[58] Zhaohui Zheng, Rongguang Ye, Qibin Hou, Dongwei Ren, Ping Wang, Wangmeng Zuo, and Ming-Ming Cheng. 2023. Localization Distillation for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 10070–10083.