Feature Fusion and Knowledge-Distilled Multi-Modal Multi-Target Detection

Ngoc Tuyen Do* and Tri Nhu Do†
*School of Information and Communications, Hanoi University of Science and Technology
†Telecom Neural Detection Lab, Polytechnique Montréal, Montreal, QC, Canada
Emails: tuyen.dn242305m@sis.hust.edu.vn, tri-nhu.do@polymtl.ca

***Abstract*—In the surveillance and defense domain, multi-target detection and classification (MTD) is considered essential yet challenging due to heterogeneous inputs from diverse data sources and the computational complexity of algorithms designed for resource-constrained embedded devices, particularly for AI-based solutions. To address these challenges, we propose a feature fusion and knowledge-distilled framework for multi-modal MTD that leverages data fusion to enhance accuracy and employs knowledge distillation for improved domain adaptation. Specifically, our approach utilizes both RGB and thermal image inputs within a novel fusion-based multi-modal model, coupled with a distillation training pipeline. We formulate the problem as a posterior probability optimization task, which is solved through a multi-stage training pipeline supported by a composite loss function. This loss function effectively transfers knowledge from a teacher model to a student model. Experimental results demonstrate that our student model achieves approximately 95% of the teacher model's mean Average Precision while reducing inference time by approximately 50%, underscoring its suitability for practical MTD deployment scenarios.**

***Index Terms*—Mutli-target detection, knowledge distillation, feature fusion, optimization, AI/ML, FLIR, thermal data, RGB**

I. Introduction

Multi-target detection (MTD), which aims to identify and classify multiple targets simultaneously, is a pivotal task in applications such as surveillance, autonomous systems, and radar-based tracking. Traditional MTD approaches predominantly relied on hand-crafted feature engineering coupled with statistical models. For instance, the Histogram of Oriented Gradients (HOG) features combined with Support Vector Machine (SVM) classifiers [1] were widely adopted for object detection and classification tasks prior to the advent of deep learning, with numerous algorithmic variants achieving notable success. Deep learning has significantly enhanced MTD capabilities by enabling the automatic extraction of rich, hierarchical, and non-linear features directly from raw data. In particular, Convolutional Neural Networks (CNNs) serve as the foundational architecture for state-of-the-art (SOTA) object detection models, such as the EfficientDet family [2] and various iterations of MobileNet [3].

Several challenges in MTD must be overcome, including aligned data fusion from heterogeneous inputs (e.g., RGB, thermal), large model sizes that deter response of edge embedding devices, and limited generalization to difficult environments. On the one hand, deep learning Knowledge Distillation (KD) approach [4] is a training technique where a large, accurate *teacher model* conveys pre-trained knowledge to a compact *student model* without significant performance degradation. Unlike single-output tasks, MTD produces multiple bounding boxes accompanied by labels and scores, which makes it challenging to define the posterior problem formulation and determine the knowledge to be transferred (e.g., classification logits, localization features). To address the complexity of the training procedure, the loss function of the proposed method requires careful architectural and algorithmic design to produce effective distillation. On the other hand, *multi-modal* sensor fusion has been explored [5] to enhance MTD by integrating complementary information from diverse input sensors, addressing adverse weather conditions, such as RGB and thermal. DeepInversion for Object Detection (DIODE) [6] developed a data-free KD model with significant improvements, including data augmentations and an automated bounding box and category sampling scheme. CrossKD [7] is a novel KD method that significantly enhances the AP of GFL ResNet-50, where the intermediate feature maps of the student model are conveyed to the teacher to receive contradictory supervision signals.

In this paper, we investigate the MTD problem in the context of autonomous driving surveillance, using realistic datasets. We aim to address several technical challenges mentioned above. To this end, we propose a *Feature Fusion and Knowledge-Distilled* (FFKD) method, which refers to a framework that integrates multi-modal data or features (fusion) and employs KD to transfer knowledge from a foundation (teacher) model to a simpler (student) model. Our contributions are twofold: (i) we provide a publicly accessible repository of implementation code,[1] and (ii) our technical contributions are detailed as follows:

- We formulate a posterior distribution-based optimization problem to rigorously characterize the MM-MTD task.
- We propose a multi-modal model combining a fusion method for diverse inputs and a distillation composite loss function to solve the formulated problem.
- The experimental results demonstrate that our approach effectively produces a lightweight model for the considered MTD scenario with heterogeneous inputs in challenging conditional environments.

[1]The code repository can be accessed at: https://github.com/TND-Lab/Feature-Fusion-Knowledge-Distilled-Multi-Modal-Multi-Target-Detection

- Specifically, our student model achieves approximately 95% of the teacher model's performance in terms of mAP index while offering approximately 50% faster inference time.

II. System Description

In the context of MTD, we consider a sophisticated sensing system designed to operate within a dynamic urban environment, tasked with the simultaneous detection and classification of multiple mobile targets, which is described as follows.

A. Multi-modal input-based MTD

The system integrates a hybrid sensing device comprising a thermal camera and an RGB camera, both co-located at coordinates (x_S, y_S, z_S). The thermal camera, characterized by a resolution of $W \times H$ pixels, captures heat emissions to produce a thermal image $I^{\text{thm}} \in \mathbb{R}^{W \times H}$, where each pixel's intensity corresponds to the thermal radiation of targets within the scene. Similarly, the RGB camera, with an identical or comparable resolution of $W \times H$ pixels, captures visible light to produce an RGB image $I^{\text{rgb}} \in \mathbb{R}^{W \times H \times 3}$, where each pixel encodes color information in three channels (red, green, blue).

The system processes a multi-modal input $(I^{\text{thm}}, I^{\text{rgb}})$, where features from the thermal and RGB images are combined to enhance detection and classification performance. The scene contains N targets (e.g., pedestrians, vehicles, traffic signs), each characterized by a 2D bounding box $[x_i, y_i, w_i, h_i]$ in the image plane, where $[x_i, y_i]$ denotes the top-left corner and w_i, h_i represent the width and height, respectively. Each target is assigned a categorical label $c_i \in \{1, 2, \ldots, C\}$, where C is the number of classes (e.g., $C = 3$ for categories such as person, car, and bike). The ground-truth annotations for a given frame are represented as

$$A = \{[x_i, y_i, w_i, h_i, c_i]\}_{i=1}^{N} \tag{1}$$

where N varies based on the number of targets present.

The system employs a deep learning model to process the multi-modal input and generate predicted annotations, formalized as

$$\hat{A} = f(I^{\text{thm}}, I^{\text{rgb}}; \theta) \tag{2}$$

where f is a parameterized model with weights θ, and the predicted annotations are

$$\hat{A} = \{[\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i, \hat{p}_i(c)]\}_{i=1}^{\hat{N}} \tag{3}$$

Here, $[\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i]$ encapsulates the predicted bounding box for the i-th target, $\hat{p}_i(c)$ represents the class probabilities across C classes, and $\hat{N}$ is the number of predicted targets. The model leverages a multi-modal feature engineering framework, as outlined in the contributions, to fuse thermal and RGB features, enhancing the accuracy of a lightweight model through a distillation training approach inspired by a foundation (teacher) model.

B. Probabilistic Characterization of MM-MTD

The multi-modal Multi-Target (MM)-MTD problem entails the simultaneous localization and categorization of multiple targets within a dynamic scene, observed by a multi-modal sensing system comprising a thermal camera and an RGB camera, as aforementioned.

Remark 1. *The considered MM-MTD is formulated as finding the most accurate approximation of the ground-truth posterior probability of the hypothesis (the set of annotations A) given the evidence (the thermal and RGB images, I^{thm} and I^{rgb}), i.e., $P(\hat{A} \mid I^{thm}, I^{rgb}) \approx P(A \mid I^{thm}, I^{rgb})$.*

The probabilistic formulation seeks the *posterior probability* $P(A \mid I^{\text{thm}}, I^{\text{rgb}})$, which quantifies the likelihood of the annotations A given the evidence provided by the thermal and RGB images. Using Bayes' theorem, the objective is to infer the set of annotations $A = \{[b_i, c_i]\}_{i=1}^{N}$, modeling the joint posterior distribution

$$P(A \mid I^{\text{thm}}, I^{\text{rgb}}) = \frac{P(I^{\text{thm}}, I^{\text{rgb}} \mid A) P(A)}{P(I^{\text{thm}}, I^{\text{rgb}})}, \tag{4}$$

where $P(I^{\text{thm}}, I^{\text{rgb}} \mid A)$ is the *likelihood*, modeling the probability of observing the images given the annotations; $P(A)$ is the *prior*, capturing prior knowledge about the annotations (e.g., number of targets, bounding box distributions, class frequencies); and $P(I^{\text{thm}}, I^{\text{rgb}})$ is the *evidence*, a normalizing constant ensuring the posterior is a valid probability distribution.

The objective of the detection problem is to find the most likely set of annotations A, typically by maximizing the posterior probability, i.e., Maximum A Posteriori (MAP) estimation, or by characterizing the full posterior distribution for probabilistic inference.

1) Joint Likelihood Probability of Multi-Modal Inputs: Assuming conditional independence between thermal and RGB observations given the annotations, the likelihood factorizes as

$$P(I^{\text{thm}}, I^{\text{rgb}} \mid A) = P(I^{\text{thm}} \mid A) P(I^{\text{rgb}} \mid A). \tag{5}$$

The likelihood for each modality models pixel values as conditionally independent given the target annotations as

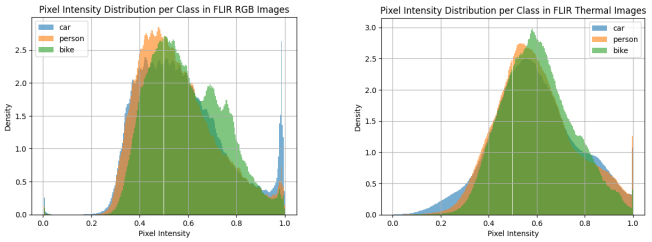
$$P(I^{\text{img}} \mid A) = \prod_{k \in \Omega} P(I^{\text{img}}(k) \mid A), \tag{6}$$

where $\text{img} \in \{\text{thm}, \text{rgb}\}$ denotes the modality, Ω is the set of pixels, k indexes a pixel at (u_k, v_k), $I^{\text{thm}}(k)$ represents the thermal intensity, and $I^{\text{rgb}}(k) \in \mathbb{R}^3$ represents the RGB color vector at pixel k. It is noted that $P(I^{\text{thm}} \mid A)$ and $P(I^{\text{rgb}} \mid A)$ can be numerically determined, e.g., using the KDE method, as illustrated in Fig. 1 for the considered FLIR dataset [8].

2) Prior Distribution of the Desired MTD Output: The prior distribution $P(A)$ models the number of targets, their bounding boxes, and class labels as

$$P(A) = P(N) \prod_{i=1}^{N} P(b_i) P(c_i). \tag{7}$$

where N is the number of targets N; bounding boxes are uniformly distributed over the image plane as $P(b_i) = \frac{1}{WHW_{\max}H_{\max}}$, where $W_{\max}, H_{\max}$ are maximum dimensions; class labels follow a categorical distribution as $P(c_i) = \pi_{c_i}$, where π_{c_i} reflects the expected class frequency, as illustrated in Fig. 2.



(a) RGB Images (b) Thermal Images

Fig. 1. Pixel intensity distribution per class in (a) RGB images and (b) thermal images in the utilized dataset [8].

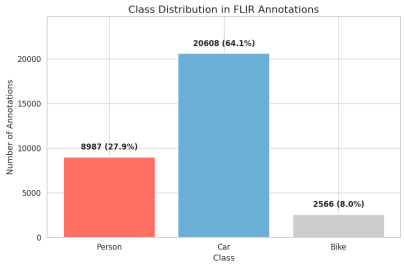


Fig. 2. Class distribution in the utilized dataset [8].

C. Optimization Problem Formulation for FFKD in MM-MTD

During distillation training, the student model learns to imitate the teacher's predictions by utilizing the teacher's soft labels, which encode richer inter-class relationships than one-hot ground-truth labels [4]. Additionally, to capture contextual knowledge, the student model replicates the intermediate feature maps of the teacher model. The optimization training objective combines two loss components a KD loss [4], which measures the difference between the student and teacher outputs, and a feature distillation (FD) loss [9], which aligns the student's internal representations with those of the teacher.

With a dataset $\mathcal{D} = \{(I^{thm,(k)}, I^{rgb,(k)}, A^{(k)})\}_{k=1}^{K}$, where $A^{(k)} = \{[b_i^{(k)}, c_i^{(k)}]\}_{i=1}^{N^{(k)}}$ are the ground-truth annotations for the k-th sample, the MM-MTD problem involves inferring a set of annotations $\hat{A}^{(k)} = \{[\hat{b}_i^{(k)}, \hat{c}_i^{(k)}]\}_{i=1}^{N^{(k)}}$, where $\hat{b}_i^{(k)} = [\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i]$ represents a bounding box and $\hat{c}_i^{(k)} \in \{1, \ldots, C\}$ is a class label, given thermal images $I^{thm,(k)} \in \mathbb{R}^{W \times H}$ and RGB images $I^{rgb,(k)} \in \mathbb{R}^{W \times H \times 3}$. Distillation training aims to train a smaller student model parameterized by θ_S that could closely mimic L level transformed feature maps $F_S^{(L)}$ with those $F_T^{(L)}$ of the teacher model, and to approximate the probabilistic outputs of a larger teacher model, leveraging soft class probabilities and a set of bounding boxes incorporating ground-truth annotations.

1) Knowledge Distillation Loss: The distillation loss encourages the student to mimic the teacher's probabilistic outputs [4], comprising class probability and bounding box components.

a) Class Probability Distillation: For each target i, the teacher provides a softened class probability distribution [4] using a temperature $\tau > 1$, which can be expressed as

$$p_{T,i}^{(k)}(\hat{c}_{T,i}^{(k)} = j | I^{thm,(k)}, I^{rgb,(k)}; \tau) = \frac{\exp(z_{T,j}^{(k)}(i)/\tau)}{\sum_{j'=1}^{C} \exp(z_{T,j'}^{(k)}(i)/\tau)}, \tag{8}$$

where $z_{T,j}^{(k)}(i)$ is the teacher's logit for class j of k-th sample. The student's probabilities similarly can be expressed as

$$p_{S,i}^{(k)}(\hat{c}_{S,i}^{(k)} = j | I^{thm,(k)}, I^{rgb,(k)}; \theta_S, \tau) = \frac{\exp(z_{S,j}^{(k)}(i;\theta_S)/\tau)}{\sum_{j'=1}^{C} \exp(z_{S,j'}^{(k)}(i;\theta_S)/\tau)}, \tag{9}$$

The class distillation loss is the Kullback-Leibler (KL) divergence between the softened class probability distributions predicted by the teacher and student models is expressed as

$$\mathcal{L}_{\text{class-distill}}^{(k)} = \tau^2 \sum_{i=1}^{N^{(k)}} \text{KL}(p_{T,i}^{(k)}(\hat{c}_{T,i}^{(k)}|\cdot;\tau) || p_{S,i}^{(k)}(\hat{c}_{S,i}^{(k)}|\cdot;\theta_S,\tau)), \tag{10}$$

where the KL divergence is formulated as

$$\text{KL}(p_{T,i}^{(k)} || p_{S,i}^{(k)}) = \sum_{j=1}^{C} p_{T,i}^{(k)}(\hat{c}_{T,i} = j) \log \left(\frac{p_{T,i}^{(k)}(\hat{c}_{T,i}=j)}{p_{S,i}^{(k)}(\hat{c}_{S,i}=j)} \right). \tag{11}$$

b) Bounding Box Distillation: For k-th sample, the model predicts $\hat{B}_i^{(k)} = \{\hat{b}_i^{(k)}\}_{i=1}^{N^{(k)}}$ The bounding box distillation loss is of the FFKD-MM-MTD is expressed as

$$\mathcal{L}_{\text{box-distill}}^{(k)} = \sum_{i=1}^{N^{(k)}} \text{smooth}_{L1}(\hat{B}_{T,i}^{(k)}, \hat{B}_{S,i}^{(k)}), \tag{12}$$

where $\hat{B}_{T,i}^{(k)}$ and $\hat{B}_{S,i}^{(k)}$ is the teacher and student predicted boxes respectively and

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x| - 0.5 & \text{otherwise.} \end{cases} \tag{13}$$

Thus, the KD loss of our problem is formulated as

$$\mathcal{L}_{\text{knowledge-distill}}^{(k)} = \lambda_{\text{cls}} \mathcal{L}_{\text{class-distill}}^{(k)} + \lambda_{\text{reg}} \mathcal{L}_{\text{box-distill}}^{(k)}, \tag{14}$$

where $\lambda_{\text{clss}}, \lambda_{\text{reg}} \in [0, 1]$ balances classification and regression components, and are also design optimization parameters.

2) Ground-Truth Loss: The ground-truth loss measures the student's error relative to true annotations. The classification loss is characterized as a cross-entropy loss as

$$\mathcal{L}_{\text{class-CE}}^{(k)} = -\sum_{i=1}^{N^{(k)}} \sum_{j=1}^{C} \mathbf{1}\{\hat{c}_{S,i}^{(k)} = j\} \times \log p_{S,i}(\hat{c}_{S,i}^{(k)} = j | I^{thm,(k)}, I^{rgb,(k)}; \theta_S, \tau = 1). \tag{15}$$

The bounding box loss is characterized as a smooth $L1$ loss for bounding box predictions, which is expressed as

$$\mathcal{L}_{\text{box-reg}}^{(k)} = \sum_{i=1}^{N^{(k)}} \text{smooth}_{L1}(\hat{b}_{S,i}^{(k)} - \mu_S(i;\theta_S)), \tag{16}$$

where the smooth function is defined in (13). Thus, the total ground-truth loss of the FFKD-MM-MTD is formulated as

$$\mathcal{L}^{(k)}_{\text{ground-truth}} = \gamma \mathcal{L}^{(k)}_{\text{class-CE}} + (1-\gamma)\mathcal{L}^{(k)}_{\text{box-reg}}, \tag{17}$$

where $\gamma \in [0,1]$ balances classification and regression.

3) Feature Distillation Loss: The objective is to have the student model replicate the intermediate representations produced by the teacher model [9] as

$$\mathcal{L}_{\text{feature-distill}} = d(F_S^{(L)}(\theta_S), F_T^{(L)}(\theta_T)) \tag{18}$$

where d denotes a distance function of transform features from L feature map levels. The optimization problem is to learn student parameters θ_S that minimize the distance to the teacher parameters θ_T.

The considered FFKD-MM-MTD problem is characterized via optimizing the following problem

$$\underset{\theta_S,\alpha,\beta,\gamma}{\text{minimize}} \quad \Big(\alpha\mathcal{L}_{\text{feature-distill}} + \beta\mathcal{L}_{\text{ground-truth}} + \gamma\mathcal{L}_{\text{knowledge-distill}}\Big) \tag{19a}$$

$$\text{subject to} \quad p_{S,i}(\hat{c}_{S,i}=j) \geq 0, \sum_{j=1}^{C} p_{S,i}(\hat{c}_{S,i}=j) = 1, \forall i \tag{19b}$$

$$\mu_S(i;\theta_S) \in [0,W]\times[0,H]\times[0,W_{\max}]\times[0,H_{\max}], \forall i, \tag{19c}$$

where constraint (19b) is for class probabilities and constraint (19c) is for Bounding box validity.

III. Proposed FFKD-based Training Pipeline for MM-MTD

To address the FFKD-MM-MTD optimization problem (19), we propose a training algorithm and pipeline that leverages KD to train a lightweight student model for efficient and accurate detection and tracking, as illustrated in Fig 3. The pipeline processes paired RGB and thermal inputs, employs Bi-Directional Feature Pyramid Networks (BiFPN) [2] for feature enhancement, ensuring performance on resource-constrained devices. The fused features across modalities by Convolutional Block Attention Module (CBAM) [10], and generate detections through classification and regression heads.

A. Mathematical Description of the Neural Network Architectures for Teacher and Student Models

To support the Distillation Training-Based pipeline outlined in Section III, we apply a teacher model based on EfficientDet-D1 [2] and a student model based on MobileNetV3 as the backbone for feature extraction on multi-scale [3].

1) Teacher Model - EfficientDet-D1: With the EfficientNet-B1 backbone which uses a method called compound scaling (uniform scalability of the resolution, depth, and width), the teacher model [2] has parameters θ_{tc}, is a high capacity architecture designed for precise detection.

2) Student Model - MobileNetV3: By utilizing hardware-aware network architecture search (NAS) and the NetAdapt algorithm in a complementary way, MobileNetV3 [3] is introduced as a CNNs optimized for mobile CPUs and embedded devices. It introduces novel architecture with parameters θ_{st}, including hard swish activation [11] and squeeze-and-excitation (SE) modules [12] in MBConv blocks, to improve performance.

B. Input Description and Processing

The pipeline operates on a dataset $\mathcal{D} = \{(I_t, Y_t)\}_{t=1}^{T}$. Both models process an input at time t is $I_t = \{I_t^{\text{thm}}, I_t^{\text{rgb}}\}$, where $I_t^{\text{rgb}} \in \mathbb{R}^{H\times W\times 3}$ and $I_t^{\text{thm}} \in \mathbb{R}^{H\times W\times 1}$ by each backbone modality independently and extract multi-scale features. For each modality $m \in \{\text{rgb}, \text{thm}\}$:

$$F_m = f_{\text{backbone}}(I_t^m; \theta_{\text{backbone}}), \tag{20}$$

where f_{backbone} comprises convolutions blocks, which can be expressed as a composite function as follows

$$f_{\text{backbone}} = f_{\text{block}_M} \circ \cdots \circ f_{\text{block}_1}, \tag{21}$$

with each f_{block_m} including depthwise separable convolution, point convolution, batch normalization, and swish-based activation [11]. The output is $F_m = \{F_m^l\}_{l=1}^{L}$.

These multi-level features are processed through BiFPN modules [2] with learnable attention weights, which refine and aggregate features across levels iteratively, enhancing cross-scale interactions to improve detection across varying target sizes. The forward propagation of BiFPN is expressed as

$$\tilde{F}_m = \text{BiFPN}(F_m). \tag{22}$$

C. Feature Engineering via Fusion Module

The refined features are fused across modalities along channel and spatial dimension using a fusion function called CBAM [10], which produces fused feature maps and is parameterized by θ_{fuse}. Recall that $\tilde{F}_m \in \{\tilde{F}_{\text{rgb}}, \tilde{F}_{\text{thm}}\}$, the data fusion is modeled as

$$F_{\text{fused}} = f_{\text{fuse}}(\tilde{F}_m; \theta_{\text{fuse}}), \tag{23}$$

D. Final Detection MTD Output

The fused output features F_{fused} are processed by classification and regression heads to produce object bounding boxes and corresponding classes and scores, which are expressed as

$$z_{\text{cls}} = f_{\text{cls}}(F_{\text{fused}}; \theta_{\text{cls}}), \quad z_{\text{reg}} = f_{\text{reg}}(F_{\text{fused}}; \theta_{\text{reg}}), \tag{24}$$

where f_{cls},f_{reg} are convolutional layers producing classification logits $z_{\text{cls}} \in \mathbb{R}^{N\times C}$ (for N anchors and C classes) and bounding box offsets z_{reg}. The detection output in (3) can be reformulated as

$$D = f_{\text{decode}}(z_{\text{cls}}, z_{\text{reg}}) = \{(\hat{b}_{t,i}, \hat{y}_{t,i})\}. \tag{25}$$

The comprehensive model of the forward propagation of the FFKD-MM-MTD is mathematically expressed as

$$D = (f_{\text{decode}} \circ (f_{\text{cls}} \parallel f_{\text{reg}})) \circ f_{\text{fuse}} \circ (\text{BiFPN} \circ f_{\text{backbone}})^{\text{rgb,thm}}(I_t; \theta)$$

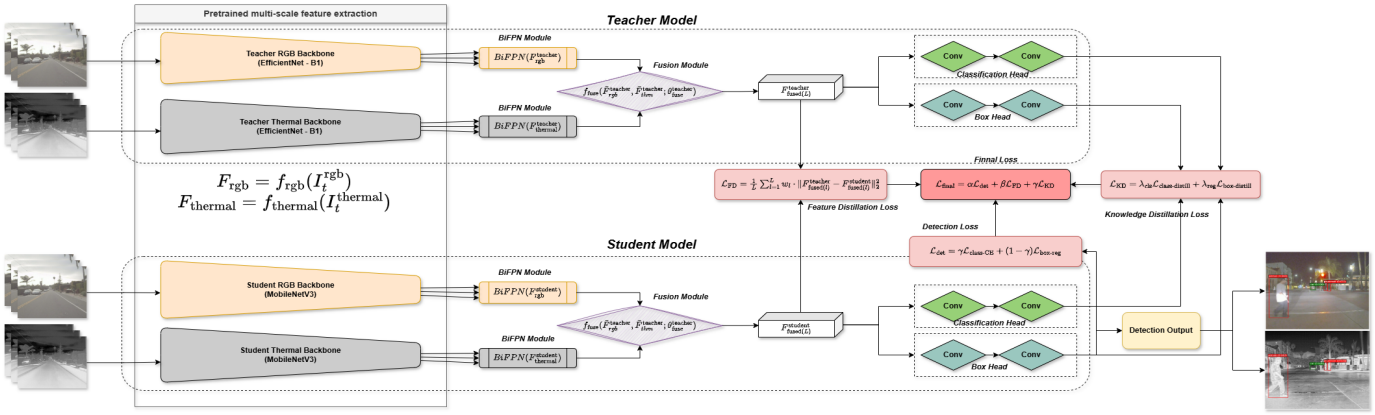


Fig. 3. Proposed FFKD-based training pipeline for the considered MM-MTD problem.

E. The Proposed FFKD-MM-MTD Loss Function Design

To address the objective function (19a), our training loss combines the actual detection loss of the student model and the transfer loss components which includes the FD loss and the KD loss as

$$\mathcal{L}_{\text{final}} = \alpha\mathcal{L}_{\text{det}} + \beta\mathcal{L}_{\text{FD}} + \gamma\mathcal{L}_{\text{KD}}, \tag{26}$$

where hyperparameters α, β, γ balance the contribution of each loss terms and $\mathcal{L}_{\text{FD}}$ is applied for the fusion feature maps mentioned in Eq. (23).

The training algorithm minimizes $\mathcal{L}_{\text{final}}$ to optimize θ_{st} of the student model, aligning with the distillation optimization problem in Section II-C. Initialized with pre-trained weights, the student parameters are updated over epochs by iterating through $\mathcal{D}$, computing teacher and student outputs, calculating $\mathcal{L}_{\text{final}}$, and applying backpropagation update as

$$\theta_{\text{st}} \leftarrow \theta_{\text{st}} - \eta\nabla_{\theta_{\text{st}}}\mathcal{L}_{\text{final}}, \tag{27}$$

using an optimizer (e.g., Adam) with learning rate η.

F. Inference of the Distillation-Trained Model on New Observations

Without relying on the teacher, the trained student model takes a new pair of input images and predicts bounding boxes and class labels directly during inference phase. The model outputs the final detections for both RGB and thermal images by applying confidence thresholding and non-maximum suppression techniques.

IV. Results and Discussions

A. Dataset

The Teledyne FLIR Thermal Dataset [13] provides 26,442 fully annotated thermal and visible spectrum frames to advance multi-target detection and classification for Advanced Driver Assistance Systems (ADAS) and autonomous vehicles. We use the FLIR Aligned Dataset [8] is a refined version of [13] that contains 5,142 paired of RGB and thermal images and includes bounding-box annotations for three classes, namely person, car, and bike. The dataset is divided into 4,129 training pairs (for training and validation set) and 1,013 testing pairs.

B. Tranining Setup

TABLE I
Parameter statistic of fusion-based multi-modal models (in millions).

Type	Backbone	BiFPN	Head	Total	Trainable
Teacher	12.2M	0.78M	0.064M	13.2M	1M
Student	3.5M	0.95M	0.064M	4.8M	1.24M

To implement our proposed pipeline, we selected EfficientDet-D1 as the teacher model due to its robust detection performance and balanced accuracy–efficiency trade-off, and MobileNetV3 Large 0.75, a lightweight and computationally efficient architecture, as the student model. Table I presents the statistics of both models. To ensure stable and optimal performance during the training, validation, and testing phases of the proposed method, we conducted a series of experiments to determine appropriate hyperparameters, as shown in our source code.

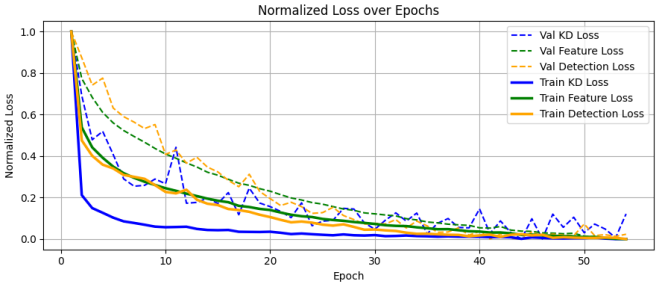


Fig. 4. Training and validation loss of each loss component in (26).

C. Inference and Demonstrations

Figure 4 shows the convergence of the training process after 50 epochs. To evaluate the effectiveness of our proposed method, we conducted experiments on the test set using mAP at various IoU thresholds. The performance of both

(a)

(b)

Fig. 5. Predicted Images from (a) teacher model and (b) student model. Please refer to the code repository for further demonstrations.

teacher and student models under different modality settings is summarized in Tables II and III.

TABLE II
PERFORMANCE ON TEST SET OF MODEL WITH MAP WHERE T- PREFIX IS TEACHER TYPE MODEL AND S- PREFIX IS STUDENT TYPE MODEL

Model	mAP@0.5:0.95	mAP@0.5	mAP@0.75
T-RGB Only	24.4	56.2	17.2
T-Thermal Only	30.7	65.2	23.7
T-Fusion	**33.0**	**69.1**	**26.2**
S-RGB Only	22.4	53.0	15.7
S-Thermal Only	29.0	61.4	23.2
S-Fusion	27.9	58.7	22.4
S-Distillation	**31.5**	**64.8**	**24.7**

The table result shows the impact of cross-modal fusion and the effectiveness of knowledge transfer in enhancing significantly smaller student model's performance compared to the single one. Moreover, the S-Distillation model achieves a remarkably high performance, with an index of mAP@0.5:0.95 at 31.5%, which is comparable to the teacher fusion model (T-Fusion) at 33.0%.

TABLE III
PERFORMANCE ON TEST SET OF MODEL WITH INFERENCE TIME

Model	Batch Size	Inference Speed (s)
T-Fusion	32	0.041
S-Distillation	32	0.023

The S-Distillation model's inference speed is computational efficiency achieving a significantly faster of 0.023 seconds per image and is approximately 50% of the inference speed of the T-Fusion model. A demonstration of the qualitative predicted images from both the fusion-based teacher model and the distilled student model is provided in Fig. 5. in which there are no significant differences in detection quality.

V. CONCLUSION

We address the problem of multi-modal multi-target detection and classification (MM-MTD) in autonomous driving using the realistic FLIR dataset. We propose a novel training pipeline that integrates fusion-based multi-modal modeling with knowledge distillation to develop an efficient and compact model for MTD. Our solution, termed as FFKD-MM-MTD pipeline, introduces a principled optimization formulation and a tailored composite loss function for the detection task. Experimental results demonstrate that our student model achieves approximately 95% of the teacher model's mAP performance while reducing inference time by approximately 50%, facilitating deployment on resource-constrained edge devices.

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[2] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 10 781–10 790.

[3] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 1314–1324.

[4] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] M. Bijelic, T. Gruber, F. Mannan, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 682–11 692.

[6] A. Chawla, P. Chattopadhyay, P. Goyal, A. Chatterjee, and A. Chakraborty, "Data-free knowledge distillation for object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1853–1862.

[7] J. Wang, Z. Dong, R. Dong, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Yuan, "Crosskd: Cross-head knowledge distillation for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[8] H. Zhang, E. Fromont, S. Lefèvre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 276–280.

[9] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1921–1930.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2018, pp. 3–19.

[11] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," in *International Conference on Learning Representations (ICLR), Workshop Track*, 2018.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *arXiv e-prints*, 2017. [Online]. Available: https://arxiv.org/abs/1709.01507

[13] *Thermal Dataset for Algorithm Training*, Teledyne FLIR, 2025. [Online]. Available: https://www.flir.ca/oem/adas/adas-dataset-form/