CrossFusionKD: A Cross-Modal Knowledge Distillation Framework for UxV Object Detection

Syed Murtaza Hussain Abidi, [a], Syed Muhammad Raza, [a], Md Masuduzzaman, [b] and Soo Young Shin, [a,*]

[a] Department of IT Convergence Engineering, Kumoh National Institute of Technology, 61 Daehak-ro, Geoui-dong, Gumi-si, 39177, Gyeongsangbuk-do, South Korea
[b] DeGroote School of Business, McMaster University, 1280 Main St W, Hamilton, L8S 4L8, Ontario, Canada

ARTICLE INFO

Keywords:
Unmanned vehicles
Knowledge Distillation
Cross-Modality
Data fusion

ABSTRACT

Unmanned vehicles (UxVs) require reliable object detection across heterogeneous imaging modalities, particularly RGB and infrared (IR). To address this challenge, we propose CrossFusion Knowledge Distillation (CrossFusionKD), a framework that improves detection performance while reducing computational complexity through model compression and cross-modal distillation. A teacher network trained on paired RGB and IR data transfers complementary multi-modal representations to a lightweight RGB-only student via channel-wise distillation, allowing deployment without IR sensing at test time. Evaluation on established benchmark datasets confirms the effectiveness of the proposed approach, achieving 96.7% mAP on LLVIP (RGB-only), 98.3% on fused RGB-IR data, 92.1% on KAIST Multispectral, and 92.8% on STUJ-SMOD. The proposed method consistently exceeds previously reported results from RGB- and IR-based approaches. Overall, CrossFusionKD provides an effective trade-off between detection quality and computational cost, making it suitable for real-time deployment on resource-limited UxV platforms.

1. Introduction

Unmanned vehicles (UxVs), including UGVs, UAVs, and UUVs, are increasingly used in mission-critical applications requiring efficient and reliable operation. These platforms may be teleoperated or deployed autonomously through integrated navigation and planning frameworks. Their adoption has expanded across domains such as disaster response [1, 2], UxV-swarm-based fruit disease segmentation [3], pedestrian detection [4], and IoT-driven systems [5, 6, 7], owing to their mobility, rapid deployment, and relatively low operational cost. At the same time, advances in AI and DL have substantially improved UxV perception, signal understanding, and autonomous decision-making. In complex dynamic environments, however, reliable operation increasingly depends on multi-sensor perception capable of providing richer scene representations.

As the integration of multiple sensors has continued to advance and real-world tasks have become more complex, research has shifted from developing uni-modal object detectors to designing multi-modal object detection frameworks. With their pixel-based representations, RGB images have enabled early object detectors to achieve accurate classification and localization in well-lit environments. However, in challenging conditions such as poor lighting, low visibility, or complete darkness, RGB-based detectors have experienced significant performance degradation. To address these limitations, incorporating complementary sensing modalities, particularly infrared (IR) imagery, has become an effective strategy for improving detection robustness. Prior studies, including [8] and [9], have shown that RGB–IR fusion frameworks can improve object detection performance under challenging environmental conditions. Nevertheless, despite these gains, several practical obstacles still restrict the deployment of such multi-modal systems on UxV platforms.

First, high-quality IR cameras have remained significantly expensive than RGB cameras, making it impractical to equip all UxVs with multi-modal sensing capabilities. Second, the complex architectures of multi-modal object detectors have required computationally complex processes such as feature extraction, feature fusion, and data synchronization, often resulting in models with millions of parameters (typically exceeding 85 million). These computational demands have posed a significant challenge for real-time deployment on computationally constrained UxV platforms. Researchers address these challenges by exploring deep learning-based model compression strategies, particularly pruning [10], quantization [11], and knowledge distillation (KD) [12].

Model compression techniques have become an important approach for maintaining strong detection performance under limited computational budgets. Among them, knowledge distillation (KD) has received considerable attention because it enables a compact student network to inherit informative representations from a more expressive teacher network, thereby improving efficiency while preserving predictive performance without requiring architectural modification of the student. KD has been adopted across a wide

*Corresponding Author: Soo Young Shin (wdragon@kumoh.ac.kr)
murtazaabidi@kumoh.ac.kr (S.M.H.A.); iamraza1998@gmail.com (S.M.R.); masud.prince@mcmaster.ca (M.M.); wdragon@kumoh.ac.kr (S.Y.S.)
ORCID(s): 0009-0005-3895-5237 (S.M.H.A.); 0000-0002-3997-4924 (S.M.R.); 0000-0001-8039-3673 (M.M.); 0000-0002-2526-2395 (S.Y.S.)

range of computer vision tasks, including image classification, object detection, and instance segmentation, and its transfer mechanisms have progressed from output-level supervision to the distillation of intermediate feature representations [13]. However, traditional KD has assumed that the teacher and student operate within the same modality. This assumption does not hold in cross-modal scenarios where the teacher model processes RGB and IR data, whereas the student model relies solely on RGB. This modality gap has often led to the transfer of redundant or misaligned features, degrading the student model's performance [14].

To address these challenges, a framework, **CrossFusionKD**, has been proposed as a cross-modal knowledge distillation approach tailored for resource-constrained UxV platforms. Unlike conventional multi-modal object detectors that require IR sensors during inference, CrossFusionKD has enabled an RGB-only student detector to achieve competitive detection performance by distilling multi-modal knowledge from an RGB-IR teacher model. Instead of performing direct RGB–IR feature fusion, CrossFusionKD factorizes multi-modal representations into channel-wise components, which facilitates the learning of modality-invariant representations by the student network. By carefully aligning RGB and IR feature representations, CrossFusionKD has mitigated the modality gap, ensuring that the transferred knowledge remains relevant and beneficial to the student model. Furthermore, unlike MTKD, which relies on multiple teachers, CrossFusionKD has employed a single-teacher approach, substantially lowering computational cost while supporting real-time deployment on UxVs. The key contributions of the proposed framework are summarized as follows:

- **CrossFusionKD** framework has been proposed as a knowledge distillation-based object detection approach designed for efficient deployment on UxVs.
- The framework has applied **channel-wise distillation**, enabling an RGB-only student detector to learn from multi-modal RGB-IR teacher models while mitigating the modality gap.
- Real-world applicability has been validated by deploying CrossFusionKD on **UAV and UGV** platforms, confirming its efficiency in edge computing environments.

2. RELATED WORK

2.1. Knowledge Distillation

Among model compression strategies, pruning, quantization, and knowledge distillation (KD) are the most widely adopted. In KD, a compact student network is optimized under the guidance of a teacher network by matching softened output distributions generated through temperature-scaled softmax, thereby improving the trade-off between predictive performance and inference efficiency without modifying the student architecture [15]. The form of transferred knowledge has progressed from output-level supervision to the distillation of richer intermediate feature representations. Nevertheless, most existing KD frameworks are developed for single-source settings, in which both the teacher and student models operate on the same input modality. Knowledge distillation (KD), as reviewed in [16], aims to transfer informative supervision from a high-capacity teacher network to a compact student network while avoiding a substantial increase in model parameters. Early KD approaches, including those discussed in [16] and [17], mainly concentrated on distilling predictive information from the final classification outputs of the teacher, which largely restricted their use to image classification problems. Subsequent studies have expanded this paradigm by exploiting information encoded in intermediate feature layers of the teacher network [18, 19, 20], thereby extending KD to more challenging vision tasks such as object detection [21, 22, 23, 24, 25] and semantic segmentation [26, 27]. For example, FitNet [28] introduced the use of intermediate teacher representations to better guide student optimization and improve final predictive performance. Building on this direction, Qin et al. [29] and Li et al. [30] further incorporated attention mechanisms to strengthen feature-level knowledge transfer. In the context of object detection, Wang et al. [31] proposed CrossKD, which transfers intermediate student features to the teacher detection head in order to reduce target inconsistency and improve detection quality. Similarly, X3KD [32] introduced a distillation framework that combines cross-modal, cross-task, and multi-stage supervision to improve multi-camera 3D object detection by exploiting knowledge from LiDAR-based detectors and instance segmentation teachers. In addition, Localization Distillation (LD) [33] was developed to improve object detection by transferring localization cues through bounding box distributions, demonstrating stronger effectiveness than conventional feature imitation strategies.

2.2. Cross-Modal Distillation

Cross-modal knowledge distillation (CKD) has gained increasing attention as a means of improving single-modal models through supervision derived from multimodal representations. In this setting, a teacher network trained with multi-source inputs guides a lighter student network that operates using only a single modality, allowing the student to approach the performance of the multimodal teacher while retaining a simpler inference pipeline. This paradigm has been adopted in a variety of application domains. For example, DisCoM-KD [34] combines disentangled representation learning with adversarial domain adaptation to extract domain-invariant features and transfer them across heterogeneous modalities. By explicitly reducing discrepancies between training and testing modalities, this framework improves the robustness of single-modal classifiers. In the domain of autonomous aerial vehicles (AAVs), CMDistill [35] was proposed to reconcile the high complexity of multimodal detectors with the limited representational

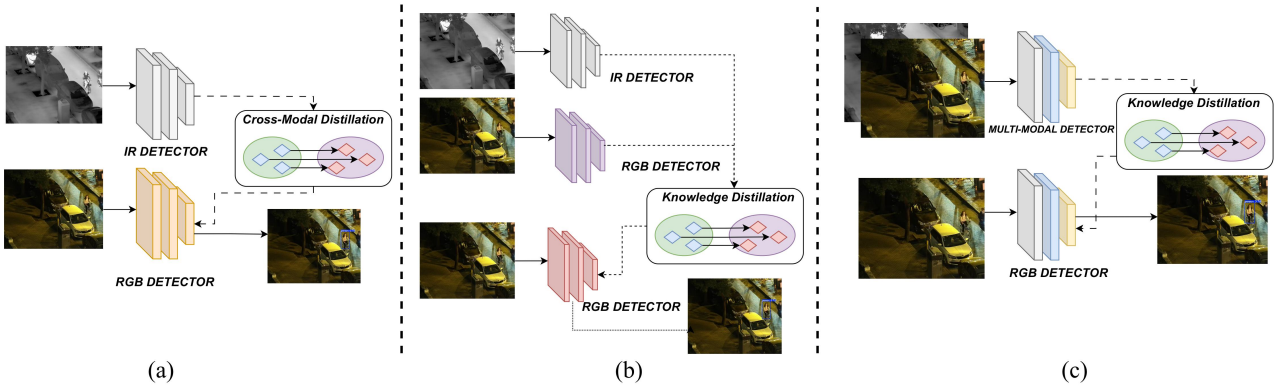


Figure 1: Comparison of representative object detection schemes. (a) **Cross-modal detector**: paired IR and RGB images are exploited during training, while deployment relies solely on RGB input. (b) **Multi-teacher detector**: multiple teacher networks, each trained on distinct modalities or tasks, jointly supervise a single student detector to enhance performance. (c) **CrossFusionKD** (proposed): a single-teacher distillation framework in which fused IR-RGB information is utilized during training to distill informative representations into an RGB-only student detector for efficient inference.

capacity of unimodal models. Its distillation objective consists of three complementary components: intermediate-layer transfer based on Pearson correlation, affinity-matrix matching over high-level semantic features to preserve relational structure, and the transfer of localization and classification cues through teacher prediction outputs. Through these mechanisms, CMDistill narrows the representation gap between modalities and improves the effectiveness of CKD for aerial perception tasks. Nevertheless, many existing CKD approaches still insufficiently address the intrinsic discrepancy between modalities. In object detection scenarios, for instance, teacher and student networks often receive heterogeneous inputs such as infrared (IR) and RGB imagery. RGB images typically provide fine-grained texture and appearance information, whereas IR images emphasize thermal signatures, resulting in substantially different feature characteristics. Effectively transferring informative knowledge across such heterogeneous representations therefore remains a central challenge in improving CKD performance [36].

2.3. Knowledge Distillation with Multiple Teachers

Multi-teacher knowledge distillation (MTKD) extends the conventional distillation paradigm by allowing a student network to learn from multiple teacher networks rather than a single teacher. By aggregating complementary supervisory signals, MTKD can enrich the representational capacity of the student and is particularly suitable for settings in which a single teacher is unable to capture the diversity of knowledge spanning multiple domains or tasks. In such frameworks, each teacher typically specializes in a particular modality, domain, or data distribution, such as RGB or infrared imagery, and its outputs are used to provide supervisory guidance or pseudo-labels for student training. Despite these advantages, MTKD introduces several practical challenges, including the effective integration of heterogeneous knowledge sources and the appropriate balancing of individual teacher contributions, both of which may otherwise lead to noisy supervision or suboptimal optimization. In addition, the computational burden associated with maintaining multiple teacher networks can restrict the feasibility of MTKD in real-time applications [37, 38, 39, 40].

3. METHODOLOGY

3.1. CrossFusionKD

The proposed framework enables cross-modal knowledge distillation by transferring complementary multimodal information to a lightweight student model. In particular, the teacher network takes paired RGB and infrared (IR) images as input and fuses them to learn enriched feature representations for object detection. By distilling this fused knowledge, the student model acquires more informative supervision, which contributes to improved detection performance. Additionally, the single-teacher setup simplifies training while ensuring that multimodal information is distilled into a lightweight RGB-based detector, as shown in Fig.1. This framework benefits edge devices and real-world applications with limited computational resources, such as Unmanned Vehicles (UxVs).

3.2. Overview of System Model

Figure 2 presents the overall architecture of the proposed framework. It consists of two object detectors: an RGB–IR detector and an RGB-only detector. The RGB–IR detector receives fused RGB and infrared (IR) images as input, while the RGB-only detector processes RGB images alone. The latter generates object detection outputs comprising bounding boxes and their associated confidence scores.

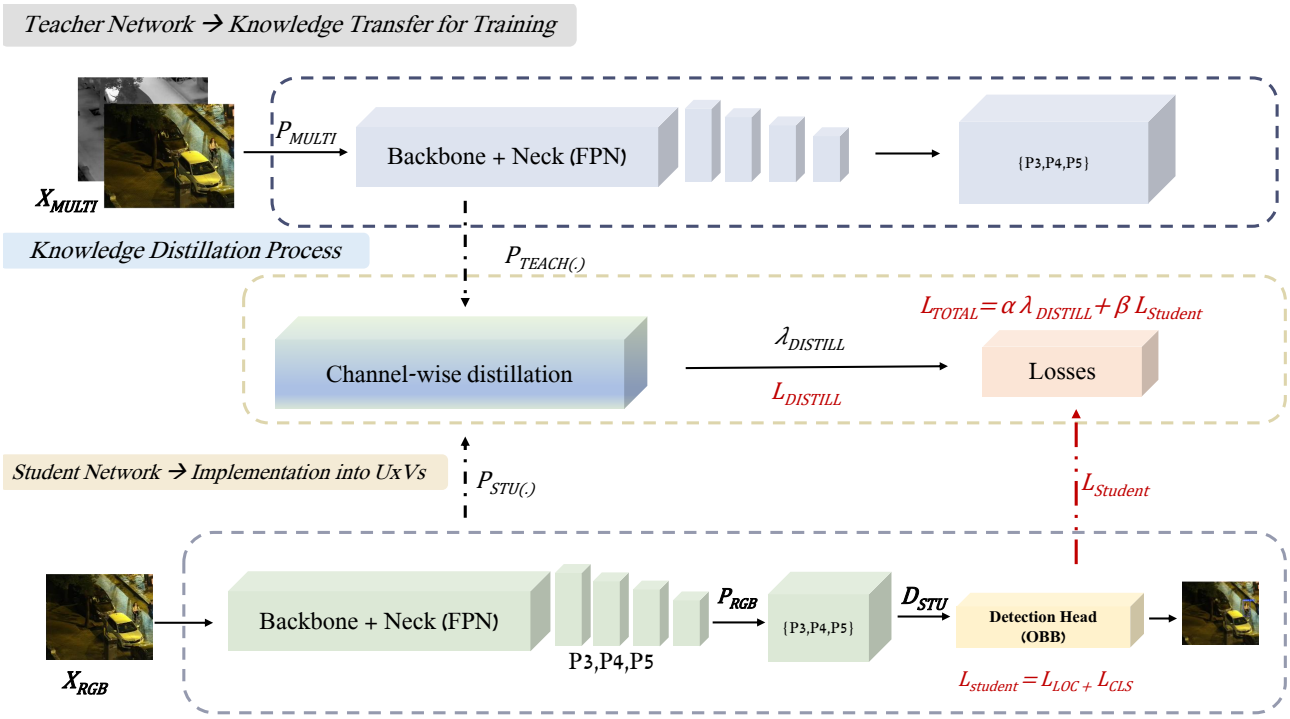


Figure 2: Overview of CrossFusionKD Object detection framework.

Table 1
Notations and Descriptions of the Framework

Notations	Descriptions
$\mathbf{X}_{\mathbf{MULTI}}$	Teacher network input images
$\mathbf{P}_{\mathbf{MULTI}}$	Teacher backbone feature maps
$\mathbf{P}_{\mathrm{TEACH}}(\cdot)$	Teacher features for alignment
$\mathbf{X}_{\mathbf{RGB}}$	Student network input images
$\mathbf{P}_{\mathbf{RGB}}$	Student backbone feature maps
$\mathbf{P}_{\mathrm{STU}}(\cdot)$	Student features for alignment
D_{STU}	Student detection predictions
$\mathbf{L}_{\mathbf{LOC}}$	Localization loss (CIOU + DFL)
$\mathbf{L}_{\mathbf{CLS}}$	Classification loss (BCE)
$\mathbf{L}_{\mathbf{STU}}$	Student prediction loss
$\mathbf{L}_{\mathbf{DISTILL}}$	Feature distillation loss
$\lambda_{\mathrm{DISTILL}}$	Weight for distillation loss
$\mathbf{L}_{\mathbf{TOTAL}}$	Combined total loss

Note: Bold symbols represent values structured as matrices or vectors.

3.2.1. Backbone Network

The backbone network serves as the feature extraction module that transforms an input image, denoted by $\mathbf{X}_{\mathrm{MULTI}}$ or $\mathbf{X}_{\mathrm{RGB}}$, into hierarchical multi-scale feature representations. In the proposed framework, ResNet50 [41] is adopted as the backbone encoder to capture discriminative visual features. Common backbone architectures used in object detection include fully convolutional networks such as ResNet [41], VGG [42], DenseNet [43], MobileNet [44], and EfficientNet [45]. These architectures provide informative representations for challenging vision tasks and are widely employed as feature encoders in detection pipelines. In many object detection frameworks, the backbone is further combined with modules such as the feature pyramid network (FPN) [46] to construct multi-level feature maps. Accordingly, the output features of the RGB–IR and RGB branches are denoted by $\mathbf{P}_{\mathrm{MULTI}}$ and $\mathbf{P}_{\mathrm{RGB}}$, respectively.

The scaling configuration of the teacher and student networks is adjusted to provide flexibility in balancing representational strength and computational cost. Given the multimodal input $\mathbf{X}_{\mathrm{MULTI}}$, the teacher network employs larger scaling coefficients of $[1.00, 1.00, 512]$, corresponding to depth, width, and maximum channel capacity, respectively. This design produces a deeper and wider architecture with 656 layers, 26.2M parameters, and a computational complexity of 91.3 GFLOPs. The resulting high-capacity model is able to learn rich and discriminative cross-modal representations, which serve as informative supervision during knowledge distillation.

In contrast, the student network processes RGB-only input, $\mathbf{X}_{\mathrm{RGB}}$, using reduced scaling coefficients of $[0.50, 0.25, 1024]$ for depth, width, and maximum channel capacity, respectively. This configuration yields a compact architecture comprising 344 layers, 2.69M parameters, and 6.9 GFLOPs. Owing to its lightweight design, the student network substantially reduces computational burden, making it more suitable for edge deployment and UxV applications while maintaining sufficient representational capability to benefit from the distilled knowledge.

3.2.2. Detection Head

The detection heads of the teacher and student networks follow a similar architectural design, with the main difference arising from the additional distillation objective applied to the student branch.

In the teacher network, the backbone processes the multimodal input $\mathbf{X}_{\text{MULTI}}$ to produce a set of feature maps, denoted by $\mathbf{P}_{\text{MULTI}}$. These representations encode multi-scale spatial and semantic information that is essential for accurate object detection. The teacher detection head further transforms these features into $\mathbf{P}_{\text{TEACH}}(\cdot)$, which denotes the refined and aligned teacher representations. As shown in Fig. 2, $\mathbf{P}_{\text{TEACH}}(\cdot)$ serves as the supervisory signal in the distillation stage, where it is matched with the corresponding student representations to facilitate effective knowledge transfer.

In the student network, the detection head operates on the backbone features $\mathbf{P}_{\text{RGB}}$. These feature maps are processed through a sequence of convolutional transformations, including upsampling, downsampling, and feature concatenation, in order to enhance multi-scale representation learning. Such hierarchical processing enables the detector to better handle objects of different sizes. The resulting detection output, denoted by D_{STU}, is used to predict oriented bounding boxes, allowing more precise localization for objects that are not aligned with the image axes. In parallel, the student branch also generates aligned intermediate representations, denoted by $\mathbf{P}_{\text{STU}}(\cdot)$, which participate in the knowledge distillation process.

The student detection pipeline is optimized through a joint objective defined over both detection supervision and distillation guidance. Specifically, the predictions from D_{STU} are compared with the ground-truth annotations using a loss function that combines localization loss for bounding box regression and classification loss for category prediction. In addition, the student representations are constrained to align with the teacher representations during distillation. This design allows the student network to inherit informative guidance from the teacher while preserving computational efficiency.

3.2.3. Knowledge Distillation Process

The knowledge distillation process is illustrated within the proposed framework, where the outputs of the teacher and student feature pyramid networks are denoted by $\mathbf{P}_{\text{TEACH}}(\cdot)$ and $\mathbf{P}_{\text{STU}}(\cdot)$, respectively. Distillation is performed by aligning the corresponding feature blocks of the teacher and student networks to facilitate feature-level knowledge transfer. Since the feature dimensions of the two branches may differ, the student feature maps $\mathbf{P}_{\text{STU}}(\cdot)$ are first projected into the teacher feature space using a learnable transformation, such as a 1×1 convolution followed by batch normalization. This operation produces dimensionally compatible representations and enables effective alignment between the two networks, where l denotes the layer index.

After alignment, both the teacher features and the transformed student features are normalized in a channel-wise manner using a softmax operation over the spatial dimension. This normalization converts the feature responses into probability distributions, thereby facilitating stable and effective feature distillation.

Feature Alignment.

$$\mathcal{P}_{\text{STU}}^{(l)} = \text{BatchNorm}\left(\text{Conv}_{1\times1}\left(\mathcal{P}_{\text{STU}}^{(l)}\right)\right) \tag{1}$$

3.2.4. Loss Calculation and Optimization

During training, the teacher network transfers supervisory knowledge to the RGB-only student detector through channel-wise distillation. The overall optimization objective, denoted by $\mathcal{L}_{\text{Total}}$, is formulated as a weighted combination of the distillation loss and the student detection loss. The distillation term, $\mathcal{L}_{\text{DISTILL}}$, measures the discrepancy between the teacher and student feature representations, whereas the student loss, $\mathcal{L}_{\text{Student}}$, consists of localization and classification components. Accordingly, the student loss is defined as

$$\mathcal{L}_{\text{Student}} = \mathcal{L}_{\text{Loc}} + \mathcal{L}_{\text{Cls}}. \tag{2}$$

The channel-wise distillation objective is used to align the intermediate feature representations of the teacher and student networks. This loss enables the student model to inherit informative supervision from the multimodal teacher while operating solely on RGB input. Specifically, the channel-wise distillation (CWD) loss is computed by applying softmax normalization over the spatial dimensions of the teacher and student feature maps, followed by the Kullback–Leibler (KL) divergence:

$$\mathcal{L}_{\text{CWD}} = \sum_{i=1}^{n} \text{KL}\left(\text{softmax}\left(T_i/\tau\right), \text{softmax}\left(S_i/\tau\right)\right)\tau^2, \tag{3}$$

where T_i and S_i denote the teacher and student feature maps at the i-th matched layer, respectively, τ is the distillation temperature, and n denotes the total number of aligned layers. Here, $\text{KL}(\cdot)$ represents the Kullback–Leibler divergence.

The total training objective is then given by

$$\mathcal{L}_{\text{Total}} = \alpha\lambda_{\text{DISTILL}}\mathcal{L}_{\text{DISTILL}} + \beta\mathcal{L}_{\text{Student}}, \tag{4}$$

where α and β are hyperparameters that balance the relative contributions of the distillation and supervised detection losses, and λ_{DISTILL} controls the strength of the distillation term.

The student network is optimized using stochastic gradient descent (SGD). At each training iteration, the gradients of $\mathcal{L}_{\text{Total}}$ are computed with respect to the student parameters, which are then updated as follows:

Table 2
Summary of RGB-IR Datasets

Dataset	Training	Validation	Testing
LLVIP [47]	12,750	961	274
KAIST[48]	76,262	14,299	4,767
SMOD [49]	6,284	1,178	392

$$\theta_{t+1} = \theta_t - \eta \frac{\partial \mathcal{L}_{\text{Total}}}{\partial \theta_t}, \tag{5}$$

where θ_t denotes the student network parameters at iteration t, and η is the learning rate.

Algorithm 1 Optimization Process: CrossFusionKD Framework

Input:
  D: RGB-IR Dataset
  $\mathbf{P}_{\text{TEACH}}(\cdot)$: Well-trained teacher features
  $\mathbf{X}_{\text{RGB}}$: Student inputs
  E: Number of epochs
  η: Learning rate
Output: Optimized parameters of the Student Network, θ_T
for epoch $e = 1$ to E **do**
  Teacher Network:
  $\mathcal{X}_{\text{MULTI}} \rightarrow \mathbf{P}_{\text{MULTI}} \rightarrow \mathbf{P}_{\text{TEACH}}(\cdot)$
  Student Network:
  $\mathcal{X}_{\text{RGB}} \rightarrow \mathbf{P}_{\text{RGB}}$
  Channel-Wise Distillation: $\mathcal{L}_{\text{DISTILL}}$
  Loss Calculation: $\mathcal{L}_{\text{STU}}, \mathcal{L}_{\text{Total}}$
  Optimization: $\theta_{t+1} \leftarrow \theta_t - \eta \frac{\partial \mathcal{L}_{\text{Total}}}{\partial \theta_t}$
end for

4. EXPERIMENT

This section includes the details of datasets, data augmentation techniques, training parameters, and real-time implementation. The framework is evaluated using various datasets, such as LLVIP, KAIST, SJTU Multispectral Object Detection, with a comparative analysis performed against the methods described in the anchor articles. This section also discusses the real-time results observed on an unmanned vehicle (UxV).

4.1. Datasets

4.1.1. LLVIP Dataset

The LLVIP dataset contains strictly aligned high-resolution visible and infrared image pairs acquired under low-light and nighttime conditions [47]. It comprises 15,985 paired RGB-IR images and is primarily designed for pedestrian detection in severely dark environments. Owing to substantial illumination variation and low-contrast visual characteristics, LLVIP provides a challenging benchmark for evaluating cross-modal knowledge transfer under adverse conditions. As summarized in Table 2, the dataset is split into 12,750 training images, 961 validation images, and 274 testing images.

4.1.2. KAIST Dataset

The KAIST Multispectral Pedestrian Dataset is a standard benchmark for multispectral pedestrian detection [48]. It contains paired RGB and IR images collected from a moving vehicle under diverse illumination settings, including daytime, low-light, and nighttime scenarios. Pedestrian instances are categorized by distance into *Near*, *Medium*, and *Far*. In addition, the annotations have been manually refined to reduce misalignment errors and improve label reliability. The dataset consists of 76,262 training images, 14,299 validation images, and 4,767 testing images.

4.1.3. SJTU Dataset

The SJTU Multispectral Object Detection (SMOD) dataset consists of synchronized RGB–IR image pairs for object detection in nighttime environments [49]. Although the original dataset includes multiple object categories, such as pedestrians, riders, bicycles, and cars, only the *person* category is retained in this study through manual annotation to ensure consistency across experiments. In total, the dataset contains 15,875 aligned infrared-visible image pairs characterized by substantial illumination variation and densely populated pedestrian scenes. The official split contains 6,284 training images, 1,178 validation images, and 392 testing images.

4.2. Data Augmentation

Data augmentation is employed to increase the diversity and robustness of the training data. The objective is to expose the model to a broader range of visual variations while preserving the underlying data distribution. For all datasets, augmentation strategies include auto-orientation, resizing to 640×640, horizontal flipping, and random noise injection affecting up to 0.3% of pixels. These transformations improve sample diversity and help mitigate overfitting during training. By presenting varied image realizations across training iterations, the model is encouraged to learn more robust and generalized feature representations. Consequently, the proposed framework achieves improved generalization on unseen data, including challenging scenarios not explicitly observed during training.

4.3. Training Parameters

All experiments were conducted on a single NVIDIA GeForce RTX 4070 Ti Super GPU, with ResNet-50 employed as the backbone network. Training was performed using a batch size of 16 and optimized with stochastic gradient descent (SGD). The main training settings included a patience value of 100, a momentum coefficient of 0.937, and a warm-up momentum of 0.8. All input images were resized to 640 × 640, and the model was trained for 100 epochs. The initial learning rate was set to $lr_0 = 0.001$, while the final learning rate was maintained at $lr_f = 0.001$. The

Table 3
Computational Complexity and Real-Time Performance on Jetson Orin NX

Model	Params	FLOPs	mAP@50	FPS	Lat. (ms)
Teacher (RGB+IR)	26.2M	91.3G	98.3	12	83
CrossFusionKD –Student	2.6M	6.9G	**96.7**	44	23

implementation was developed in PyTorch with CUDA 12.4 support.

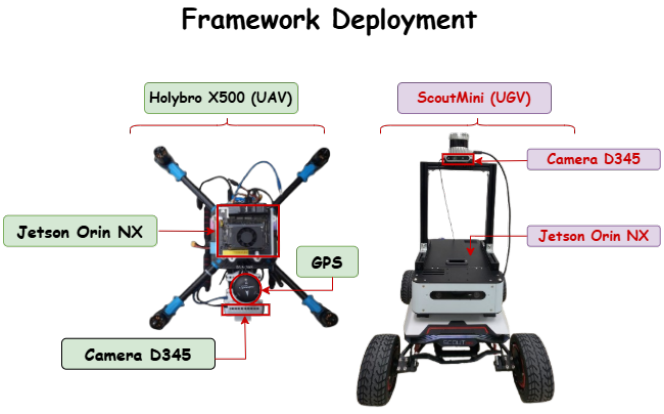


Figure 3: Deployment devices utilized for real-time detection: a SCOUT MINI UGV (right) and a Holybro X500 v2 UAV (left), each equipped with a Jetson Orion NX running ROS2.

4.4. Realtime-Implementation Details

The framework was deployed on UAV and UGV platforms using a Jetson Orin NX 16 GB module for onboard inference under ROS2 Humble, ensuring interoperability across heterogeneous robotic systems. Inference was performed with batch size = 1 using FP16 precision to reduce memory usage and improve throughput. No TensorRT-specific graph optimization was applied, ensuring a fair evaluation of the base model performance. The system was evaluated under standard power mode settings of the Jetson Orin NX during continuous real-time operation. The HOLYBRO X500 v2 UAV (Pixhawk 6C flight controller) and SCOUT MINI UGV both utilized the same onboard compute module, while an Intel RealSense D435 camera was employed for real-time image acquisition across varying environmental conditions.

4.5. Results

The proposed framework was evaluated on standard benchmark datasets, and the corresponding comparative results are reported in this section. Each dataset was trained for 100 epochs. The proposed method achieved mAP@50 scores of 92.1% on KAIST, 96.7% on LLVIP, and 93.1% on SJTU-SMOD. Although performance on KAIST was slightly lower than that obtained on the other two datasets, the overall results indicate that the framework provides a favorable balance between detection performance and computational efficiency across diverse operating conditions.

Table 4
Comparison of Performance (mAP, %) on the LLVIP Dataset

Methods	Backbone	mAP_{50} ↑	mAP ↑	Modality
RetinaNet[50]	ResNet50	88.0	42.8	RGB
Faster R-CNN[51]	ResNet50	87.0	45.1	
DDQ-DETR[52]	ResNet50	86.1	46.7	
Cascade R-CNN[53]	ResNet50	88.3	47.0	
CrossFusionKD(*)	**ResNet50**	**96.7**	**72.6**	
Halfway Fusion[54]	VGG16	91.4	55.1	RGB+IR
RSDet[55]	ResNet50	95.8	61.3	
GAFF[56]	ResNet18	94.0	55.8	
CSAA[57]	ResNet50	87.0	59.2	
ProbEn[58]	ResNet50	93.4	51.5	
CrossFusionKD(*)	**ResNet50**	**98.3**	**81.2**	

Table 5
Comparison of Performance (mAP, %) on the SMOD Dataset

Methods	Backbone	mAP50 (%)	mAP (%)
MGD [59]	ResNet18	89.9	61.7
CWD [60]	ResNet18	91.4	62.2
AFMD [61]	ResNet18	91.5	**63.8**
CrossFusionKD(*)	**ResNet18**	**93.2**	65.6

Table 4 reports the object detection performance on the LLVIP dataset under single-modality (RGB) and multi-modality (RGB+IR) configurations. Standard detection baselines, including SSD, RetinaNet, Faster R-CNN, and DDQ-DETR, operate on RGB input only, while fusion-based methods such as Halfway Fusion, RSDet, GAFF, and ProbEn leverage both RGB and IR information. CrossFusionKD attains 96.7% mAP with RGB input and 98.3% mAP with fused RGB+IR input, thereby achieving the best overall performance among all compared approaches.

Table 5 presents the performance comparison of different methods on the SMOD dataset under the RGB-only setting. The baseline methods, including MGD, CWD, and AFMD, employ ResNet18 as the backbone, whereas the proposed method uses ResNet50 and achieves the highest mAP@50 of 92.8% and mAP of 65.6%. Compared with the strongest baseline, AFMD, which reports 63.8% mAP, the proposed approach attains improved overall performance while benefiting from a deeper backbone architecture. Moreover, it outperforms both MGD and CWD in terms of mAP@50 and mAP, further demonstrating the effectiveness of the learned feature representations. These results indicate the robustness of the proposed method for RGB-based object detection on the SMOD dataset.

Fig. 4 presents qualitative results of the proposed CrossFusionKD object detection framework on benchmark datasets, namely KAIST, LLVIP, and SJTU-SMOD, together with real-world deployment examples. The KAIST and LLVIP results demonstrate reliable detection performance in low-illumination environments, whereas the SJTU-SMOD samples further verify the effectiveness of the framework under diverse nighttime conditions. Moreover, real-time deployment on both an unmanned ground vehicle (UGV) and an unmanned aerial vehicle (UAV) confirms the feasibility of

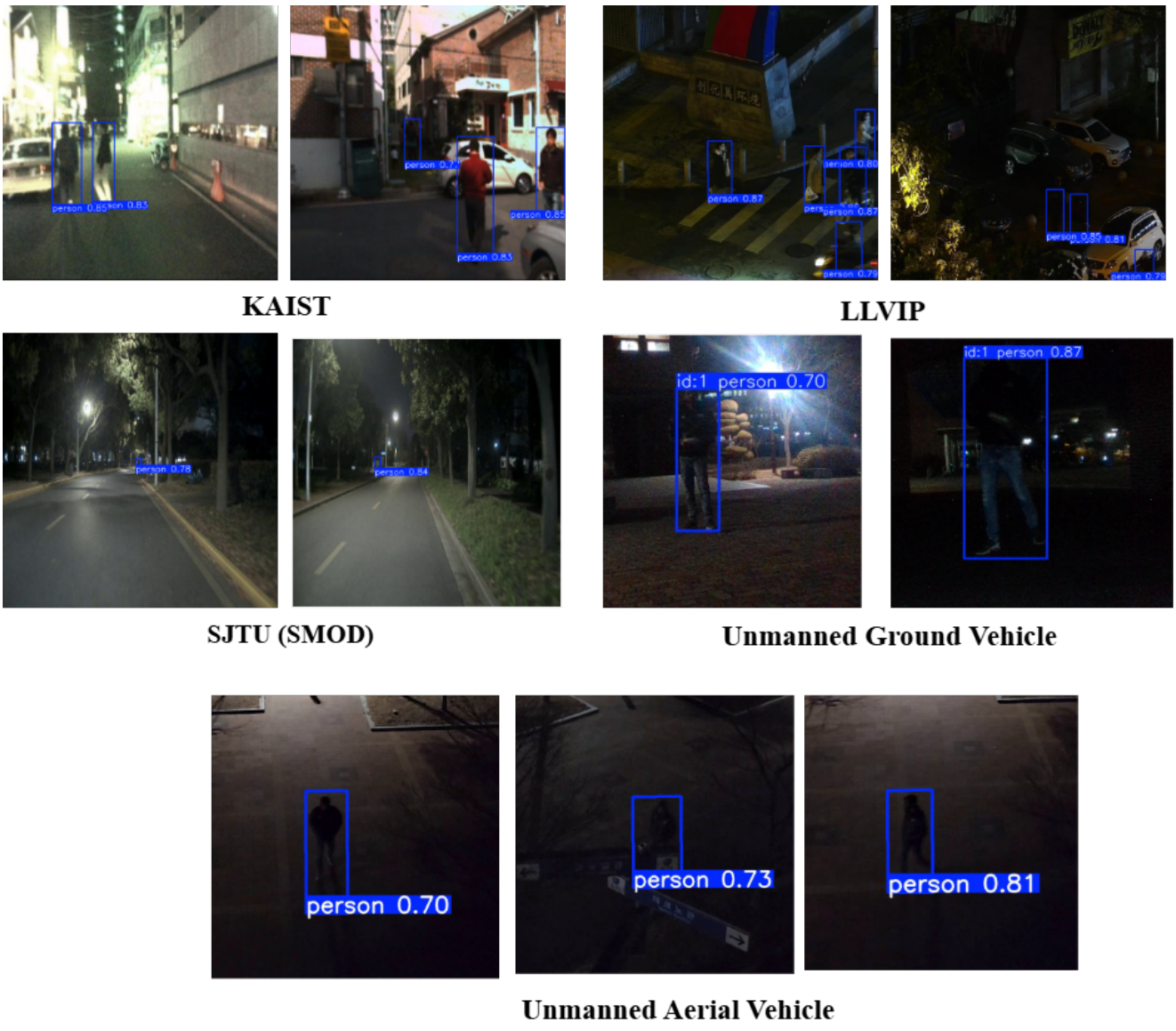


Figure 4: Qualitative comparison of the proposed CrossFusionKD object detector across standard datasets (KAIST, LLVIP, SJTU-SMOD) and real-time deployments. The top row shows results on KAIST and LLVIP datasets, The second row includes SJTU-SMOD (left) and UGV (right), and the final row illustrates detection on the UAV platform.

the proposed approach for edge inference on computationally constrained platforms.

5. Ablation Studies

This section analyzes the contribution of different components of the proposed CrossFusionKD framework. All ablation experiments were conducted under identical training settings using the same backbone architecture, optimization strategy, and dataset splits to ensure fair comparison. Unless otherwise specified, experiments were performed on the LLVIP dataset using RGB-only inference.

5.1. Effect of Distillation Strategy

To assess the effectiveness of the proposed channel-wise distillation (CWD) strategy, we compare it with two alternative settings: (i) a baseline student trained without knowledge distillation and (ii) Mask Generation Distillation (MGD), a representative feature-based distillation method. The corresponding results are reported in Table 6.

The baseline model achieves 91.2% mAP@50. Incorporating MGD improves performance to 94.3%, indicating that transferring masked feature representations from the multimodal teacher enhances the student detector. However, the proposed CWD strategy further increases performance to 96.7% mAP@50, yielding a total improvement of +5.5% over the baseline and +2.4% over MGD. This improvement

Table 6
Comparison of Distillation Strategies on the LLVIP Dataset (RGB-only Inference). Performance evaluation of different knowledge distillation approaches, including MGD and channel-wise distillation (CWD), demonstrating the effectiveness of spatial distribution alignment.

Method	mAP@50 (%)	mAP (%)	Gain
Baseline (No KD)	91.2	67.1	–
MGD	94.3	70.5	+3.1
CWD	**96.7**	**72.6**	+5.5

Table 7
Effect of the Distillation Weight $\lambda_{distill}$ on Detection Performance. Performance variation of the student model under different distillation weight settings, illustrating the balance between supervised learning and knowledge transfer.

$\lambda_{distill}$	mAP@50 (%)	mAP (%)
0.1	94.9	71.0
0.5	95.8	72.1
1.0	**96.7**	**72.6**
2.0	95.4	71.5

demonstrates that channel-wise spatial distribution alignment is more effective for cross-modal knowledge transfer than masked generative feature regression. By normalizing feature responses across spatial dimensions and aligning their probabilistic distributions, CWD enables the student model to capture modality-invariant structural cues rather than modality-specific intensity patterns. The results confirm that CWD is particularly suitable for mitigating the modality gap between RGB-IR teacher features and RGB-only student representations.

5.2. Effect of Distillation Weight

The impact of the distillation loss weight $\lambda_{distill}$ was analyzed to assess sensitivity and robustness of the proposed framework. As shown in Table 7, performance increases steadily as $\lambda_{distill}$ grows from 0.1 to 1.0, reaching optimal performance at $\lambda_{distill} = 1.0$. Beyond this value, performance slightly degrades.

When the distillation weight is too small, the influence of the teacher signal becomes limited, reducing knowledge transfer effectiveness. Conversely, overly large weights cause the student model to overemphasize feature imitation at the expense of supervised learning from ground-truth annotations. The results indicate that a balanced contribution between supervised loss and distillation loss is essential for optimal performance. The stable behavior across a wide range of $\lambda_{distill}$ values also demonstrates robustness of the proposed method.

6. Conclusion

This paper introduced CrossFusionKD, a cross-modal knowledge distillation framework developed for efficient RGB-only pedestrian detection on resource-constrained UxV platforms. By transferring modality-enriched representations from an RGB–IR teacher network to a lightweight RGB student model, the proposed framework preserves discriminative visual information while avoiding the computational cost associated with multimodal inference during deployment.

The teacher network, with a computational complexity of 91.3 GFLOPs, provides informative cross-modal supervision, whereas the distilled student model remains compact, with 2.6M parameters and 6.9 GFLOPs. Experimental results on the LLVIP, KAIST, and SMOD datasets show consistent improvements over the baseline student model, with gains of up to 5.5% mAP while maintaining real-time performance of 44 FPS on the Jetson Orin NX platform. In addition, backbone-normalized comparisons indicate that these improvements are not merely a consequence of increased architectural capacity, but rather stem from effective cross-modal knowledge transfer.

Real-world deployment on UAV and UGV platforms further demonstrates the practical applicability of the proposed method for edge-based perception tasks. Overall, the results show that CrossFusionKD achieves an effective balance between detection accuracy and computational efficiency, making it well suited for long-duration autonomous operation in dynamic and low-illumination environments. Overall, the framework establishes an effective strategy for modality-aware model compression, enabling scalable and deployable multi-spectral knowledge transfer for real-time robotic perception systems.

7. Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2026-RS-2024-00437190, 50%) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2026-RS-2020-II201612, 50%)

References

[1] Y. Wang, Z. Su, Q. Xu, R. Li, T. H. Luan, P. Wang, A secure and intelligent data sharing scheme for uav-assisted disaster rescue, IEEE/ACM Transactions on Networking 31 (6) (2023) 2422–2438.

[2] S. M. H. Abidi, S. M. Raza, S. Y. Shin, Safevision: Vision-language reasoning for context-aware safety monitoring, Neurocomputing (2025) 132479.

[3] S. M. Raza, M. Masuduzzaman, S. Y. Shin, Uxv-swarm based automated fruits disease instance segmentation for smart farming using deep learning model, in: 2024 Fifteenth International Conference on Ubiquitous and Future Networks (ICUFN), IEEE, 2024, pp. 274–279.

[4] Z. Shao, G. Cheng, J. Ma, Z. Wang, J. Wang, D. Li, Real-time and accurate uav pedestrian detection for social distancing monitoring in covid-19 pandemic, IEEE transactions on multimedia 24 (2021) 2069–2083.

[5] M. Masuduzzaman, R. Nugraha, S. Y. Shin, Uav-agv cooperated remote toxic gas sensing and automated alarming scheme in smart factory, Computer Communications 226 (2024) 107923.

[6] M. Masuduzzaman, T. Rahim, A. Islam, S. Y. Shin, Uxv-based deep-learning-integrated automated and secure garbage management scheme using blockchain, IEEE Internet of Things Journal 10 (8) (2022) 6779–6793.

[7] J. Du, G. Han, C. Lin, X. Kong, G. Shen, An autonomous underwater vehicle-assisted adaptive trust prediction model for uasns, IEEE Transactions on Intelligent Vehicles (2024).

[8] Q. Wang, Y. Chi, T. Shen, J. Song, Z. Zhang, Y. Zhu, Improving rgb-infrared object detection by reducing cross-modality redundancy, Remote Sensing 14 (9) (2022) 2020.

[9] Y. Sun, B. Cao, P. Zhu, Q. Hu, Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning, IEEE Transactions on Circuits and Systems for Video Technology 32 (10) (2022) 6700–6713.

[10] S. Vadera, S. Ameen, Methods for pruning deep neural networks, IEEE Access 10 (2022) 63280–63300.

[11] B. Rokh, A. Azarpeyvand, A. Khanteymoori, A comprehensive survey on model quantization for deep neural networks in image classification, ACM Transactions on Intelligent Systems and Technology 14 (6) (2023) 1–50.

[12] M. Ji, G. Peng, S. Li, F. Cheng, Z. Chen, Z. Li, H. Du, A neural network compression method based on knowledge-distillation and parameter quantization for the bearing fault diagnosis, Applied Soft Computing 127 (2022) 109331.

[13] A. Alkhulaifi, F. Alsahli, I. Ahmad, Knowledge distillation in deep learning and its applications, PeerJ Computer Science 7 (2021) e474.

[14] L. Zhou, D. Wu, J. Chen, X. Wei, Cross-modal collaborative communications, IEEE wireless communications 27 (2) (2019) 112–117.

[15] S. Sun, W. Ren, J. Li, R. Wang, X. Cao, Logit standardization in knowledge distillation, IEEE Transactions on Computer Vision and Pattern Recognition XX (X) (2024) 15731–15740.

[16] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, T. Zhou, A survey on knowledge distillation of large language models, arXiv preprint arXiv:2402.13116 (2024).

[17] R. Miles, I. Elezi, J. Deng, Vkd: Improving knowledge distillation using orthogonal projections, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15720–15730.

[18] Z. Huang, S. Yang, M. Zhou, Z. Li, Z. Gong, Y. Chen, Feature map distillation of thin nets for low-resolution object recognition, IEEE Transactions on Image Processing 31 (2022) 1364–1379.

[19] D. Soydaner, Attention mechanism in neural networks: where it comes and where it goes, Neural Computing and Applications 34 (16) (2022) 13371–13385.

[20] X. Li, H. Xiong, H. Wang, Y. Rao, L. Liu, Z. Chen, J. Huan, Delta: Deep learning transfer using feature map with attention for convolutional networks, arXiv preprint arXiv:1901.09229 (2019).

[21] Z. Li, P. Xu, X. Chang, L. Yang, Y. Zhang, L. Yao, X. Chen, When object detection meets knowledge distillation: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (8) (2023) 10555–10579.

[22] G. Li, X. Li, Y. Wang, S. Zhang, Y. Wu, D. Liang, Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 36, 2022, pp. 1306–1313.

[23] S. Ye, Y. Zhang, J. Hu, L. Cao, S. Zhang, L. Shen, J. Wang, S. Ding, R. Ji, Distilpose: Tokenized pose regression with heatmap distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2163–2172.

[24] A. Chawla, H. Yin, P. Molchanov, J. Alvarez, Data-free knowledge distillation for object detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3289–3298.

[25] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, C. Yuan, Focal and global knowledge distillation for detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4643–4652.

[26] Y. Mo, Y. Wu, X. Yang, F. Liu, Y. Liao, Review the state-of-the-art technologies of semantic segmentation based on deep learning, Neurocomputing 493 (2022) 626–646.

[27] D. Ji, H. Wang, M. Tao, J. Huang, X.-S. Hua, H. Lu, Structural and statistical texture knowledge distillation for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16876–16885.

[28] P. Chen, S. Liu, H. Zhao, J. Jia, Distilling knowledge via knowledge review, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5008–5017.

[29] Y. Qin, B. Yang, S. Ke, P. Liu, F. Rong, X. Xia, M-fanet: Multi-feature attention convolutional neural network for motor imagery decoding, IEEE Transactions on Neural Systems and Rehabilitation Engineering (2024).

[30] X. Li, H. Xiong, Z. Chen, J. Huan, J. Liu, C.-Z. Xu, D. Dou, Knowledge distillation with attention for deep transfer learning of convolutional networks, ACM Transactions on Knowledge Discovery from Data (TKDD) 16 (3) (2021) 1–20.

[31] J. Wang, Y. Chen, Z. Zheng, X. Li, M.-M. Cheng, Q. Hou, Crosskd: Cross-head knowledge distillation for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16520–16530.

[32] S. Kim, Y. Kim, S. Hwang, H. Jeong, D. Kum, Labeldistill: Label-guided cross-modal knowledge distillation for camera-based 3d object detection, in: European Conference on Computer Vision, Springer, 2024, pp. 19–37.

[33] Z. Zheng, R. Ye, Q. Hou, D. Ren, P. Wang, W. Zuo, M.-M. Cheng, Localization distillation for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (8) (2023) 10070–10083.

[34] D. Ienco, C. F. Dantas, Discom-kd: Cross-modal knowledge distillation via disentanglement representation and adversarial learning, arXiv preprint arXiv:2408.07080 (2024).

[35] J. Hu, Y. Huang, X. Hu, Y. Xu, Enhancing the perceived emotional intelligence of conversational agents through acoustic cues, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–7.

[36] A. Li, S. Ni, Y. Chen, J. Chen, X. Wei, L. Zhou, M. Guizani, Cross-modal object detection via uav, IEEE Transactions on Vehicular Technology 72 (8) (2023) 10894–10905.

[37] C. Yang, X. Yu, H. Yang, Z. An, C. Yu, L. Huang, Y. Xu, Multi-teacher knowledge distillation with reinforcement learning for visual recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 9148–9156.

[38] D. P. Do, T. Kim, J. Na, J. Kim, K. Lee, K. Cho, W. Hwang, D3t: Distinctive dual-domain teacher zigzagging across rgb-thermal gap for domain-adaptive object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23313–23322.

[39] Y. Jiang, C. Feng, F. Zhang, D. Bull, Mtkd: Multi-teacher knowledge distillation for image super-resolution, in: European Conference on Computer Vision, Springer, 2024, pp. 364–382.

[40] Y.-e. Lin, S. Yin, Y. Ding, X. Liang, Atmkd: Adaptive temperature guided multi-teacher knowledge distillation, Multimedia Systems 30 (5) (2024) 292.

[41] S. Chen, P. Sun, Y. Song, P. Luo, Diffusiondet: Diffusion model for object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 19830–19843.

[42] K. Song, Y. Zhao, L. Huang, Y. Yan, Q. Meng, Rgb-t image analysis technology and application: A survey, Engineering Applications of Artificial Intelligence 120 (2023) 105919.

[43] L. Yang, G. Chen, W. Ci, Multiclass objects detection algorithm using darknet-53 and densenet for intelligent vehicles, EURASIP Journal on Advances in Signal Processing 2023 (1) (2023) 85.

[44] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, Z. Liu, Mobile-former: Bridging mobilenet and transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5270–5279.

[45] H. He, H. Xu, Y. Zhang, K. Gao, H. Li, L. Ma, J. Li, Mask r-cnn based automated identification and extraction of oil well sites, International Journal of Applied Earth Observation and Geoinformation 112 (2022) 102875.
[46] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al., Yolov6: A single-stage object detection framework for industrial applications, arXiv preprint arXiv:2209.02976 (2022).
[47] X. Jia, C. Zhu, M. Li, W. Tang, W. Zhou, Llvip: A visible-infrared paired dataset for low-light vision, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3496–3504.
[48] S. Hwang, J. Park, N. Kim, Y. Choi, I. S. Kweon, Multispectral pedestrian detection: Benchmark dataset and baselines, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[49] Z. Chen, Y. Qian, x. Yang, C. Wang, M. Yang, Amfd: Distillation via adaptive multimodal fusion for multispectral pedestrian detection, arXiv preprint arXiv:2405.12944 (2024).
[50] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
[51] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: IEEE transactions on pattern analysis and machine intelligence, 2016, pp. 1137–1149.
[52] S. Zhang, X. Wang, J. Wang, J. Pang, C. Lyu, W. Zhang, P. Luo, K. Chen, Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr), in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7329–7338.
[53] Z. Cai, N. Vasconcelos, Cascade r-cnn: High quality object detection and instance segmentation, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, pp. 1483–1498.
[54] J. Liu, S. Zhang, S. Wang, D. N. Metaxas, Multispectral deep neural networks for pedestrian detection, arXiv preprint arXiv:1611.02644 (2016).
[55] T. Zhao, M. Yuan, F. Jiang, N. Wang, X. Wei, Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion, arXiv preprint arXiv:2401.10731 (2024).
[56] H. Zhang, E. Fromont, S. Lefèvre, B. Avignon, Guided attentive feature fusion for multispectral pedestrian detection, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 72–80.
[57] Y. Cao, J. Bin, J. Hamari, E. Blasch, Z. Liu, Multimodal object detection by channel switching and spatial attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 403–411.
[58] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, S. Kong, Multimodal object detection via probabilistic ensembling, in: European Conference on Computer Vision, Springer, 2022, pp. 139–158.
[59] Z. Yang, Z. Li, M. Shao, D. Shi, Z. Yuan, C. Yuan, Masked generative distillation, in: European Conference on Computer Vision, Springer, 2022, pp. 53–69.
[60] C. Shu, Y. Liu, J. Gao, Z. Yan, C. Shen, Channel-wise knowledge distillation for dense prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5311–5320.
[61] Z. Chen, Y. Qian, X. Yang, C. Wang, M. Yang, Amfd: Distillation via adaptive multimodal fusion for multispectral pedestrian detection, arXiv preprint arXiv:2405.12944 (2024).



Syed Murtaza Hussain Abidi received a Master's degree in IT Convergence Engineering from the Wireless and Emerging Network Systems Laboratory (WENS Lab), Kumoh National Institute of Technology (KIT), South Korea. He is currently a graduate fellow and researcher at KIT. His research interests include computer vision, vision-language models, and ROS.



Syed Muhammad Raza received an M.S. degree in IT Convergence Engineering from the Wireless and Emerging Network Systems Laboratory (WENS Lab), Kumoh National Institute of Technology (KIT), South Korea. He is currently an AI Engineer with Neubility, Seoul, South Korea. His research interests include image processing, lightweight deep learning models, and unmanned vehicles (UxVs).



Md Masuduzzaman is a Postdoctoral Fellow at McMaster University, Canada. He received the Ph.D. degree in IT Convergence Engineering from the Wireless and Emerging Network Systems Laboratory (WENS Lab), Kumoh National Institute of Technology (KIT), South Korea. He also received the B.Sc. and M.Sc. degrees in Computer Science from the American International University-Bangladesh (AIUB). His research interests include blockchain, deep learning, the Internet of Things (IoT), network security, and unmanned vehicles (UxVs).



Soo Young Shin (M'07–SM'17) received the Ph.D. degree in Electrical Engineering and Computer Science from Seoul National University in 2006. He worked at Samsung Electronics from 2007 to 2010 and joined Kumoh National Institute of Technology, where he is currently an Associate Professor. His research interests include wireless communications, signal processing, unmanned mobility, quantum neural networks, the Internet of Things, and AI/deep learning applications.