Progressive Multi-modal Knowledge Distillation for Multi-spectral Object Re-identification

Aihua Zheng[1,2,5], **Pengyu Li**[1], **Zi Wang**[3,4], **Jin Tang**[4,5*]
[1]School of Artificial Intelligence, Anhui University
[2]State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology
[3]School of Biomedical Engineering, Anhui Medical University
[4]Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University
[5]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University
ahzheng214@foxmail.com, lpy2285901773@163.com, ziwang1121@foxmail.com, tangjin@ahu.edu.cn

Abstract

In the field of multi-spectral object re-identification (ReID), multi-modal knowledge and modal-specific knowledge exhibit complementary advantages when handling hard samples, but existing methods rarely integrate this collaborative information. Knowledge distillation is a direct approach for transferring information, however, heterogeneity in model architectures and variations in sample hardness can undermine the stability and controllability of knowledge transfer. To alleviate these limitations, we propose the novel **P**rogressive **M**ulti-modal **K**nowledge **D**istillation (**PMKD**) framework that enables multi-stage knowledge transfer guided by hard sample awareness. In the multi-modal knowledge transfer stage, the source model (pre-trained on multi-modal data) disseminates its learned multi-modal collaborative knowledge to multiple independently modal-specific target models, guiding their adaptation to hard samples within training batches. In the modal-specific knowledge retention stage, the independent models enriched with multi-modal knowledge guide the training phase. The architectural consistency between source-target models ensures more lossless knowledge transfer, effectively mitigating the risk of capability drift, and preserving inherent competence. Moreover, the entire progressive multi-modal knowledge distillation is regulated by the proposed hardness-aware distillation loss, which automatically adapts distillation intensity through hard sample mining, thereby ensuring stable transfer of hard sample handling capabilities. Extensive experiments on benchmark multi-spectral ReID datasets validate the effectiveness and superior performance of the proposed method.

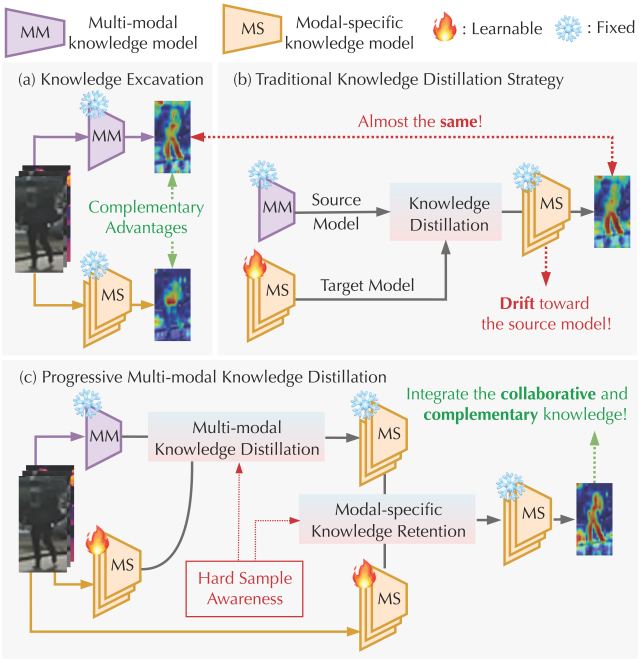

Figure 1: (a) Different models focus on distinct knowledge of the input. (b) Traditional knowledge distillation leads to ability drifting in the target model. (c) The proposed PMKD framework integrates the complementary strengths of heterogeneous models, emphasizing collaborative information.

Introduction

Object re-identification (ReID) aims to match objects across non-overlapping camera views, yet single-modal (visible light) methods perform poorly under poor illumination, occlusion, and adverse weather (Zhang et al. 2023; Wang et al. 2024b; Liu et al. 2024; Li, Sun, and Li 2023; He et al. 2021; Liu et al. 2023, 2021; Yu et al. 2024). Multi-spectral object ReID has emerged as the promising solution (Zheng et al. 2021; Wang et al. 2022, 2024c,a; Zhang et al. 2024; Wang et al. 2025b; Qiu et al. 2023), due to the fusion of complementary spectral data yields more comprehensive object representations and improved recognition in challenging environments.

Existing models designed for multi-spectral object ReID primarily focus on extracting two types of information: modal-specific and multi-modal knowledge. Specifically, (1) Some methods (Wang et al. 2022; Li et al. 2020; Wang et al. 2024a) typically adopt independent backbones to capture fine-grained but heterogeneous modal-specific knowledge from each modality. (2) Some approaches (Wang et al. 2024c; Zhang et al. 2024; Li et al. 2025) leverage the shared backbone to exploit collaborative multi-modal knowledge. For hard samples, these two types of models offer com-

*Corresponding author

plementary advantages, as illustrated in the class activation maps (Selvaraju et al. 2017) in Fig. 1 (a). Multi-modal knowledge models focus on the broader and semantically aligned areas but overlook certain specific nuances, which are the strength of modal-specific knowledge models. This inherent trade-off limits the ability of either paradigm to robustly address hard samples. Knowledge Distillation (KD) can transfer the rich collaborative representations of the multi-modal model to the modal-specific model, thereby integrating the strengths of both paradigms. However, applying KD to multi-spectral ReID still faces two key limitations: First, the architectural heterogeneity between multi-modal and modal-specific models often leads to representation misalignment and information drift (Huang et al. 2025; Hao et al. 2023). As shown in Fig. 1 (b), traditional KD may cause catastrophic capability drift in the target model, resulting in class activation maps that are almost identical to the source model. Secondly, hard samples in the dataset urgently require the complementary abilities of different models. Traditional KD strategies treat training samples equally, and this scheme may lead to overfitting of models, thereby hindering the effective transfer of the capabilities required to robustly handle hard samples.

To alleviate these limitations, we propose the novel Progressive Multi-modal Knowledge Distillation (PMKD) framework that enables multi-stage knowledge transfer guided by hard sample awareness. First, the multi-modal knowledge transfer aims to introduce multi-modal collaborative understanding into the target models and enhance their ability to tackle hard samples. We jointly train a multi-modal source model on full-modality data and distill its collaborative representations into independent modal-specific target models through the Hardness-aware Distillation Mechanism (HDM). Next, the modal-specific knowledge retention focuses on mitigating the potential capability drift caused by architectural heterogeneity and preserving the modal-specific representations. We employ distilled independent models enriched with multi-modal knowledge as new source models, and similarly employ HDM to guide the training of the new target models with the same architecture. Notably, HDM is optimized by the ReID loss and the proposed hardness-aware distillation loss (HDL) to prevent target models from overfitting and preserve inherent capabilities. HDL performs hard sample mining based on the distance map within each batch to estimate sample hardness and then assigns larger distillation weights to harder samples while applying weaker constraints to easier ones.

As a result, the final model, obtained through two rounds of HDM, preserves modal-specific information while gaining enhanced capabilities from the multi-modal model, producing more accurate and collaborative attention when handling hard samples, as shown in Fig. 1 (c). Sufficient experimental results and in-depth analyses demonstrate the advantages of the proposed method. Our main contributions can be summarized as follows.

- We design the Progressive Multi-modal Knowledge Distillation framework for multi-spectral object ReID, which integrates multi-modal knowledge and modal-specific knowledge to enhance the capability of handling hard samples and break the performance bottleneck.
- We propose Multi-modal Knowledge Transfer introduces multi-modal collaborative understanding into modal-specific models and Modal-specific Knowledge Retention leverages architecture-consistent distillation to mitigate capability drift.
- We introduce the Hardness-aware Distillation loss, which guides the target model's learning toward challenging instances while preventing overfitting to the source model, thereby preserving its inherent capabilities and ensuring stable transfer of hard-sample handling ability.
- Extensive experiments on three benchmark multi-spectral ReID datasets demonstrate that our framework consistently surpasses state-of-the-art approaches and can be seamlessly integrated into existing methods to further enhance their performance.

Related Work

Multi-spectral Re-identification

Multi-spectral object Re-identification (ReID) aims to obtain comprehensive and discriminative representations by fusing visual features from different spectral bands. Existing approaches can be categorized into modal-specific knowledge models and multi-modal knowledge models, each with distinct advantages and limitations. Modal-specific models employ independent backbones for each modality to preserve fine-grained, modal-specific details. For example, IEEE (Wang et al. 2022) enhances modal-specific information via cross-modal interaction and multi-modal margin loss, while TOP-ReID (Wang et al. 2024a) leverages modality permutation and reconstruction to facilitate multi-modal interaction, and FACENet (Zheng et al. 2025) restores flare-corrupted features using thermal infrared guidance. Although these methods retain modal-specific cues, they often lack deep multi-modal synergy, which is crucial for identifying hard samples. Multi-modal models, in contrast, adopt shared backbones to promote deep multi-modal synergy. For instance, ICPL-ReID (Li et al. 2025) leverages an identity prototype-based conditional prompt learning strategy to steer semantic acquisition across different modalities, and IDEA (Wang et al. 2025c) guides model learning via projecting textual features into the visual domain and deformable aggregation. DeMo (Wang et al. 2025b) further employs a mixture-of-experts to adaptively handle modality information confusion. While these approaches excel at multi-modal fusion, they often suppress subtle modal-specific cues that are critical for distinguishing fine-grained identities. Despite these advancements, modal-specific models lack multi-modal reasoning for hard samples, while multi-modal models compromise fine-grained details. These limitations motivate our Progressive Multi-modal Knowledge Distillation (PMKD), which integrates the strengths of both paradigms to effectively handle challenging samples.

Knowledge Distillation

Knowledge distillation transfers knowledge from a powerful teacher to a student model, originally aimed at model

compression (Park et al. 2019; Wang and Yoon 2021; Tung and Mori 2019; Jia et al. 2024), but increasingly adapted for enhancing multi-modal learning. Some methods explore sample hardness to guide more effective knowledge transfer. For example, DKD (Zhao et al. 2022) disentangles target and non-target class knowledge to separately model dark knowledge and sample difficulty. BTKD (Zhang et al. 2025a) treats hard samples as biased teacher outputs and employs a progressive strategy that shifts focus from easy to hard samples, enhancing learning efficiency and generalization. In the ReID domain, particularly in cross-modal (Shi et al. 2024a, 2023, 2024b; Lu, Zou, and Zhang 2023; Chen et al. 2023) settings, KD has been adapted to facilitate knowledge exchange between each modality branch. IDKL (Ren and Zhang 2024) distills implicit discriminative knowledge from modal-specific features into modal-shared representations, improving cross-modality alignment. TSKD (Shi et al. 2025) enhances visible-infrared ReID by aligning modal-specific features through self-mimic and mutual-distillation learning. However, these methods still face limitations in multi-spectral ReID. Most existing hard-sample-aware distillation approaches are misaligned with the open-set, embedding-driven nature of ReID. In contrast, feature-based distillation methods avoid this mismatch but typically fail to account for sample hardness, limiting their ability to handle challenging instances effectively. Meanwhile, cross-modal ReID approaches mainly align modalities and often overlook the complementary cues across richer multi-modal inputs. To address these challenges, we propose the Progressive Multi-modal Knowledge Distillation framework, which effectively integrates both multi-modal and modal-specific knowledge to handle hard samples better.

Method

In this section, we present the proposed Progressive Multi-modal Knowledge Distillation method in detail. As shown in Fig. 2, PMKD comprises three key components: (1) Progressive Multi-modal Knowledge Distillation, which transfers multi-modal knowledge to the target model in a staged manner. (2) Hardness-aware Distillation Mechanism, which provides the unified optimization and transfer pipeline across both distillation stages. (3) Hardness-aware Distillation Loss based on hard sample mining and distillation weight calculation, which identifies challenging samples and modulates distillation strength accordingly.

Progressive Knowledge Distillation Strategy

Our progressive knowledge distillation strategy unfolds in three distinct phases. We first detail the knowledge excavation approach, which focuses on excavating multi-modal knowledge and modal-specific knowledge. Subsequently, we elaborate on the two specialized distillation stages: Multi-modal Knowledge Transfer and Modal-specific Knowledge Retention, which progressively optimize multi-modal synergistic understanding and modal-specific fine-grained details.

Knowledge Excavation To capture collaborative understanding across different modalities, our models employ the shared backbone, denoted as ϕ_{sha}. This allows for learning multi-modal representations f_R^{mul}, f_N^{mul}, f_T^{mul} of each modality input, which are obtained from $\phi_{sha}(I_R)$, $\phi_{sha}(I_N)$, $\phi_{sha}(I_T)$, where I_R, I_N, I_T denote the input of RGB, Near-Infrared, and Thermal Infrared modalities. This shared encoder is crucial for distilling integrated multi-modal insights. Concurrently, to capture modal-specific knowledge that is often more unique and fine-grained, we utilize distinct, modality-independent backbones for each modality. We denote a set of such backbones as $\phi_{ind} = \{\phi_R, \phi_N, \phi_T\}$. Thus, modal-specific knowledge features f_R^{spe}, f_N^{spe}, f_T^{spe} are acquired from $\phi_R(I_R)$, $\phi_N(I_N)$, $\phi_T(I_T)$. Each f^{spe} is designed to retain specific characteristics inherent to its source modality, providing heterogeneous information.

Crucially, all models within our PMKD framework, including the initial source and subsequent target models, consistently employ this feature extraction approach. Then, we elaborate on the specific stages of our progressive knowledge distillation strategy.

Multi-modal Knowledge Transfer The progressive distillation strategy begins by establishing a robust initial source multi-modal model with shared backbone ϕ_{sha}, which is trained using all modalities' data with standard ReID losses (Cross-Entropy loss (Szegedy et al. 2016) and Triplet loss (Hermans, Beyer, and Leibe 2017)) to ensure strong multi-modal collaborative understanding. In the first stage, termed Multi-modal Knowledge Transfer, aims to transfer the robust multi-modal collaborative knowledge from ϕ_{sha} to newly initialized set of independent target models ϕ_{ind}. During this stage, the parameters of ϕ_{sha} are fixed, while ϕ_{ind} is optimized to learn from the integrated representations of ϕ_{sha}. The training objective combines the standard ReID loss L_{reid} and the distillation loss L_{dis}. Upon completion, the final checkpoint of ϕ_{ind} is saved.

Modal-specific Knowledge Retention The second stage of our progressive distillation strategy is Modal-specific Knowledge Retention, which aims to mitigate the capability drift introduced during the last distillation stage and further preserve the modal-specific knowledge. In this phase, the independent models ϕ_{ind} enriched with multi-modal knowledge now serve as the new source model. A fresh set of independent backbones, denoted as $\phi_{ind}^{'}$, is initialized as the target model, ensuring architectural consistency with the source model. Similar to the earlier stage, the parameters of ϕ_{ind} are fixed, while $\phi_{ind}^{'}$ is optimized using the total loss comprising the standard ReID loss L_{reid} and the distillation loss L_{dis}. This homogeneous setup facilitates a more precise and stable knowledge transfer, allowing $\phi_{ind}^{'}$ to maintain and refine more modal-specific information, ultimately mitigating capability drift. Notably, during inference, only the final $\phi_{ind}^{'}$ is required, with no additional distillation stages.

Hardness-aware Distillation Mechanism

This section details the Hardness-aware Distillation Mechanism, which encompasses the unified optimization and knowledge transfer pipeline applied to the target model dur-

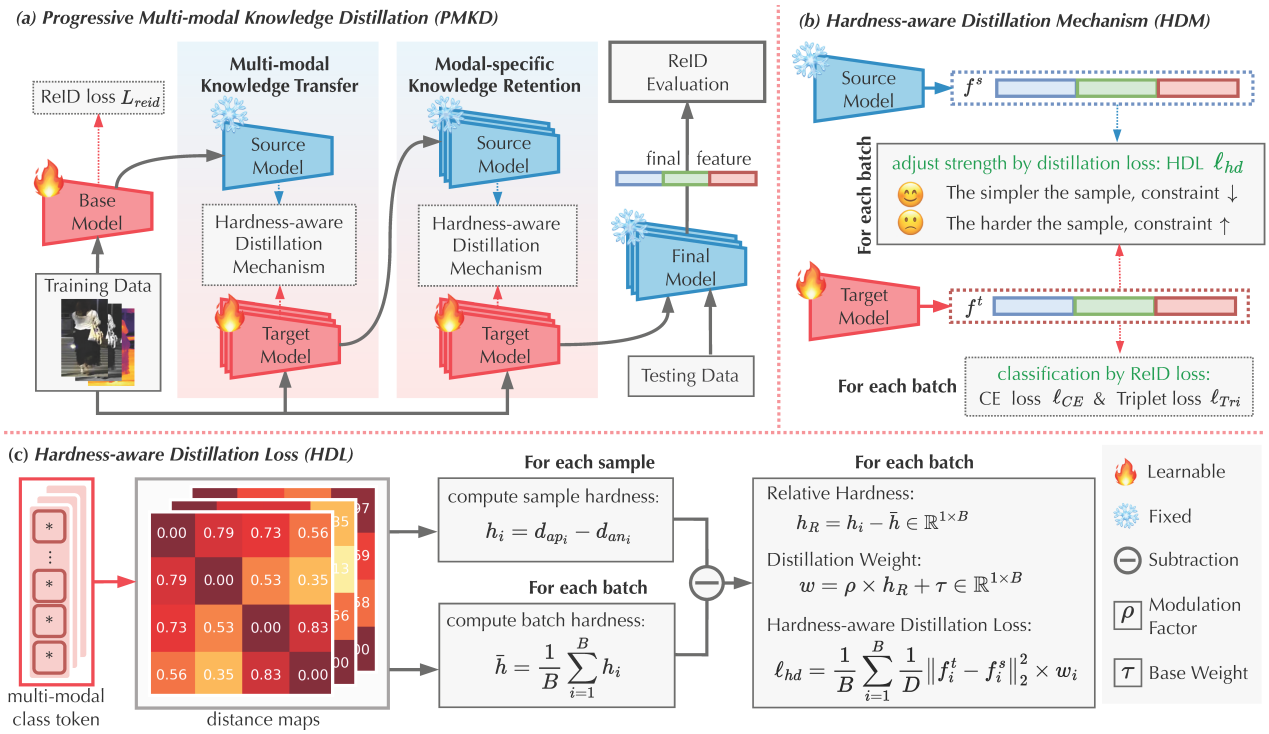

Figure 2: (a) PMKD achieves multi-modal knowledge transfer and modal-specific knowledge retention progressively through Hardness-aware Distillation Mechanism. (b) Hardness-aware Distillation Mechanism measures sample hardness using the HDL ℓ_{hd} to adjust the distillation strength, while the ReID loss guides identity classification. (c) HDL is computed by performing hard sample mining based on the distance map of each batch, and then calculating the distillation weight for adaptive modulation.

ing each distillation stage. This mechanism defines how extracted features are processed, aggregated, and how knowledge is transferred from source model to target models, leveraging the feature extraction methods outlined in the section of Knowledge Excavation.

Feature Aggregation and Prediction After extracting features using either the shared backbone (ϕ_{sha}) or independent backbones (ϕ_{ind}), each modality feature f_M is first processed by Batch Normalization (BN) layer BN_M and then fed into classifier CLS_M to produce the respective modality identity prediction:

$$f'_M = BN_M(f_M), \quad p_M = CLS_M(f'_M). \tag{1}$$

To derive unified global representation encompassing all modalities, we construct the fused feature by concatenating all modality features along the channel dimension:

$$f_{cat} = \text{Concat}(f_R, f_N, f_T), \tag{2}$$

which is normalized via BN and subsequently fed into the classifier:

$$f'_{cat} = BN_{cat}(f_{cat}), \quad p_{cat} = CLS_{cat}(f'_{cat}). \tag{3}$$

In contrast, for the test phase, which focuses on feature retrieval, the identity prediction step is omitted; we directly utilize the global feature f_{cat} to compute similarity scores for final retrieval evaluation.

Loss Optimization During each distillation stage, the target model is trained with the combination of standard ReID losses and hardness-aware knowledge distillation losses, applied to each respective modality representation or their fused global representation.

The target models' training aims to optimize both their respective and fusion discriminative power using ReID losses. The ReID loss is the combination of Cross-Entropy loss (ℓ_{CE}) for identity prediction and Triplet loss (ℓ_{Tri}) for metric learning as follows:

$$\begin{aligned} L_{reid,M} &= \ell_{CE}(p_M, y_i) + \ell_{Tri}(f_M, y_i), \\ L_{reid,cat} &= \ell_{CE}(p_{cat}, y_i) + \ell_{Tri}(f_{cat}, y_i), \end{aligned} \tag{4}$$

where f_M and f_{cat} denote the modality features and the fusion spectral global feature. Respectively, p_M and p_{cat} represent the identity predictions for each modality or the final fusion. The total ReID loss L_{reid} is the sum of these losses.

Then, we independently apply hardness-aware knowledge distillation losses on each modality. The loss of distillation measures the discrepancy between the features extracted by the target model and the source model, giving higher weights to the harder samples. The per-modality distillation loss for a batch of samples is defined as:

$$\ell_{hd} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{D}\left\|f_i^t - f_i^s\right\|_2^2 \times w_i, \tag{5}$$

where f_i^t and f_i^s denote the target and source feature for each modality M and sample i, and w_i is the dynamically assigned weight, which is higher for harder samples and lower for simpler ones, ensuring more effective knowledge transfer for challenging instances. The details of this process will be described in the next section. And the total distillation loss L_{dis} is the sum of per-modality distillation losses.

Finally, the overall training objective for the target model in the distillation stage is the sum of the total ReID loss and the total Hardness-aware KD loss:

$$L = L_{reid} + L_{dis}. \tag{6}$$

Hardness-aware Distillation Loss

Although feature-based knowledge distillation is commonly used to facilitate knowledge transfer, its uniform treatment of all samples can cause the target model to overfit to the source model, resulting in limited generalization. To overcome this, we introduce the Hardness-Aware Distillation Loss that adaptively focuses on more informative hard samples via a dynamic weighting mechanism.

We begin by defining the base distillation loss. For any given sample i and modality M, the feature-based mean squared error is defined as:

$$\ell_{mse} = \frac{1}{D} \left\| f_i^t - f_i^s \right\|_2^2, \tag{7}$$

where D is the feature dimension and $\|.\|$ is the mean squared error. To adaptively modulate the importance of each sample during distillation, we introduce the hardness-aware weight w_i:

$$w_i = \rho \times h_R + \tau, \tag{8}$$

where τ is the base weight that ensures some level of distillation for all samples, and ρ is the modulation factor controlling the influence of sample hardness. The relative hardness h_R of sample i is defined as its deviation from the batch average:

$$h_R = h_i - \bar{h}, \tag{9}$$

where h_i is the hardness metric for sample i, calculated based on in-batch triplet distances from the target model's feature f_i^t, and $\bar{h}$ denotes the average of the hardness metrics across all samples in the batch:

$$h_i = d_{ap_i} - d_{an_i},\ \bar{h} = \frac{1}{B} \sum_{i=1}^{B} h_i, \tag{10}$$

where B denotes the batch size, d_{ap_i} and d_{an_i} denote the Euclidean distances from the anchor feature f_i^t to its hardest positive and hardest negative samples within the batch, respectively. The hardest samples are selected based on pairwise distance maps:

$$\begin{aligned} d_{ap_i} &= \max_{p \in \mathcal{P}(i)} ||f_i^t - f_p^t||_2, \\ d_{an_i} &= \min_{n \in \mathcal{N}(i)} ||f_i^t - f_n^t||_2, \end{aligned} \tag{11}$$

where $\mathcal{P}(i)$ and $\mathcal{N}(i)$ denote the in-batch sets of positive and negative samples of anchor i. The final hardness-aware distillation loss is calculated as follows:

$$\ell_{hd} = \frac{1}{B} \sum_{i=1}^{B} \frac{1}{D} \left\| f_i^t - f_i^s \right\|_2^2 \times w_i. \tag{12}$$

The overall distillation loss L_{dis} is obtained by summing ℓ_{hd} across all modalities.

Experiments

Datasets and Evaluation Protocols

Datasets. We evaluate the proposed method on three widely recognized multi-spectral object ReID benchmarks: RGBNT201 (Zheng et al. 2021), a multi-spectral person ReID dataset comprising 4,787 aligned RGB, NIR, and TIR images from 201 identities; RGBNT100 (Li et al. 2020), a large-scale multi-spectral vehicle ReID dataset with 17,250 image triples covering diverse challenging conditions; and WMVeID863 (Zheng et al. 2025), another multi-spectral vehicle ReID dataset comprising 4,709 image triples collected in complex environments, featuring challenges such as drastic background variations and intense glare.

Evaluation Protocols. To comprehensively assess performance, we utilize standard ReID metrics: mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-K ($K = 1, 5, 10$).

Implementation Details

Our model is implemented using PyTorch (Paszke et al. 2019) and trained on an NVIDIA 5090 GPU. We utilize Transformer architectures pre-trained with DINO-v2 (Oquab et al. 2023) as backbones. All images within triples are resized to 224×224 for RGBNT201, RGBNT100, and WMVeID863 datasets. For data augmentation, we apply standard techniques including random horizontal flipping, cropping, and erasing (Zhong et al. 2020). For all datasets, we use a mini-batch size of 32, where each mini-batch comprises 4 randomly sampled object identities, with 8 images per identity. The proposed network is fine-tuned using the Adam optimizer, with an initial learning rate of $4.5e^{-5}$. Training is conducted for 50 epochs.

Comparison with State-of-the-art Methods

Experiments on person ReID datasets. We compare our proposed method with several approaches on the RGBNT201 dataset as presented in Table 1, which can be broadly divided into two categories: (i) **modal-specific models**, which adopt independent backbones for each modality to preserve modal-specific cues, and (ii) **multi-modal models**, which leverage shared backbone architectures to enhance multi-modal collaborative understanding. Among those modal-specific models, such as HAMNet, IEEE, and TOP-ReID, perform well, with TOP-ReID achieving an impressive 72.3% mAP, which represents the best performance reported to date for methods initialized with ViT-pretrained weights. However, their ability to integrate multi-modal understanding is limited by their independent backbone designs. On the other hand, multi-modal models (Crawford et al. 2023; Wang et al. 2024c; Zhang

Methods	mAP	R-1	R-5	R-10
HAMNet (AAAI20)	27.7	26.3	41.5	51.7
PFNet (AAAI21)	38.5	38.9	52.0	58.4
IEEE (AAAI22)	47.5	44.4	57.1	63.6
LRMM (ESWA25)	52.3	53.4	64.6	73.2
DENet (TNNLS23)	42.4	42.2	55.3	64.5
TIENet (TNNLS25)	54.4	54.4	66.2	71.0
UniCat* (NIPSW23)	57.0	55.7	-	-
RSCNet* (TCSVT24)	68.2	72.5	-	-
HTT* (AAAI24)	71.1	73.4	83.1	87.3
TOP-ReID* (AAAI24)	72.3	76.6	84.7	89.4
EDITOR* (CVPR24)	66.5	68.3	81.1	88.2
WTSF-ReID* (ESWA25)	67.9	72.2	83.4	89.7
ICPL-ReID† (TMM25)	75.1	77.4	84.2	87.9
PromptMA† (TIP25)	78.4	80.9	87.0	88.9
MambaPro† (AAAI25)	78.9	83.4	89.8	91.9
DeMo† (AAAI25)	79.0	82.3	88.8	92.0
DeMo‡ (AAAI25)	81.8	84.4	89.7	92.1
IDEA† (CVPR25)	80.2	82.1	90.0	**93.3**
IDEA‡ (CVPR25)	82.0	85.0	91.0	93.0
PMKD‡	**84.7**	**88.9**	**91.0**	92.2

Table 1: Performance comparison on RGBNT201 (in %). The symbol ‡ denotes DINOv2-based methods, the symbol † denotes CLIP-based methods, and the symbol * indicates ViT-based methods, and others are CNN-based methods.

et al. 2024; Yu et al. 2025b,a; Li et al. 2025; Zhang et al. 2025b; Wang et al. 2025a,b,c) such as ICPL, DeMo, and IDEA, benefit from deeper multi-modal collaboration, resulting in significant performance gains. Among these, to ensure fair comparison with our approach, we reproduced the results of DeMo and IDEA using DINOv2 (Oquab et al. 2023) pre-trained weights (marked with ‡). Our PMKD surpasses all existing methods, establishing new state-of-the-art results and outperforms IDEA‡ by 2.7% mAP and 3.9% R-1. This improvement underscores the multi-modal understanding and modal-specific details preservation achieved by our progressive multi-modal knowledge distillation framework, establishing PMKD as the new state-of-the-art in multi-spectral object ReID.

Experiments on vehicle ReID datasets. To further verify the robustness and generalizability of PMKD, we also evaluate it on two challenging multi-spectral vehicle ReID datasets: RGBNT100 and WMVEID863, with the results summarized in Table 2. Following the same categorization discussed in the person ReID experiments, the listed methods include both modal-specific and multi-modal models. A similar trend is observed: multi-modal models demonstrate stronger robustness in complex vehicle scenes by effectively exploiting multi-modal understanding, whereas modal-specific models, although capable of preserving modal-specific cues, perform poorly under such challenging conditions. Building on this observation, PMKD delivers consistent and substantial improvements across both datasets. Compared with the strongest prior methods, PMKD narrows the failure cases observed in challenging samples and establishes the state-of-the-art. These results further highlight that our design principles—balancing collaborative multi-modal understanding with fine-grained modality details, which generalize effectively to the vehicle ReID domain.

Methods	*RGBNT100*		*WMVEID863*	
	mAP	R-1	mAP	R-1
HAMNet (AAAI20)	74.5	93.3	45.6	48.5
PFNet (AAAI21)	68.1	94.1	50.1	55.9
IEEE (AAAI22)	61.3	87.8	45.9	48.6
CCNet (INFFUS23)	77.2	96.3	50.3	52.7
TOP-ReID* (AAAI24)	81.2	96.4	67.7	75.3
EDITOR* (CVPR24)	82.1	96.4	65.6	73.8
FACENet* (INFFUS25)	81.5	96.9	69.8	77.0
ICPL-ReID† (TMM25)	87.0	**98.6**	67.2	74.0
PromptMA† (TIP25)	85.3	97.4	-	-
MambaPro† (AAAI25)	83.9	94.7	69.5	76.9
DeMo† (AAAI25)	86.2	97.6	68.8	77.2
IDEA† (CVPR25)	87.2	96.5	-	-
PMKD‡	**91.6**	98.0	**71.9**	**79.5**

Table 2: Experimental results of our method on multi-spectral vehicle ReID datasets (in %).

Strategy					Metrics			
MM	MS	MKT	MKR	HDL	mAP	R-1	R-5	R-10
✓	×	×	×	×	82.5	83.7	88.8	91.7
×	✓	×	×	×	81.5	82.7	88.8	91.3
✓	✓	✓	×	×	84.3	86.4	90.3	92.1
✓	✓	✓	✓	×	84.6	87.7	90.4	92.1
✓	✓	✓	✓	✓	**84.7**	**88.9**	**91.0**	**92.2**

Table 3: Ablation study of PMKD on RGBNT201 (in %).

Ablation Study

To validate the contribution of each component in our Progressive Multi-modal Knowledge Distillation (PMKD) framework, we conduct the detailed ablation study on the RGBNT201 dataset. The results are presented in Table 3. We begin by examining two fundamental baselines: The Multi-Modal knowledge model (MM) and the Modal-Specific knowledge model (MS), which represent the trade-off between multi-modal collaborative understanding and the preservation of modal-specific details.

(1) **Effect of Multi-modal Knowledge Transfer (MKT).** Introducing the MKT module leads to a clear performance gain over the baseline models. Compared to MS alone, adding MKT improves mAP by +2.8% and R-1 by +3.7%. This demonstrates the benefit of distilling rich multi-modal knowledge into the target model. The enhanced multi-modal understanding empowers the target model to resolve challenging samples that it previously failed to handle effectively. (2) **Effect of Modal-specific Knowledge Retention (MKR).** Adding the second-stage MKR further enhances

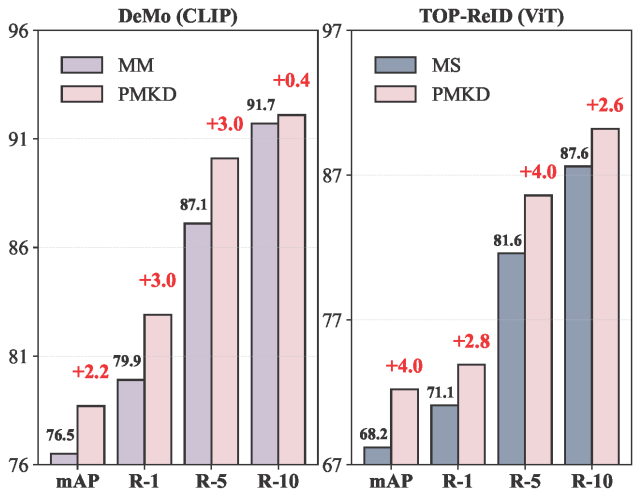

Figure 3: Performance of applying PMKD to typical MM (Multi-Modal knowledge model) and MS (Modal-Specific knowledge model) on RGBNT201 (in %). All results are reproduced for fair comparison.

performance, yielding an improvement of $+0.3\%$ in mAP and $+1.3\%$ in R-1. These results highlight the importance of modal-specific knowledge in preserving fine-grained discriminative cues and mitigating the capability drift introduced by the initial-stage heterogeneous distillation. (3) **Effect of Hardness-aware Distillation Loss (HDL).** Incorporating HDL results in the most significant R-1 improvement with an additional $+1.2\%$ gain, highlighting the importance of emphasizing hard samples during distillation. This helps the model focus on challenging instances, leading to more discriminative feature representations. Overall, these results confirm that each component of PMKD contributes synergistically to the more robust and discriminative representation, leading to superior accuracy.

Discussion on PMKD

To further validate the robustness and applicability of our proposed Progressive Multi-modal Knowledge Distillation (PMKD), we apply it to two representative multi-spectral ReID methods that reflect different knowledge excavation paradigms: TOP-ReID (Wang et al. 2024a), a Modal-Specific (MS) knowledge model that employs ViT-based (Dosovitskiy et al. 2020) independent backbones to preserve modal-specific cues, and DeMo (Wang et al. 2025b), a Multi-Modal (MM) knowledge model that leverages shared CLIP (Radford et al. 2021) encoders to capture multi-modal collaborative understanding. As shown in Figure 3, PMKD consistently improves performance, demonstrating its superiority across different knowledge excavation paradigms. Moreover, since TOP-ReID and DeMo are respectively initialized with ViT and CLIP pretraining, the results also highlight the robustness of PMKD across different pretraining paradigms. These results further demonstrate the generalizability of our progressive distillation strategy, which effectively integrates multi-modal synergy and modal-specific discriminative cues to address hard samples and break through performance bottlenecks.

Methods	mAP	R-1	R-5	R-10
Decoupling-based distillation				
Zhao et al. (2022)	82.6	87.2	90.7	**93.2**
Bias-mitigation distillation				
Zhang et al. (2025a)	76.4	77.6	84.8	88.8
Hardness-aware distillation				
HDM (Ours)	**84.7**	**88.9**	**91.0**	92.2

Table 4: Comparison of knowledge distillation strategies on RGBNT201 (in %).

Discussion on HDM

To further demonstrate the advantage of our Hardness-aware Distillation Mechanism (HDM), we compare it with representative knowledge distillation strategies that consider hard samples: decoupling-based distillation (e.g., DKD (Zhao et al. 2022)) and bias-mitigation distillation (e.g., BTKD (Zhang et al. 2025a)). The detailed results are shown in Table 4. Compared to DKD, our HDM achieves the 2.1% improvement in mAP and the 1.7% gain in R-1, demonstrating its superior ability to discriminate hard samples under the open-set ReID setting. It also significantly surpasses BTKD, further validating the effectiveness of our hardness-aware distillation strategy. Notably, unlike prior methods that rely on the source model's logits to define sample hardness, HDM dynamically estimates hardness based on the target model's own learning status, enabling more adaptive and context-aware supervision. This target-centric formulation is particularly well-suited to ReID, where stable and robust retrieval depends on the model's ability to adapt its learning strategy to its own state rather than rigidly following source-driven signals.

Conclusion

In this paper, we worked on the challenge of balancing multi-modal synergy with preserving modal-specific details in multi-spectral object ReID. We propose the Progressive Multi-modal Knowledge Distillation (PMKD), a novel approach that effectively integrates modal-specific knowledge with multi-modal knowledge to better handle hard samples and overcome performance bottlenecks. PMKD transfers multi-modal collaborative understanding to modal-specific models through Multi-modal Knowledge Transfer, while the Modal-specific Knowledge Retention stage employs architecture-consistent distillation to mitigate capability drift. Moreover, the Hardness-aware Distillation Loss focuses learning on challenging instances while avoiding overfitting, ensuring stable knowledge transfer. Experiments on three benchmark datasets demonstrate PMKD's state-of-the-art performance and generalizability. In the future, we will explore real-time capability evaluation to select the optimal model to further improve the final model's robustness and adaptability to unseen data.

Acknowledgements

This research is supported in part by the National Natural Science Foundation of China under Grants 62372003, the Natural Science Foundation of Anhui Province under Grants 2308085Y40, and the Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University (Grant No. 2024A004).

References

Chen, J.; Gao, C.; Sun, L.; and Sang, N. 2023. Ccsd: cross-camera self-distillation for unsupervised person re-identification. *VI*, 1(1): 27.

Crawford, J.; Yin, H.; McDermott, L.; and Cummings, D. 2023. Unicat: Crafting a stronger fusion baseline for multimodal re-identification. *arXiv preprint arXiv:2310.18812*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Hao, Z.; Guo, J.; Han, K.; Tang, Y.; Hu, H.; Wang, Y.; and Xu, C. 2023. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *NIPS*, 36: 79570–79582.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *ICCV*, 15013–15022.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Huang, Y.; Hu, K.; Zhang, Y.; Chen, Z.; and Gao, X. 2025. Distilling Knowledge from Heterogeneous Architectures for Semantic Segmentation. In *AAAI*, 3824–3832.

Jia, Z.; Sun, S.; Liu, G.; and Liu, B. 2024. Mssd: multi-scale self-distillation for object detection. *VI*, 2(1): 8.

Li, H.; Li, C.; Zhu, X.; Zheng, A.; and Luo, B. 2020. Multi-spectral vehicle re-identification: A challenge. In *AAAI*, 11345–11353.

Li, S.; Li, C.; Zheng, A.; Tang, J.; and Luo, B. 2025. ICPL-ReID: Identity-Conditional Prompt Learning for Multi-Spectral Object Re-Identification. *arXiv preprint arXiv:2505.17821*.

Li, S.; Sun, L.; and Li, Q. 2023. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *AAAI*, 1405–1413.

Liu, X.; Yu, C.; Zhang, P.; and Lu, H. 2023. Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification. *TNNLS*, 35(10): 13753–13763.

Liu, X.; Zhang, P.; Yu, C.; Lu, H.; and Yang, X. 2021. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, 13334–13343.

Liu, X.; Zhang, P.; Yu, C.; Qian, X.; Yang, X.; and Lu, H. 2024. A video is worth three views: Trigeminal transformers for video-based person re-identification. *TITS*, 25: 12818–12828.

Lu, H.; Zou, X.; and Zhang, P. 2023. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *AAAI*, 1835–1843.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *CVPR*, 3967–3976.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NIPS*.

Qiu, L.; Chen, S.; Yan, Y.; Xue, J.-H.; Wang, D.-H.; and Zhu, S. 2023. High-Order Structure Based Middle-Feature Learning for Visible-Infrared Person Re-Identification. arXiv:2312.07853.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Ren, K.; and Zhang, L. 2024. Implicit discriminative knowledge learning for visible-infrared person re-identification. In *CVPR*, 393–402.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Shi, J.; Yin, X.; Chen, Y.; Zhang, Y.; Zhang, Z.; Xie, Y.; and Qu, Y. 2024a. Multi-Memory Matching for Unsupervised Visible-Infrared Person Re-Identification. *arXiv preprint arXiv:2401.06825*.

Shi, J.; Yin, X.; Zhang, D.; Zhang, Z.; Xie, Y.; and Qu, Y. 2025. Two-stage knowledge distillation for visible-infrared person re-identification. *PR*, 169: 111850.

Shi, J.; Yin, X.; Zhang, Y.; Xie, Y.; Qu, Y.; et al. 2024b. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. *NIPS*, 99715–99734.

Shi, J.; Zhang, Y.; Yin, X.; Xie, Y.; Zhang, Z.; Fan, J.; Shi, Z.; and Qu, Y. 2023. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *ICCV*, 11218–11228.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Tung, F.; and Mori, G. 2019. Similarity-preserving knowledge distillation. In *ICCV*, 1365–1374.

Wang, L.; and Yoon, K.-J. 2021. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *TPAMI*, 44(6): 3048–3068.

Wang, Y.; Liu, X.; Yan, T.; Liu, Y.; Zheng, A.; Zhang, P.; and Lu, H. 2025a. Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt. In *AAAI*, 8150–8158.

Wang, Y.; Liu, X.; Zhang, P.; Lu, H.; Tu, Z.; and Lu, H. 2024a. Top-reid: Multi-spectral object re-identification with token permutation. In *AAAI*, 5758–5766.

Wang, Y.; Liu, Y.; Zheng, A.; and Zhang, P. 2025b. Decoupled feature-based mixture of experts for multi-modal object re-identification. In *AAAI*, 8141–8149.

Wang, Y.; Lv, Y.; Zhang, P.; and Lu, H. 2025c. Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In *CVPR*, 29701–29710.

Wang, Y.; Zhang, P.; Wang, D.; and Lu, H. 2024b. Other tokens matter: Exploring global and local features of Vision Transformers for Object Re-Identification. *CVIU*, 244: 104030.

Wang, Z.; Huang, H.; Zheng, A.; and He, R. 2024c. Heterogeneous test-time training for multi-modal person re-identification. In *AAAI*, 5850–5858.

Wang, Z.; Li, C.; Zheng, A.; He, R.; and Tang, J. 2022. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In *AAAI*, 2633–2641.

Yu, C.; Liu, X.; Wang, Y.; Zhang, P.; and Lu, H. 2024. TF-CLIP: Learning text-free CLIP for video-based person re-identification. In *AAAI*, 6764–6772.

Yu, Z.; Huang, Z.; Hou, M.; Pei, J.; Yan, Y.; Liu, Y.; and Sun, D. 2025a. Representation selective coupling via token sparsification for multi-spectral object re-identification. *TCSVT*, 35(4): 3633–3648.

Yu, Z.; Huang, Z.; Hou, M.; Yan, Y.; and Liu, Y. 2025b. WTSF-ReID: Depth-driven Window-oriented Token Selection and Fusion for multi-modality vehicle re-identification with knowledge consistency constraint. *ESA*, 274: 126921.

Zhang, G.; Zhang, Y.; Zhang, T.; Li, B.; and Pu, S. 2023. PHA: Patch-Wise High-Frequency Augmentation for Transformer-Based Person Re-Identification. In *CVPR*, 14133–14142.

Zhang, J.; Gao, Y.; Liu, R.; Cheng, X.; Zhang, H.; and Chen, S. 2025a. Can students beyond the teacher? distilling knowledge from teacher's bias. In *AAAI*, 22434–22442.

Zhang, P.; Wang, Y.; Liu, Y.; Tu, Z.; and Lu, H. 2024. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *CVPR*, 17117–17126.

Zhang, S.; Luo, W.; Cheng, D.; Xing, Y.; Liang, G.; Wang, P.; and Zhang, Y. 2025b. Prompt-based modality alignment for effective multi-modal object re-identification. *TIP*, 34: 2450–2462.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *CVPR*, 11953–11962.

Zheng, A.; Ma, Z.; Sun, Y.; Wang, Z.; Li, C.; and Tang, J. 2025. Flare-aware cross-modal enhancement network for multi-spectral vehicle Re-identification. *INFFUS*, 116: 102800.

Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; and Tang, J. 2021. Robust Multi-Modality Person Re-identification. In *AAAI*, 3529–3537.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *AAAI*, 13001–13008.