DeMo: Decoupled Feature-Based Mixture of Experts for Multi-Modal Object Re-Identification

Yuhao Wang[1], **Yang Liu**[1,3], **Aihua Zheng**[2,3], **Pingping Zhang**[1,3*]

[1]School of Future Technology, School of Artificial Intelligence, Dalian University of Technology
[2]School of Artificial Intelligence, Anhui University
[3]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University
924973292@mail.dlut.edu.cn, {ly, zhpp}@dlut.edu.cn, ahzheng214@foxmail.com

Abstract

Multi-modal object Re-IDentification (ReID) aims to retrieve specific objects by combining complementary information from multiple modalities. Existing multi-modal object ReID methods primarily focus on the fusion of heterogeneous features. However, they often overlook the dynamic quality changes in multi-modal imaging. In addition, the shared information between different modalities can weaken modality-specific information. To address these issues, we propose a novel feature learning framework called DeMo for multi-modal object ReID, which adaptively balances decoupled features using a mixture of experts. To be specific, we first deploy a Patch-Integrated Feature Extractor (PIFE) to extract multi-granularity and multi-modal features. Then, we introduce a Hierarchical Decoupling Module (HDM) to decouple multi-modal features into non-overlapping forms, preserving the modality uniqueness and increasing the feature diversity. Finally, we propose an Attention-Triggered Mixture of Experts (ATMoE), which replaces traditional gating with dynamic attention weights derived from decoupled features. With these modules, our DeMo can generate more robust multi-modal features. Extensive experiments on three object ReID benchmarks verify the effectiveness of our methods.

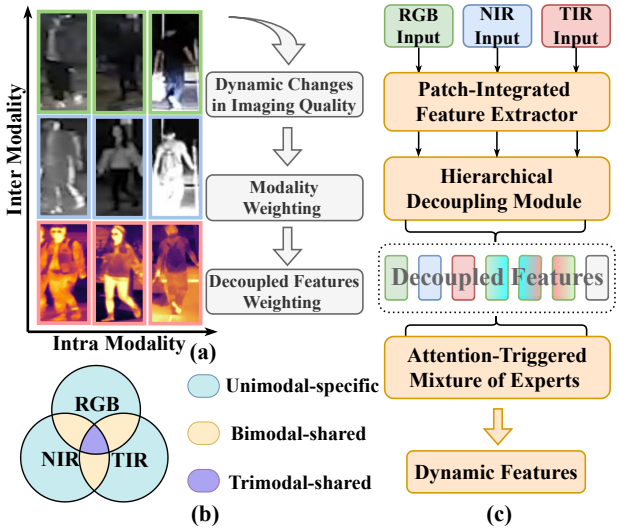

Figure 1: (a) The prevalent dynamic quality changes in multi-modal imaging. (b) Hierarchical feature decoupling. (c) The proposed modules and the framework of our DeMo.

Introduction

Object Re-IDentification (ReID) aims to retrieve the same object across different camera views. Over the past decade, single-modal object ReID (Liu et al. 2023; Wang et al. 2024a; Liu et al. 2024b, 2021; Zhang et al. 2021; Yu et al. 2024a), primarily based on RGB images, has made significant progress. However, RGB imaging is highly susceptible to adverse conditions such as darkness and glare, leading to poor generalization in complex environments. Fortunately, multi-modal imaging, which introduces diverse information from different modalities (Chen et al. 2024; Shi et al. 2024a, 2023, 2024b; Zheng et al. 2021; Lu, Zou, and Zhang 2023), has emerged as a promising solution to enhance the feature robustness in challenging scenarios. By integrating complementary information from different modalities, existing multi-modal object ReID methods (Zhang et al. 2024a; Yang, Chen, and Ye 2024, 2023) achieve remarkable performance. However, they often overlook dynamic quality changes inherent in multi-modal imaging. As shown in Fig. 1 (a), images from RGB, Near Infrared (NIR) and Thermal Infrared (TIR) modalities are presented, respectively. Horizontally, environmental and imaging interferences lead to content fluctuations within the same modality. Vertically, the relevance of each modality varies under identical imaging conditions. For example, in the second column, the RGB image provides limited information due to darkness, whereas the NIR and TIR images clearly show details such as glasses and bags. Consequently, the model should prioritize the NIR and TIR modalities in such scenarios. This highlights the need for adaptive modality weighting to address the dynamic changes in multi-modal imaging quality.

Recently, Mixture of Experts (MoE) (Cai et al. 2024) have gained attention for their effectiveness in adaptively weighting expert features. Inspired by this, we treat each modality as an expert and dynamically adjust the weights of the experts based on the importance of each modality. However, directly weighting multi-modal features risks modality information confusion (Zhang et al. 2024b). As depicted

*Corresponding author (zhpp@dlut.edu.cn).

in Fig. 1 (b), multi-modal features can be categorized into three types: modality-specific, bimodal-shared and trimodal-shared features. Directly weighting features from RGB, NIR and TIR modalities can amplify modality-shared information while suppressing modality-specific information. Fortunately, decoupling multi-modal features can effectively separates modality-specific information (Wei et al. 2024), preserving discriminative details. Motivated by these observations, we propose DeMo, a novel feature learning framework that adaptively assigns weights to decoupled features with a mixture of experts for multi-modal object ReID.

As shown in Fig. 1 (c), our DeMo consists of three components: Patch-Integrated Feature Extractor (PIFE), Hierarchical Decoupling Module (HDM) and Attention-Triggered Mixture of Experts (ATMoE). We first utilize the PIFE to extract multi-granularity representations from multi-modal inputs. Specifically, the PIFE integrates high-level patch tokens with class tokens to generate robust features. This synergy between global and local information significantly enhances the feature discrimination for each modality. Then, we introduce the HDM to guide the hierarchical decoupling of multi-modal features. We first categorize these features into three hierarchical types based on their degree of information overlap. By utilizing learnable queries with different combinations of multi-modal tokens, HDM employs cross-attentions to effectively decouple features into corresponding hierarchical levels, preserving modality-specific information and enhancing feature diversity. Finally, we introduce the ATMoE to replace traditional gating with attention-guided interactions between decoupled features, allowing for more accurate and context-aware weighting of each expert. Moreover, the multi-head mechanism in ATMoE enhances the model's adaptability to dynamic imaging conditions. With the above modules, our DeMo can extract robust representations across various scenarios. Even in extreme cases where one or more modalities are missing, our DeMo can still achieve competitive performances. Extensive experiments on three multi-modal object ReID datasets validate the effectiveness of our proposed method.

In summary, our contributions are as follows:

- We introduce DeMo, a novel framework for multi-modal object ReID. To our best knowledge, our proposed DeMo is the first attempt to address dynamic changes in multi-modal imaging with decoupled feature-based MoE.
- We develop a Hierarchical Decoupling Module (HDM) to effectively decouple multi-modal features into hierarchical types, increasing the decoupled features' diversity.
- We propose an Attention-Triggered Mixture of Experts (ATMoE), which utilizes attention-guided interactions for accurate expert weighting and a multi-head mechanism for adaptability to dynamic imaging conditions.
- Extensive experiments on three multi-modal object ReID datasets demonstrate the effectiveness of our method.

Related Work

Multi-Modal Object Re-Identification

Multi-modal object ReID has attracted increasing attention due to its robustness in practical scenarios. Existing methods primarily focus on integrating complementary information from different modalities. For multi-modal person ReID, Zheng *et al.* (Zheng et al. 2021) propose to learn robust features with a progressive fusion. Wang *et al.* (Wang et al. 2022) introduce an interact-embed-enlarge framework to boost the modality-specific knowledge. In addition, Zheng *et al.* (Zheng et al. 2023) address the modal-missing problem with a pixel reconstruction method. For multi-modal vehicle ReID, Li *et al.* (Li et al. 2020) propose to fuse multi-modal features with a coherence loss. Afterwards, many CNN-based methods (He et al. 2023; Guo et al. 2022; Zheng et al. 2022) have been proposed to enhance the feature robustness with modality generation, graph learning and instance sampling, etc. With the strong generalization ability of vision Transformer (Dosovitskiy et al. 2020) (ViT), many Transformer-based methods (Pan et al. 2023; Crawford et al. 2023; Wang et al. 2023, 2024b; Zhang et al. 2024a) have been proposed to further improve the performance of multi-modal object ReID. Among them, Wang *et al.* (Wang et al. 2024b) propose to mine the modality interactions in test-time training. Recently, Zhang *et al.*. (Zhang et al. 2024a) propose to select diverse tokens and suppress the influence of backgrounds. Although these methods achieve remarkable performance, they often overlook the dynamic quality changes in multi-modal imaging. Meanwhile, they lack the ability to adaptively balance the multi-modal features based on instance characteristics. In contrast, our porposed DeMo can effectively address these issues by adaptively weighting decoupled features, enhancing the model's robustness.

Mixtures of Experts

The Mixture of Experts (MoE) (Jacobs et al. 1991) is designed to tackle complex tasks by combining the specialized knowledge of multiple experts. Recently, MoE has advanced significantly in various fields, including natural language processing (Dai et al. 2024), computer vision (Hwang et al. 2023; Chowdhury et al. 2023) and multi-modal learning (Li et al. 2024; Lin et al. 2024). In object ReID, MoE has been applied to domain generalizable ReID (Xu et al. 2022; Kuang et al. 2024) and unsupervised ReID (Li et al. 2023), but its potential in multi-modal object ReID remains unexplored. Meanwhile, many approaches (Chen and Wang 2024; Gui et al. 2024) directly apply MoE without explicitly decoupling the expert inputs, which may lead to feature entanglement and limit the effectiveness of MoE. In contrast, we perform hierarchical decoupling of multi-modal features, providing MoE with more flexible and specialized expert inputs. Besides, existing methods (Liu et al. 2024a) often rely on simple techniques to generate gating weights, which may not fully capture the intricate relationships between experts. To address this issue, we introduce ATMoE, which replaces traditional gating methods with attention-guided interactions between decoupled features. It ensures more accurate and context-aware expert weighting, enhancing the model's adaptability to dynamic imaging conditions.

Methodology

As shown in Fig. 2, our proposed DeMo is composed of three main components: Patch-Integrated Feature Extrac-

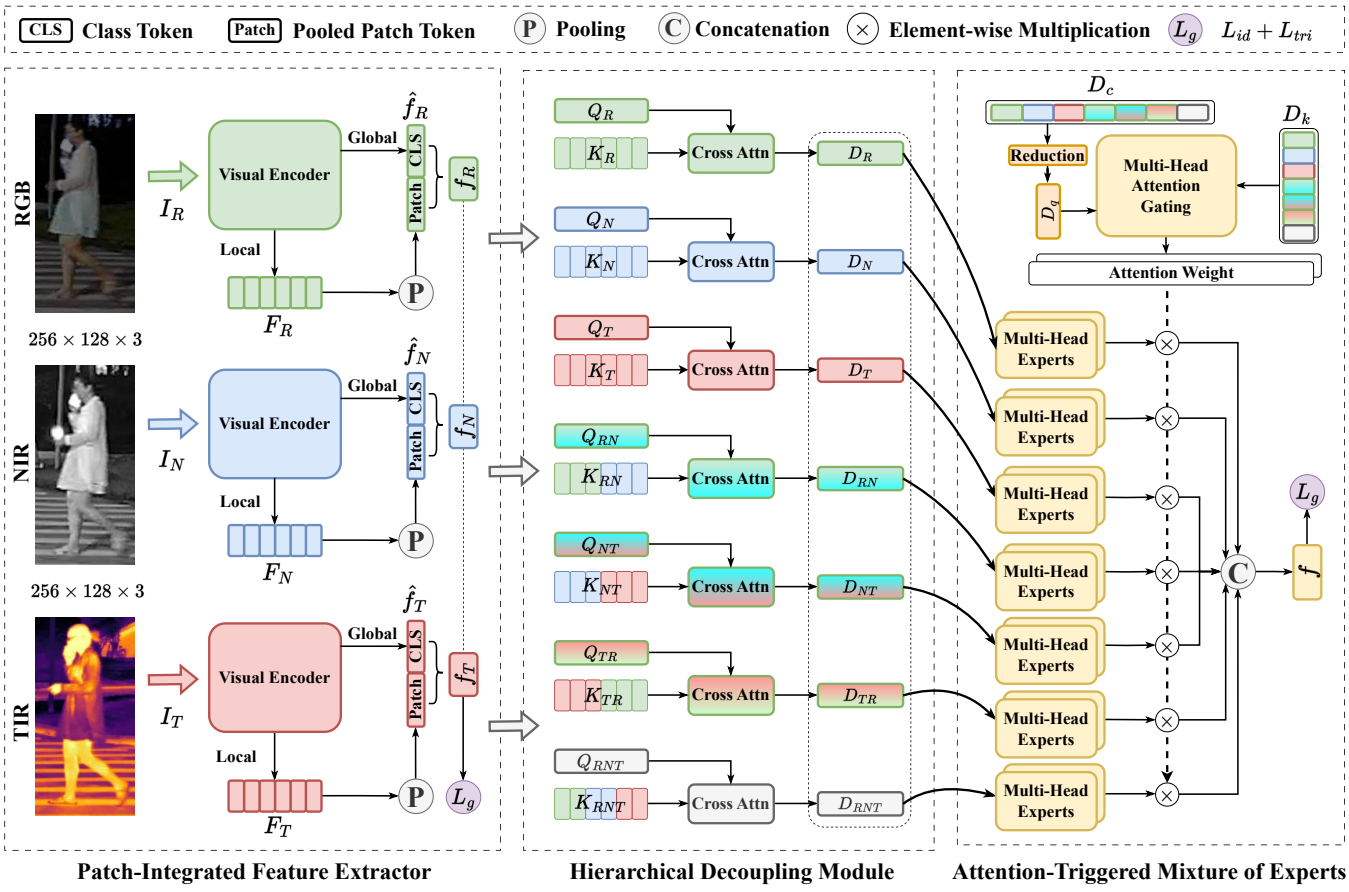

Figure 2: The overall framework of our DeMo. We first employ a Patch-Integrated Feature Extractor (PIFE) to extract multi-granularity features from each modality. Then, the Hierarchical Decoupling Module (HDM) decouples multi-modal features into different levels with learnable query tokens. Finally, the Attention-Triggered Mixture of Experts (ATMoE) adaptively balances the decoupled features with accurate and context-aware weights, generating robust multi-modal features.

tor (PIFE), Hierarchical Decoupling Module (HDM) and Attention-Triggered Mixture of Experts (ATMoE).

Patch-Integrated Feature Extractor

To fully extract discriminative information from different modalities, we propose a Patch-Integrated Feature Extractor (PIFE) to capture multi-granularity features from each modality. More specifically, the multi-modal inputs I_m ($m \in \{R, N, T\}$) are fed into the visual encoder Θ to produce patch tokens $F_m \in \mathbb{R}^{N_p \times C}$ and class token $\hat{f}_m \in \mathbb{R}^C$:

$$F_m, \hat{f}_m = \Theta(I_m). \tag{1}$$

Here, R, N and T represent the RGB, NIR and TIR modalities, respectively. N_p denotes the number of patch tokens and C is the embedding dimension. As shown in the left part of Fig. 2, we pool the patch tokens F_m and concatenate them with the corresponding class token $\hat{f}_m$. The concatenated features are then passed through a projection layer to obtain the modality-specific features f_m as follows:

$$f_m = \omega(W_{\text{pro}}(\text{LN}([\hat{f}_m, P(F_m)]))), \tag{2}$$

where $P(\cdot)$ is the average pooling operation and $[\cdot]$ means concatenation. $\text{LN}(\cdot)$ represents the layer normalization (Ba, Kiros, and Hinton 2016). $W_{\text{pro}} \in \mathbb{R}^{2C \times C}$ is the projection matrix. $\omega(\cdot)$ is the GELU activation function (Hendrycks and Gimpel 2016). By integrating global and local information, we obtain multi-granularity features for each modality, enhancing the subsequent multi-modal fusion.

Hierarchical Decoupling Module

Multi-modal features consist of both modality-specific and modality-shared information. Previous methods (Wang et al. 2023, 2024b; Zhang et al. 2024a) often neglect the mutual interference between modalities, leading to weakened modality-specific features and decreased diversity. To address these issues, we propose a Hierarchical Decoupling Module (HDM). As shown in Fig. 2, the decoupling process involves comprehensive interactions using cross-attentions. Specifically, the HDM can be divided into three processes: unimodal-specific, bimodal-shared and trimodal-shared feature decoupling. Details of each process are as follows.

Unimodal-specific Feature Decoupling. In the first three rows of HDM in Fig. 2, we show the unimodal-specific fea-

ture decoupling process. For each modality, we first initialize a learnable query token $Q_{m_1} \in \mathbb{R}^C$ and key tokens $K_{m_1} \in \mathbb{R}^{(N_p+1)\times C}$, where $m_1 \in \{R, N, T\}$. Here, key tokens K_{m_1} are constructed by concatenating the enhanced token and patch tokens from the corresponding modality m_1:

$$K_{m_1} = [f_{m_1}, F_{m_1}]. \tag{3}$$

Then, Q_{m_1} is used to interact with K_{m_1} to obtain the decoupled unimodal-specific feature D_{m_1} as follows:

$$D_{m_1} = \Phi(Q_{m_1}, K_{m_1}), \tag{4}$$

where Φ denotes the multi-head cross-attention mechanism (Vaswani et al. 2017). This approach leverages the learnable query token Q_{m_1} to dynamically focus on and highlight crucial modality-specific information, enabling a refined and context-aware extraction of unimodal features.

Bimodal-shared Feature Decoupling. As shown in the 4-6 rows of HDM in Fig. 2, we illustrate the bimodal-shared feature decoupling process. Similar to the unimodal-specific feature decoupling, we generate a learnable query token $Q_{m_2} \in \mathbb{R}^C$ and key tokens $K_{m_2} \in \mathbb{R}^{(2N_p+2)\times C}$ for paired modalities $m_2 \in \{RN, NT, TR\}$. Here, the key tokens K_{m_2} are constructed by concatenating all tokens from the corresponding paired modalities m_2 as follows:

$$K_{m_2} = [f_{m_2[0]}, F_{m_2[0]}, f_{m_2[1]}, F_{m_2[1]}], \tag{5}$$

where $m_2[0]$ and $m_2[1]$ represent the two modalities in the paired modalities m_2. If $m_2 = RN$, then $m_2[0] = R$ and $m_2[1] = N$. After that, we use Q_{m_2} to extract the decoupled bimodal-shared feature D_{m_2} from K_{m_2} as follows:

$$D_{m_2} = \Phi(Q_{m_2}, K_{m_2}). \tag{6}$$

Through this interaction, the learnable query token Q_{m_2} integrates discriminative information from paired modalities, enhancing the representation of modality-shared features.

Trimodal-shared Feature Decoupling. In the last row of HDM in Fig. 2, we show the trimodal-shared feature decoupling process. The only difference is that the key token $K_{m_3} \in \mathbb{R}^{(3N_p+3)\times C}$, where $m_3 = RNT$, is constructed by concatenating all tokens from three modalities as follows:

$$K_{m_3} = [f_{m_3[0]}, F_{m_3[0]}, f_{m_3[1]}, F_{m_3[1]}, f_{m_3[2]}, F_{m_3[2]}], \tag{7}$$

where $m_3[0]$, $m_3[1]$ and $m_3[2]$ represent the R, N and T modalities, respectively. Then, we use Q_{m_3} to extract the decoupled trimodal-shared feature D_{m_3} from K_{m_3} as:

$$D_{m_3} = \Phi(Q_{m_3}, K_{m_3}). \tag{8}$$

If there are highly shared discriminative regions among three modalities, the query token will prioritize these regions, assigning higher weights to them. Thus, cross-attention helps D_{m_3} better capture shared discriminative information.

Finally, we obtain the decoupled features D_{m_1}, D_{m_2} and D_{m_3} for unimodal-specific, bimodal-shared and trimodal-shared information, respectively. By separating modality-specific and modality-shared information, the model can prevent interferences between modalities, preserving each modality's unique strengths and enhancing feature diversity. Additionally, it provides the MoE with more options to select the most suitable experts under varying imaging conditions, thereby improving the model's generalization ability.

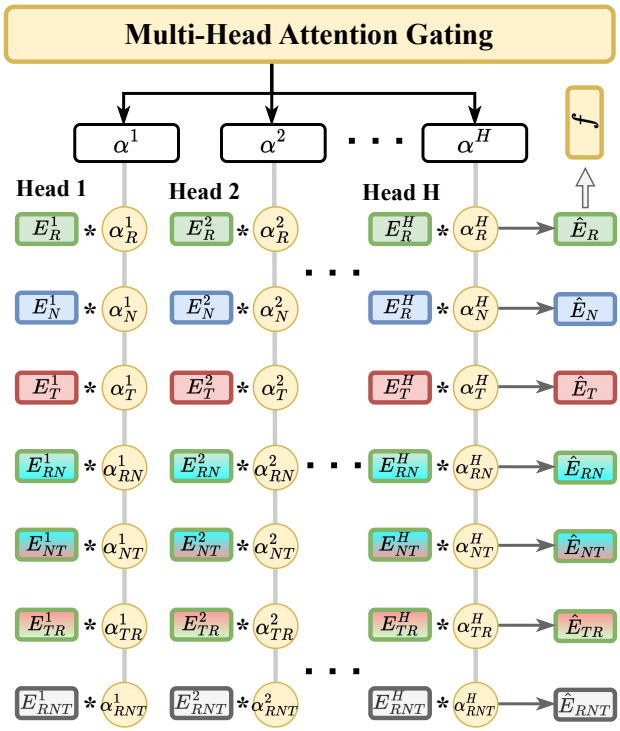

Figure 3: Detailed structure of ATMoE.

Attention-Triggered Mixture of Experts

To address the dynamic imaging quality and appropriately balance decoupled features across different instances, we introduce an Attention-Triggered Mixture of Experts (ATMoE). Unlike traditional MoE (Liu et al. 2024a) where weights are directly generated from decoupled features, we incorporate an attention mechanism. With attention-guided interactions between the integrated feature and each decoupled feature, ATMoE can assign more accurate and context-aware weights to the decoupled experts. To be specific, as depicted in the top part of ATMoE in Fig. 2, we first concatenate the decoupled features and send them to a reduction layer φ to obtain the query $D_q \in \mathbb{R}^C$ as follows:

$$D_q = \varphi([D_R, D_N, D_T, D_{RN}, D_{NT}, D_{TR}, D_{RNT}]), \tag{9}$$

$$\varphi(\mathcal{X}) = \mathrm{BN}(\omega(W_{\mathrm{red}}(\mathcal{X}))), \tag{10}$$

where $W_{\mathrm{red}} \in \mathbb{R}^{n_d C\times C}$ is the reduction matrix and $\mathrm{BN}(\cdot)$ denotes the batch normalization (Ioffe and Szegedy 2015). Here, n_d represents the number of decoupled features. Meanwhile, we stack the decoupled features to generate the key tokens $D_k \in \mathbb{R}^{n_d\times C}$. Then, we project D_q and D_k to generate $Q \in \mathbb{R}^C$ and $K \in \mathbb{R}^{n_d\times C}$ as follows:

$$Q = W_q D_q, K = W_k D_k, \tag{11}$$

where $W_q \in \mathbb{R}^{C\times C}$ and $W_k \in \mathbb{R}^{C\times C}$ are the projection matrices. After that, we use the multi-head attention mechanism to generate weights $A \in \mathbb{R}^{H\times n_d}$ as follows:

$$A = \left[\alpha^1, \alpha^2, \cdots, \alpha^H\right], \tag{12}$$

where H is the number of heads and $\alpha^h \in \mathbb{R}^{n_d}$ is the attention weight for the h-th head. To be specific, the attention weight α^h is calculated as follows:

$$\alpha^h = \delta\left(\frac{Q^h K^{h\top}}{\sqrt{c}}\right). \tag{13}$$

Here, $\delta(\cdot)$ is the softmax function and $c = \frac{C}{H}$ is the dimension of the head. $Q^h \in \mathbb{R}^c$ and $K^h \in \mathbb{R}^{n_d \times c}$ are the query and key tokens for the h-th head, respectively.

As shown in Fig. 3, after sending each decoupled feature to the corresponding expert, we chunk the experts' output into H parts and multiply them with the corresponding attention weights. Without loss of generality, we take D_R as an example to obtain the expert output E_R as follows:

$$E_R = \mathrm{BN}(\omega(W_{\text{exp}}(D_R))), \tag{14}$$

where $W_{\text{exp}} \in \mathbb{R}^{C \times C}$ is the expert matrix. Then, we chunk E_R into H parts and multiply them with the corresponding attention weights to generate the weighted experts $\hat{E}_R$:

$$\hat{E}_R = [E_R^1 * \alpha_R^1, E_R^2 * \alpha_R^2, \cdots, E_R^H * \alpha_R^H], \tag{15}$$

where $E_R^h \in \mathbb{R}^{N_p \times c}$ is the h-th chunk of E_R. Similarly, we can obtain weighted experts for other decoupled features. Finally, we concatenate the outputs of all the weighted experts to form the final feature $f \in \mathbb{R}^{n_d C}$. Through the attention-guided interactions and the multi-head machanism, ATMoE can assign more accurate and context-aware weights to the decoupled experts, enhancing the feature robustness.

Objective Functions

As shown in Fig. 2, we optimize the model using multiple losses. Features after PIFE and ATMoE are supervised by the label smoothing cross-entropy loss (Szegedy et al. 2016) and triplet loss (Hermans, Beyer, and Leibe 2017) as:

$$\mathcal{L}_g(\mathcal{X}) = \mathcal{L}_{cross}(\mathcal{X}) + \mathcal{L}_{triplet}(\mathcal{X}), \tag{16}$$

where $\mathcal{X}$ represents input features for supervision. Finally, the overall loss $\mathcal{L}$ for our framework can be given by:

$$\mathcal{L} = \mathcal{L}_g([f_R, f_N, f_T]) + \mathcal{L}_g(f). \tag{17}$$

Experiments

Datasets and Evaluation Protocols

Datasets. We evaluate the proposed method on three multi-modal object ReID benchmarks. To be specific, RGBNT201 (Zheng et al. 2021) is a multi-modal person ReID dataset, consisting of 4,787 aligned RGB, NIR and TIR images from 201 identities. RGBNT100 (Li et al. 2020) is a large-scale multi-modal vehicle ReID dataset with 17,250 image triples, covering a wide range of challenging visual conditions. MSVR310 (Zheng et al. 2022) is a small-scale multi-modal vehicle ReID dataset with 2,087 image triples, featuring high-quality images captured across diverse environments and time spans.

Evaluation Protocols. We use the mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-K ($K = 1, 5, 10$) to assess performance and present trainable parameters and FLOPs for complexity analysis.

	Methods	mAP	R-1	R-5	R-10
Single	OSNet (Zhou et al. 2019)	25.4	22.3	35.1	44.7
	CAL (Rao et al. 2021)	27.6	24.3	36.5	45.7
	PCB (Sun et al. 2018)	32.8	28.1	37.4	46.9
Multi	HAMNet (Li et al. 2020)	27.7	26.3	41.5	51.7
	PFNet (Zheng et al. 2021)	38.5	38.9	52.0	58.4
	DENet (Zheng et al. 2023)	42.4	42.2	55.3	64.5
	IEEE (Wang et al. 2022)	47.5	44.4	57.1	63.6
	LRMM (Wu et al. 2025)	52.3	53.4	64.6	73.2
	UniCat* (Crawford et al. 2023)	57.0	55.7	-	-
	HTT* (Wang et al. 2024b)	71.1	73.4	83.1	87.3
	TOP-ReID* (Wang et al. 2023)	72.3	76.6	84.7	89.4
	EDITOR* (Zhang et al. 2024a)	66.5	68.3	81.1	88.2
	RSCNet* (Yu et al. 2024b)	68.2	72.5	-	-
	DeMo*	<u>73.7</u>	<u>80.5</u>	<u>88.3</u>	<u>91.5</u>
	DeMo†	**79.0**	**82.3**	**88.8**	**92.0**

Table 1: Performance comparison on RGBNT201. The best and second results are in bold and underlined, respectively. The symbol † denotes CLIP-based methods, $*$ indicates ViT-based methods and others are CNN-based methods.

	Methods	RGBNT100		MSVR310	
		mAP	R-1	mAP	R-1
Single	PCB (Sun et al. 2018)	57.2	83.5	23.2	42.9
	OSNet (Zhou et al. 2019)	75.0	95.6	28.7	44.8
	AGW (Ye et al. 2021)	73.1	92.7	28.9	46.9
	TransReID* (He et al. 2021)	75.6	92.9	18.4	29.6
Multi	GAFNet (Guo et al. 2022)	74.4	93.4	-	-
	GPFNet (He et al. 2023)	75.0	94.5	-	-
	PFNet (Zheng et al. 2021)	68.1	94.1	23.5	37.4
	HAMNet (Li et al. 2020)	74.5	93.3	27.1	42.3
	CCNet (Zheng et al. 2022)	77.2	96.3	36.4	<u>55.2</u>
	LRMM (Wu et al. 2025)	78.6	<u>96.7</u>	36.7	49.7
	PHT* (Pan et al. 2023)	79.9	92.7	-	-
	HTT* (Wang et al. 2024b)	75.7	92.6	-	-
	TOP-ReID* (Wang et al. 2023)	81.2	96.4	35.9	44.6
	EDITOR* (Zhang et al. 2024a)	82.1	96.4	39.0	49.3
	RSCNet* (Yu et al. 2024b)	82.3	96.6	<u>39.5</u>	49.6
	DeMo*	<u>82.4</u>	96.0	39.1	48.6
	DeMo†	**86.2**	**97.6**	**49.2**	**59.8**

Table 2: Performance on RGBNT100 and MSVR310.

Implementation Details

Our model is implemented using PyTorch with an NVIDIA A100 GPU. We use the pre-trained ViT (Dosovitskiy et al. 2020) or CLIP (Radford et al. 2021) as the visual encoder. The number of experts n_d is set to 7. Images in triples are resized to 256×128 for RGBNT201 and 128×256 for RGBNT100/MSVR310. For data augmentation, we apply random horizontal flipping, cropping and erasing (Zhong et al. 2020). For RGBNT201 and MSVR310, the mini-batch size is set to 64, sampling 8 images per identity. For RGBNT100, the mini-batch size is 128 with 16 images per identity. We fine-tune the proposed modules using the Adam optimizer with a learning rate of $3.5\mathrm{e}^{-4}$ and a smaller learning rate of $5\mathrm{e}^{-6}$ for the visual encoder. The total number of training epochs is 50. The detailed configurations and results are available at https://github.com/924973292/DeMo.

Methods	M (RGB)		M (NIR)		M (TIR)		M (RGB+NIR)		M (RGB+TIR)		M (NIR+TIR)		Average	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
TOP-ReID	54.4	57.5	64.3	67.6	51.9	**54.5**	35.3	35.4	26.2	**26.0**	34.1	31.7	44.4	45.4
DeMo	**63.3**	**65.3**	**72.6**	**75.7**	**56.2**	54.1	**45.6**	**46.5**	**26.3**	24.9	**40.3**	**38.5**	**50.7**	**50.8**

Table 3: Performance of missing-modality settings on RGBNT201. "M (X)" means missing the X image modality.

Methods	M (RGB)		M (NIR)		M (TIR)		M (RGB+NIR)		M (RGB+TIR)		M (NIR+TIR)		Average	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
TOP-ReID	70.6	90.6	77.9	94.5	64.0	81.5	42.5	69.3	45.9	65.4	55.4	77.8	59.4	79.9
DeMo	**81.0**	**94.5**	**84.1**	**96.5**	**71.1**	**87.6**	**50.2**	**73.7**	**59.6**	**78.1**	**66.3**	**82.8**	**68.7**	**85.5**

Table 4: Performance of missing-modality settings on RGBNT100.

Comparison with State-of-the-Art Methods

Multi-modal Person ReID. In Tab. 1, we compare our proposed DeMo with single-modal and multi-modal methods on RGBNT201. Generally, multi-modal methods significantly outperform single-modal methods by integrating complementary information from different modalities. Among them, models based on ViT and CLIP perform better than those based on CNNs. Specifically, DeMo* shows a 1.4% improvement in mAP and a 3.9% improvement in Rank-1 compared with TOP-ReID*. This highlights DeMo's effectiveness in dynamically integrating multi-modal information in complex environments. Additionally, DeMo† utilizes CLIP's pre-trained knowledge to improve the robustness of multi-modal features, achieving a 6.7% improvement in mAP and a 5.7% improvement in Rank-1 over TOP-ReID*. These results confirm DeMo's capability in managing dynamic changes in modality fusion.

Multi-modal Vehicle ReID. As shown in Tab. 2, TransReID* achieves an mAP of 75.6% on the large-scale RGBNT100 dataset. However, on the smaller MSVR310 dataset, it performs worse than AGW and OSNet, which are better suited for small-scale datasets. Among multi-modal methods, EDITOR* demonstrates significant improvements on both datasets. Our DeMo*, with its simpler structure, delivers competitive results and avoids the instability issues present in EDITOR*. Additionally, DeMo† achieves 86.2% mAP on RGBNT100, outperforming EDITOR* by 4.1%. On MSVR310, DeMo† exceeds EDITOR* and TOP-ReID* by over 10.2% in mAP and 10.5% in Rank-1. These results highlight DeMo's robustness in integrating multi-modal information across dynamic environments.

Multi-modal Object ReID with Missing Modalities. To assess DeMo's robustness in missing-modality scenarios, we conduct experiments on RGBNT201 and RGBNT100. As shown in Tab. 3 and Tab. 4, DeMo consistently outperforms TOP-ReID in these settings. Despite lacking specific designs like the reconstruction modules in TOP-ReID, DeMo achieves competitive performance through automatic feature weighting. In all missing-modality settings, DeMo achieves an average mAP of 50.7% on RGBNT201, which is 6.3% higher than TOP-ReID. On RGBNT100, DeMo achieves an average mAP of 68.7%, surpassing TOP-ReID by 9.3%. These results fully validate the robustness of our proposed DeMo in handling diverse and complex ReID scenarios.

Index	Modules			Metrics		Params	FLOPs
	PIFE	HDM	ATMoE	mAP	R-1	M	G
A	×	×	×	70.7	72.4	86.41	34.28
B	✓	×	×	73.0	75.8	87.99	34.28
C	✓	✓	×	74.4	77.5	95.96	35.09
D	✓	✓	✓	76.8	79.8	98.79	35.10
E	✓	✓	✓	**79.0**	**82.3**	98.79	35.10

Table 5: Comparison with different modules. Model D infers with f, while Model E further incorporates f_m.

Ablation Studies

We evaluate the effectiveness of different modules on the RGBNT201 dataset. To be specific, our baseline model only utilizes the class tokens from the visual encoders.

Effects of Key Modules. Tab. 5 shows the performance comparison with different modules. Model A is the baseline model, achieving an mAP of 70.7% and Rank-1 of 72.4%. With PIFE, Model B increases the performance to an mAP of 73.0%. Model C further incorporates HDM, boosting mAP to 74.4% and Rank-1 to 77.5%, indicating the robustness of decoupled multi-modal features. Model D introduces ATMoE, delivering an mAP of 76.8% and Rank-1 of 79.8%. Finally, Model E combines features after PIFE and ATMoE, achieving the best results with an mAP of 79.0% and Rank-1 of 82.3%. As for the complexity analysis, our proposed modules introduce a minor increase in learnable parameters (less than 13MB). In addition, the increase of FLOPs is rather small when compared with the baseline model. These results demonstrate the effectiveness of our methods.

Effects of Gating Methods. Tab. 6 compares the performance of different gating methods. The "Simple" method generates weights through a linear transformation and a softmax function applied to the decoupled features. The direct addition of weighted experts in "Simple^A" leads to worse performances, while the simple concatenation in "Simple^C" achieves better results. In contrast, our ATMoE leverages attention mechanisms to generate more accurate weights. Especially, ATMoE with a single head achieves an mAP of 77.6%, outperforming the "Simple" method. Further improvements are observed with the number of heads increased to 2 and 4, resulting in mAP scores of 78.1% and 79.0%, respectively. These results clearly highlight the effectiveness of ATMoE in generating more accurate weights.

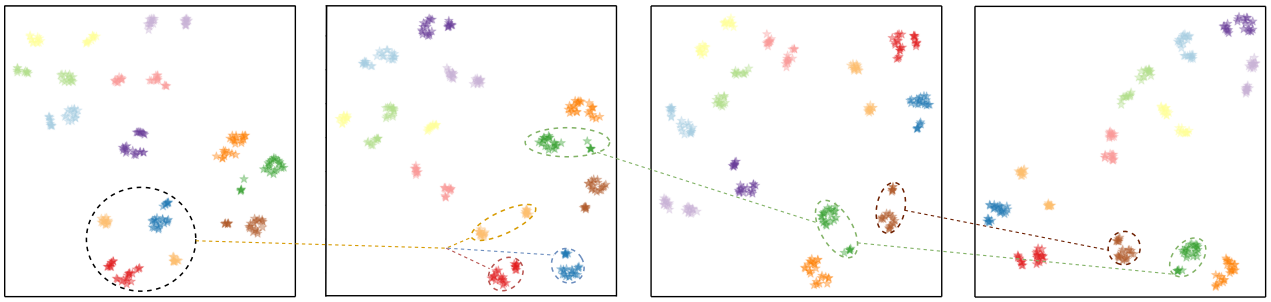

Figure 4: Feature distributions with t-SNE (Van der Maaten and Hinton 2008). Different colors refer to different IDs.

Gating	Head	mAP	Rank-1	Rank-5	Rank-10
Simple^A	-	75.2	76.7	84.6	89.7
Simple^C	-	76.7	77.4	85.2	90.1
Attention^C	1	77.6	81.6	87.7	90.3
Attention^C	2	78.1	81.8	88.4	91.7
Attention^C	4	**79.0**	82.3	**88.8**	**92.0**
Attention^C	8	78.2	**82.5**	88.2	90.4

Table 6: Comparison of gating methods. The symbol A and C indicate addition and concatenation of weighted experts.

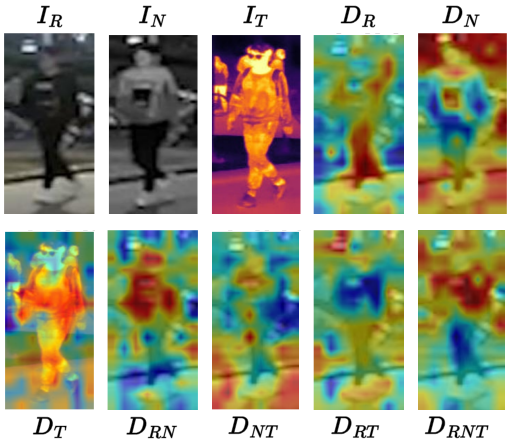

Figure 5: Activation maps of decoupled features.

Visualization Analysis

Multi-modal Feature Distributions. In Fig. 4, we visualize the discriminative feature distributions of different modules. From Fig. 4 (a) to Fig. 4 (b), challenging samples are better separated. Compared with Fig. 4 (b), HDM further narrows the distance between instances of the same ID, thereby enhancing feature discrimination, as shown in Fig. 4 (c). Finally, with ATMoE in Fig. 4 (d), features for each ID become more compact and the gap between different IDs increases. These visualizations fully validate the effectiveness of our proposed modules in improving feature discrimination.

Activation Maps of Decoupled Features. In Fig. 5, we visualize the activation maps of decoupled features. Different features focus on distinct regions of the input image. Notably, D_{RN} highlights areas that differ from those in D_R and D_N, which are shared between I_R and I_N. This suggests that HDM effectively promotes the decoupling of multi-modal features. Similar phenomena can be observed in the D_R, D_T and D_{RT} triplet, demonstrating HDM's effectiveness in enhancing multi-modal feature diversity.

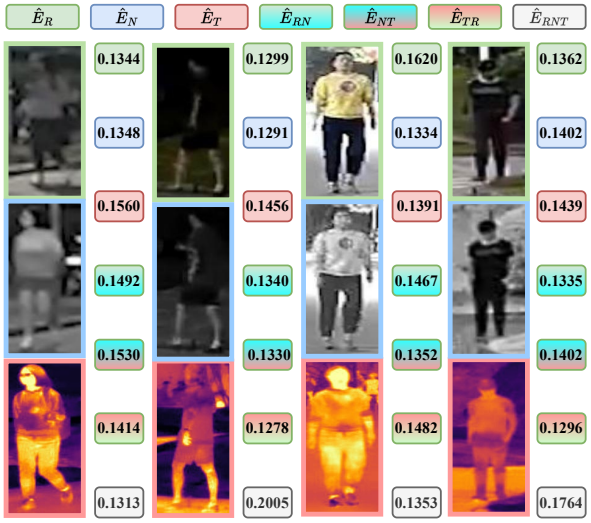

Figure 6: Visualization of dynamic weights across instances. Different colors correspond to distinct decoupled features.

Dynamic Weight Visualizations. As shown in Fig. 6, we present the dynamic weights for different instances. The weights of various decoupled features fluctuate across instances, highlighting the capability of ATMoE to adjust feature importance based on instance characteristics. Notably, modalities that contain more details receive greater attention, enhancing the robustness against imaging variations.

Conclusion

In this paper, we present a novel framework named DeMo for multi-modal object ReID. Our approach starts with a Patch-Integrated Feature Extractor (PIFE) to capture multi-granular features from diverse modalities. Then, we introduce the Hierarchical Decoupling Module (HDM) to separate modality-specific information. Finally, the Attention-Triggered Mixture of Experts (ATMoE) assigns accurate and context-aware weights to the decoupled experts. DeMo effectively enhances feature robustness against variations in imaging quality across modalities. Extensive experiments on three benchmarks validate the effectiveness of our DeMo.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62101092, 62476044, 62388101), Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No.MMC202102, MMC202407) and Fundamental Research Funds for the Central Universities (No.DUT23BK050, DUT23YG232).

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; and Huang, J. 2024. A Survey on Mixture of Experts. *arXiv preprint arXiv:2407.06204*.

Chen, L.; Sun, R.; Yu, Y.; Du, Y.; and Zhang, X. 2024. Visible thermal person re-identification via multi-branch modality residual complementary learning. *IVC*, 105201.

Chen, Y.; and Wang, L. 2024. eMoE-Tracker: Environmental MoE-based Transformer for Robust Event-guided Object Tracking. *arXiv preprint arXiv:2406.20024*.

Chowdhury, M. N. R.; Zhang, S.; Wang, M.; Liu, S.; and Chen, P.-Y. 2023. Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks. In *ICML*, 6074–6114.

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*.

Dai, D.; Deng, C.; Zhao, C.; Xu, R.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

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*.

Gui, Y.; Chen, M.; Su, Y.; Luo, G.; and Yang, Y. 2024. EEGMamba: Bidirectional State Space Models with Mixture of Experts for EEG Classification. *arXiv preprint arXiv:2407.20254*.

Guo, J.; Zhang, X.; Liu, Z.; and Wang, Y. 2022. Generative and attentive fusion for multi-spectral vehicle re-identification. In *ICSP*, 1565–1572.

He, Q.; Lu, Z.; Wang, Z.; and Hu, H. 2023. Graph-Based Progressive Fusion Network for Multi-Modality Vehicle Re-Identification. *TITS*, 1–17.

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

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

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

Hwang, C.; Cui, W.; Xiong, Y.; Yang, Z.; Liu, Z.; Hu, H.; Wang, Z.; Salas, R.; Jose, J.; Ram, P.; et al. 2023. Tutel: Adaptive mixture-of-experts at scale. *MLS*, 5: 269–287.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 448–456. pmlr.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Kuang, Z.; Zhang, H.; Cheng, L.; Liu, Y.; Huang, Y.; and Ding, X. 2024. Unity in Diversity: Multi-expert Knowledge Confrontation and Collaboration for Generalizable Vehicle Re-identification. *arXiv preprint arXiv:2407.07351*.

Li, H.; Li, C.; Zhu, X.; Zheng, A.; and Luo, B. 2020. Multi-spectral vehicle re-identification: A challenge. In *AAAI*, volume 34, 11345–11353.

Li, X.; Li, Q.; Liang, F.; and Wang, W. 2023. Multi-granularity pseudo-label collaboration for unsupervised person re-identification. *CVIU*, 227: 103616.

Li, Y.; Jiang, S.; Hu, B.; Wang, L.; Zhong, W.; Luo, W.; Ma, L.; and Zhang, M. 2024. Uni-MoE: Scaling Unified Multimodal LLMs with Mixture of Experts. *arXiv preprint arXiv:2405.11273*.

Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; Zhang, J.; Ning, M.; and Yuan, L. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Liu, T.; Liu, H.; Shang, F.; Yu, L.; Han, T.; and Wan, L. 2024a. Completed Feature Disentanglement Learning for Multimodal MRIs Analysis. *arXiv preprint arXiv:2407.04916*.

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*.

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. 2024b. A video is worth three views: Trigeminal transformers for video-based person re-identification. *TITS*.

Lu, H.; Zou, X.; and Zhang, P. 2023. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *AAAI*, volume 37, 1835–1843.

Pan, W.; Huang, L.; Liang, J.; Hong, L.; and Zhu, J. 2023. Progressively Hybrid Transformer for Multi-Modal Vehicle Re-Identification. *Sensors*, 23(9): 4206.

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. PMLR.

Rao, Y.; Chen, G.; Lu, J.; and Zhou, J. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *ICCV*, 1025–1034.

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, Y.; Xie, Y.; Qu, Y.; et al. 2024b. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 480–496.

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

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Wang, Y.; Liu, X.; Zhang, P.; Lu, H.; Tu, Z.; and Lu, H. 2023. TOP-ReID: Multi-spectral Object Re-Identification with Token Permutation. *arXiv preprint arXiv:2312.09612*.

Wang, Y.; Zhang, P.; Wang, D.; and Lu, H. 2024a. 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. 2024b. Heterogeneous Test-Time Training for Multi-Modal Person Re-identification. In *AAAI*, volume 38, 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*, volume 36, 2633–2641.

Wei, S.; Luo, Y.; Wang, Y.; and Luo, C. 2024. Robust Multimodal Learning via Representation Decoupling. *arXiv preprint arXiv:2407.04458*.

Wu, D.; Liu, Z.; Chen, Z.; Gan, S.; Tan, K.; Wan, Q.; and Wang, Y. 2025. LRMM: Low rank multi-scale multi-modal fusion for person re-identification based on RGB-NI-TI. *ESWA*, 263: 125716.

Xu, B.; Liang, J.; He, L.; and Sun, Z. 2022. Mimic embedding via adaptive aggregation: Learning generalizable person re-identification. In *ECCV*, 372–388. Springer.

Yang, B.; Chen, J.; and Ye, M. 2023. Towards Grand Unified Representation Learning for Unsupervised Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11069–11079.

Yang, B.; Chen, J.; and Ye, M. 2024. Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16870–16879.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 44(6): 2872–2893.

Yu, C.; Liu, X.; Wang, Y.; Zhang, P.; and Lu, H. 2024a. TF-CLIP: Learning text-free CLIP for video-based person re-identification. In *AAAI*, volume 38, 6764–6772.

Yu, Z.; Huang, Z.; Hou, M.; Pei, J.; Yan, Y.; Liu, Y.; and Sun, D. 2024b. Representation Selective Coupling via Token Sparsification for Multi-Spectral Object Re-Identification. *TCSVT*.

Zhang, G.; Zhang, P.; Qi, J.; and Lu, H. 2021. Hat: Hierarchical aggregation transformers for person re-identification. In *ACM MM*, 516–525.

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

Zhang, Q.; Wei, Y.; Han, Z.; Fu, H.; Peng, X.; Deng, C.; Hu, Q.; Xu, C.; Wen, J.; Hu, D.; et al. 2024b. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

Zheng, A.; He, Z.; Wang, Z.; Li, C.; and Tang, J. 2023. Dynamic Enhancement Network for Partial Multi-modality Person Re-identification. *arXiv preprint arXiv:2305.15762*.

Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; and Tang, J. 2021. Robust multi-modality person re-identification. In *AAAI*, volume 35, 3529–3537.

Zheng, A.; Zhu, X.; Ma, Z.; Li, C.; Tang, J.; and Ma, J. 2022. Multi-spectral vehicle re-identification with cross-directional consistency network and a high-quality benchmark. *arXiv preprint arXiv:2208.00632*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *AAAI*, volume 34, 13001–13008.

Zhou, K.; Yang, Y.; Cavallaro, A.; and Xiang, T. 2019. Omni-scale feature learning for person re-identification. In *ICCV*, 3702–3712.