AdapDeFormer: Adaptive Deformable Attention Aggregation with Hierarchical Query Learning for Multi-Modal Object Re-identification

Xingan Ma[1], Yuhao Wang[2,†], Jinhui Yi[1], Juergen Gall[1,†]
[1]University of Bonn [2]Dalian University of Technology
maxingan1993@gmail.com, 924973292@mail.dlut.edu.cn, {yij, gall}@iai.uni-bonn.de

***Abstract*—Multi-modal object Re-Identification (ReID) exploits complementary cues from heterogeneous sensors to recognize identities under challenging environmental conditions. However, existing methods suffer from three key limitations. First, modality reliability can vary substantially across instances and environments, rendering static fusion suboptimal. Second, sensor-induced spatial misalignment often corrupts cross-modal correspondences and degrades feature fusion. Third, naive feature aggregation lacks a progressive mechanism to integrate identity cues from fine-grained details to high-level semantics. To address these issues, we propose AdapDeFormer, a unified framework that combines Enhanced Deformable Attention (EDA) with a Hierarchical Adaptive Query Network (HAQN) for robust multi-modal representation learning. Specifically, EDA reweights modalities via Adaptive Modal Weighting (AMW) and corrects spatial discrepancies through Multi-Scale Offset Fusion (MSOF) to enable alignment-aware fusion. Building on the enhanced features, HAQN performs progressive identity feature extraction with hierarchical query learning, where cross-layer query propagation and modality-adaptive aggregation facilitate discriminative representation learning. Extensive experiments on three multi-modal object ReID benchmarks demonstrate the effectiveness and robustness of AdapDeFormer.**

***Index Terms*—Multi-modal object re-identification, Deformable attention aggregation, Hierarchical query learning**

I. Introduction

Object Re-Identification (ReID) aims to retrieve the same object across different camera views. Over the past decade, single-modal object ReID [1]–[4], 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. Therefore, multi-modal imaging introduces complementary information from RGB, Near Infrared (NIR) and Thermal Infrared (TIR) modalities [5], [6], which can enhance representation robustness in challenging scenarios.

Despite great progress, many of them [7] overlook the dynamic quality changes [8] inherent in multi-modal imaging. Environmental factors and sensing noise cause content fluctuations even within the same modality [9]. The relative reliability of each modality can vary across instances under similar imaging conditions. For example, in dark scenarios, the RGB image provides limited information, whereas the NIR and TIR images clearly show discriminative details [10]. Thus, adaptive modality weighting is crucial to account for the instance-wise variation in modality reliability. Moreover, multi-modal images captured by different sensors often exhibit spatial misalignments. Directly fusing them can introduce erroneous correspondences and degrade feature robustness [8]. Beyond adaptive fusion and alignment, another limitation of existing methods lies in their flat feature representation. As shown in Fig. 1, most methods extract features at a single layer or simply stack multi-layer features, making it difficult to simultaneously capture fine-grained details and global semantics. However, robust representation requires multi-scale cues ranging from local discriminative regions to holistic appearance. Therefore, a hierarchical progressive mechanism is required to effectively aggregate identity-relevant information.

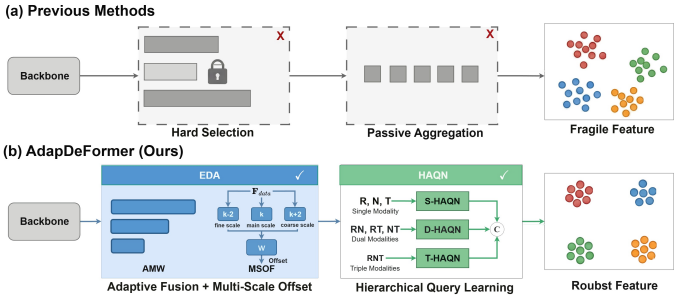

Fig. 1. Motivation for AdapDeFormer. (a) **Hard selection and passive aggregation** can overemphasize unreliable cues, yielding fragile representations under varying conditions. (b) **Adaptive fusion with alignment and hierarchical queries** adaptively fuses aligned modalities with multi-scale offsets and hierarchical query learning, producing robust representations.

Motivated by these observations, we propose AdapDeFormer for multi-modal object ReID. Our AdapDeFormer consists of two main components: Enhanced Deformable Attention (EDA) and Hierarchical Adaptive Query Network (HAQN). Specifically, EDA dynamically evaluates the contribution of each modality through Adaptive Modal Weighting (AMW), which assigns weights based on instance-specific characteristics. Meanwhile, Multi-Scale Offset Fusion (MSOF) generates multi-scale sampling offsets to correct spatial misalignments via deformable sampling. On top of EDA, HAQN employs Hierarchical Queries (HQ) for progressive feature extraction from fine to coarse, using more queries in early layers to capture details and fewer queries in deeper layers to extract semantics. Query Propagation (QP) ensures effective cross-layer information flow, while Modality-Specific archi-

†Corresponding authors.

tectures (MS) and Adaptive Fusion (AF) further strengthen multi-modal representations. As illustrated in Fig. 1, previous methods rely on hard selection and passive aggregation, which often leads to fragile representations. In contrast, our framework achieves adaptive modality weighting and spatial alignment through EDA and performs progressive feature aggregation through HAQN, resulting in more robust identity representations. Extensive experiments on three multi-modal object ReID benchmarks demonstrate the effectiveness of AdapDeFormer. In summary, our contributions are as follows:

- We introduce AdapDeFormer, a unified framework that combines Enhanced Deformable Attention (EDA) and a Hierarchical Adaptive Query Network (HAQN) to address key limitations in multi-modal object ReID.
- We design EDA with Adaptive Modal Weighting (AMW) for instance-adaptive modality balancing and Multi-Scale Offset Fusion (MSOF) for alignment-aware fusion.
- We develop HAQN with Hierarchical Queries (HQ), Query Propagation (QP), Modality-Specific (MS) and Adaptive Fusion (AF) for progressive aggregation.
- Extensive experiments on three multi-modal object ReID benchmarks validate the effectiveness of our method.

II. Related Work

A. Multi-modal Object Re-Identification

Multi-modal object ReID has attracted increasing attention for its robustness in real-world scenarios, with existing efforts exploiting complementary cues via fusion, modality-specific learning, or modality completion. For multi-modal person ReID, Zheng et al. [5] introduce progressive fusion to learn robust representations, while Wang et al. [10] strengthen modality-specific knowledge with an interact-embed-enlarge framework. To handle missing modalities, Zheng et al. [11] adopt pixel reconstruction and for multi-modal vehicle ReID, Li et al. [6] propose coherence-based fusion. Subsequent methods further improve robustness through cross-modal interaction [12]. Benefiting from Vision Transformers [13], Transformer-based approaches have been explored for multi-modal ReID [7], [14]–[16]. For example, TOP-ReID [7] guides modality fusion via token permutation, whereas EDITOR [16] suppresses background interference through diverse token selection. More recently, CLIP-based methods have achieved strong performance, such as DeMo [9] which employs MoE for adaptive weighting of decoupled features and IDEA [8] which leverages text semantics to guide feature learning. Despite these advances, adaptive modality fusion and spatial alignment are often treated implicitly. In contrast, our AdapDeFormer explicitly performs adaptive modality weighting and spatial misalignment correction via EDA, and introduces hierarchical query learning in HAQN for progressive multi-scale identity aggregation.

B. Multi-Modal Feature Aggregation

Multi-modal feature aggregation aims to integrate complementary information from heterogeneous sources to improve representation robustness. In general vision tasks, attention bottlenecks [17] provide a compact interface for aggregating multi-modal signals and synergistic residual prompts [18] further promote joint learning across modalities. In multi-modal object ReID, Wang et al. [7] leverage complementary reconstruction to reduce cross-modal distribution gaps, while Zhang et al. [16] introduce pairwise background consistency to align background features. However, pairwise alignment is difficult to extend to simultaneous alignment across multiple modalities. Moreover, existing approaches rarely unify modality weighting and spatial correspondence modeling within a single framework. More importantly, many methods rely on flat aggregation, which limits the exploitation of identity cues across semantic levels. To bridge these critical gaps, we propose EDA for adaptive weighting and alignment and HAQN for hierarchical feature aggregation.

III. Method

As illustrated in Fig. 2, AdapDeFormer builds upon a pre-trained CLIP vision encoder to extract modality-specific features from RGB, NIR and TIR images. It then performs two-stage representation learning: Enhanced Deformable Attention (EDA) enhances and aligns multi-modal features by modeling instance-wise modality reliability and sensor-induced spatial misalignment, while the Hierarchical Adaptive Query Network (HAQN) progressively aggregates identity cues across semantic levels and modality combinations.

A. Enhanced Deformable Attention

EDA fuses multi-modal features by addressing modality reliability and spatial misalignment.

Adaptive Modal Weighting (AMW). AMW learns to assign higher weights to more informative modalities based on instance-specific characteristics. Technically, AMW employs a gating network $\mathcal{G}_{gate} : \mathbb{R}^{B\times 3C} \rightarrow \mathbb{R}^{B\times 3}$ composed of two convolutional layers to analyze pooled concatenated features and predict instance-specific scores. These scores are combined with learnable global weights $\boldsymbol{\theta}_{global} \in \mathbb{R}^3$ to compute the final modality weights with softmax normalization:

$$\mathbf{w} = \text{Softmax}(\boldsymbol{\theta}_{global} \odot \mathcal{G}_{gate}(\text{GAP}(\mathbf{F}_{concat}))), \quad (1)$$

where $\odot$ denotes element-wise multiplication, $\text{GAP}(\cdot)$ denotes global average pooling and $\mathbf{F}_{concat} = [\mathbf{F}_R, \mathbf{F}_N, \mathbf{F}_T]$ denotes the concatenated features along the channel dimension. $\mathbf{F}_R$, $\mathbf{F}_N$ and $\mathbf{F}_T$ represent the features from RGB, NIR and TIR modalities, respectively. We then obtain the weighted features:

$$\mathbf{F}'_m = \mathbf{F}_m \odot w_m, \quad m \in \{R, N, T\}, \quad (2)$$

where w_m is the m-th element of $\mathbf{w} \in \mathbb{R}^{B\times 3}$.

Multi-Scale Offset Fusion (MSOF). While AMW adjusts modality importance, spatial misalignments from varied sensor properties may persist. To address this, we propose Multi-Scale Offset Fusion (MSOF), which combines modality-independent offset learning with multi-scale receptive fields to capture spatial contexts at different granularities. For each modality $m \in \{R, N, T\}$, we employ a dedicated set of three parallel offset generators $\{\mathcal{O}_{m,s}\}_{s\in\mathcal{S}}$ with scales

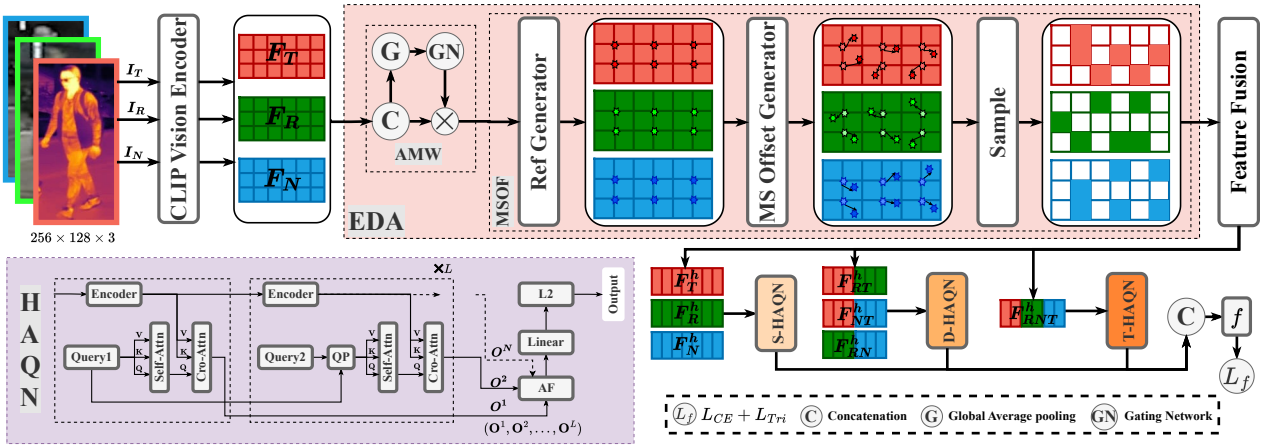

Fig. 2. Overview of AdapDeFormer. Given tri-modal inputs, a pre-trained CLIP vision encoder first extracts modality-specific features. **Enhanced Deformable Attention (EDA)** then enhances and aligns these features via Adaptive Modal Weighting (AMW) and deformable attention guided by Multi-Scale Offset Fusion (MSOF). Next, the enhanced features are concatenated to form seven modality groups corresponding to all modality subsets. **Hierarchical Adaptive Query Network (HAQN)** processes these groups in parallel with three branches using transformer encoders for hierarchical query learning. With EDA and HAQN, AdapDeFormer effectively performs alignment-aware adaptive fusion and progressive feature aggregation for robust multi-modal representation learning.

$\mathcal{S} = \{main, coarse, fine\}$, operating independently on the weighted features $\mathbf{F}'_m$ as follows:

$$\boldsymbol{\delta}_{m,s} = \mathcal{O}_{m,s}(\mathbf{F}'_m; k_s), \quad s \in \mathcal{S}, \tag{3}$$

where $\boldsymbol{\delta}_{m,s} \in \mathbb{R}^{H_s \times W_s \times 2}$ denotes the 2D sampling offset for modality m at scale s and k_s is the kernel size with $k_{main} = k$, $k_{coarse} = k + 2$, $k_{fine} = \max(k - 2, 1)$ for base, larger and smaller receptive fields. Each generator $\mathcal{O}_{m,s}$ follows a three-layer convolutional architecture: $\mathcal{C}^{(1)}$ for channel projection, depthwise $\mathcal{C}^{(2)}$ with kernel k_s and $\mathcal{C}^{(3)}$ for 2D offset generation, with GELU activations. The multi-scale offsets are fused via learnable weights as follows:

$$\boldsymbol{\delta}_m = \sum_{s \in \mathcal{S}} \alpha_s \cdot \boldsymbol{\delta}_{m,s}, \tag{4}$$

where $\boldsymbol{\alpha} = \text{Softmax}(\boldsymbol{\theta}_{scale})$ with learnable $\boldsymbol{\theta}_{scale} \in \mathbb{R}^3$. Each modality then obtains its own sampling positions as follows:

$$\mathbf{P}_m = \mathbf{P}_{ref} + \gamma \cdot \tanh(\boldsymbol{\delta}_m), \tag{5}$$

where $\mathbf{P}_{ref} \in \mathbb{R}^{H \times W \times 2}$ denotes the normalized reference grid and γ controls the offset range. The enhanced features are obtained via bilinear sampling at modality-specific locations:

$$\mathbf{F}^e_m = \text{Sample}(\mathbf{F}'_m, \mathbf{P}_m), \quad m \in \{R, N, T\}. \tag{6}$$

This design enables each modality to learn optimal spatial correspondences with multi-scale context.

B. Hierarchical Adaptive Query Network (HAQN)

After obtaining enhanced multi-modal features $\{\mathbf{F}^e_m\}_{m \in \{R,N,T\}}$, HAQN aggregates information across different abstraction levels and modality combinations. Specifically, we denote the input features of HAQN as $\mathbf{F}^h_m = \mathbf{F}^e_m$. First, we create seven distinct feature groups by concatenating these features to represent all single-modal ($\mathbf{F}^h_R, \mathbf{F}^h_N, \mathbf{F}^h_T$), dual-modal ($\mathbf{F}^h_{RN}, \mathbf{F}^h_{RT}, \mathbf{F}^h_{NT}$) and triple-modal ($\mathbf{F}^h_{RNT}$) combinations, as shown in Fig. 2. This allows HAQN to learn from all modality combinations through four components: Hierarchical Queries (HQ) for progressive multi-scale extraction, Query Propagation (QP) for cross-layer continuity, Modality-Specific architectures (MS) for complexity-adapted processing, and Adaptive Fusion (AF) for discriminative layer weighting.

Hierarchical Queries (HQ). Unlike flat query methods that use a fixed number of queries per layer, our HQ strategy assigns more queries to early layers for capturing fine-grained details and fewer queries to deeper layers for extracting high-level semantic information as follows:

$$M^l = \max(8, \lfloor M \cdot (1.0 - \frac{l}{L-1} \cdot 0.3) \rfloor), \tag{7}$$

where M^l is the number of queries at layer l, M is the initial query count, L is the network depth and 0.3 is empirically selected to ensure gradual capacity reduction while retaining sufficient queries at each layer (see Tab. V for sensitivity analysis of key hyperparameters).

Query Propagation (QP). To build consistent representations, we propagate learned queries from each layer to guide feature aggregation in subsequent layers. When consecutive layers have different query counts, we use a linear adapter to align them for seamless propagation as follows:

$$\hat{\mathbf{Q}}^{l-1} = \begin{cases} \mathbf{Q}^{l-1} & \text{if } M^{l-1} = M^l \\ \mathcal{A}^l(\mathbf{Q}^{l-1}) & \text{if } M^{l-1} \neq M^l \end{cases} \tag{8}$$

where $\mathcal{A}^l$ is a learnable linear transformation at layer l that adapts the query count from M^{l-1} to M^l. We then integrate the propagated queries with the current layer's initial queries through adaptive gating:

$$\mathbf{Q}^l = \alpha^l \cdot \mathbf{Q}^l_0 + (1 - \alpha^l) \cdot \hat{\mathbf{Q}}^{l-1}, \tag{9}$$

where α^l is a learnable gating parameter and $\mathbf{Q}^l_0$ denotes the initial queries at layer l. This mechanism enables the network to balance between preserving rich identity patterns from earlier layers and incorporating new layer-specific information needed for the current abstraction level.

Modality-Specific Architectures (MS). Since fusing more modalities increases complexity, we design three HAQN branches with shared parameters but different query capacities to handle single, dual and triple modal feature groups. These branches share network weights for parameter efficiency, while using modality-specific queries to match the complexity of different combinations: single-modal (RGB, NIR, TIR), dual-modal (RGB-NIR, RGB-TIR, NIR-TIR) and triple-modal (RGB-NIR-TIR). Each branch's capacity is adjusted through its base query count as formulated below:

$$M_{\text{single}} = M, \quad M_{\text{dual}} = \lfloor 1.2 \times M \rfloor, \quad M_{\text{triple}} = \lfloor 1.5 \times M \rfloor \tag{10}$$

where M_{single}, M_{dual} and M_{triple} represent the base query numbers for single-modal, dual-modal and triple-modal processing networks, respectively.

Adaptive Fusion (AF). Unlike previous methods [7] that simply concatenate all layer outputs, we propose Adaptive Fusion (AF) to dynamically weight the contribution of each hierarchical layer within each HAQN network, ensuring that more discriminative layers contribute more to the final representation. Given the query outputs $\{\mathbf{O}^l\}_{l=1}^{L}$ from L layers within a single HAQN, where $\mathbf{O}^l \in \mathbb{R}^{B \times M^l \times D}$ with varying query numbers M^l per layer, we compute layer-wise importance weights:

$$w^l = \text{Sigmoid}(\mathcal{F}_{\text{att}}(\text{GAP}(\mathbf{O}^l))), \quad l = 1, 2, \ldots, L, \tag{11}$$

where $\mathcal{F}_{\text{att}}$ is a two-layer attention network. Then, the fused representation for each HAQN is computed as follows:

$$\tilde{\mathbf{F}}_m = \mathcal{L}_2\left(\mathcal{P}\left(\text{Concat}_{l=1}^{L}(\mathbf{O}^l \cdot w^l)\right)\right), \tag{12}$$

where w^l scales all queries in layer l, $\mathcal{P}(\cdot)$ represents linear projection and $\mathcal{L}_2(\cdot)$ denotes L2 normalization. Then, this procedure is applied independently to S-HAQN, D-HAQN and T-HAQN, yielding seven L2-normalized feature vectors that are concatenated as the final HAQN representation for robust and discriminative multi-modal object ReID.

C. Objective Function

We optimize our AdapDeFormer using features from both the backbone and HAQN outputs. Each supervision level employs label smoothing cross-entropy loss and triplet loss:

$$\mathcal{L}_f(\mathcal{X}) = \mathcal{L}_{ce}(\mathcal{X}) + \mathcal{L}_{triplet}(\mathcal{X}), \tag{13}$$

where $\mathcal{X}$ means the input features. The overall loss is:

$$\mathcal{L} = \mathcal{L}_f([\mathbf{F}_R, \mathbf{F}_N, \mathbf{F}_T]) + \mathcal{L}_f(\tilde{\mathcal{F}}), \tag{14}$$

where $[\mathbf{F}_R, \mathbf{F}_N, \mathbf{F}_T]$ represents the backbone features from RGB, NIR and TIR modalities and $\tilde{\mathcal{F}} = \{\tilde{\mathbf{F}}_S \mid S \subseteq \{R, N, T\},\ S \neq \emptyset\}$ denotes the seven features from HAQN.

IV. Experiments

A. Datasets and Evaluation Protocols

We evaluate AdapDeFormer on three multi-modal object ReID benchmarks. **RGBNT201** [5] is a multi-modal person ReID dataset containing 4,787 aligned RGB, NIR and TIR images from 201 identities. **RGBNT100** [6] is a large-scale multi-modal vehicle ReID dataset with 17,250 image triplets covering diverse challenging visual conditions. **MSVR310** [12] is a small-scale multi-modal vehicle ReID dataset with 2,087 image triplets captured across diverse environments. Following common practice, we report mean Average Precision (mAP) and Rank-K accuracy ($K = 1, 5, 10$), with mAP and Rank-1 as the primary metrics in our analysis. Besides, we report the parameter count and FLOPs to validate efficiency.

TABLE I
PERFORMANCE COMPARISON. BEST RESULTS ARE IN **BOLD**, SECOND BEST ARE UNDERLINED. † DENOTES CLIP-BASED METHODS, * INDICATES VIT-BASED, OTHERS ARE CNN-BASED.

Methods	RGBNT201				RGBNT100		MSVR310	
	mAP	R-1	R-5	R-10	mAP	R-1	mAP	R-1
HAMNet [6]	27.7	26.3	41.5	51.7	74.5	93.3	27.1	42.3
PFNet [5]	38.5	38.9	52.0	58.4	68.1	94.1	23.5	37.4
DENet [11]	42.4	42.2	55.3	64.5	-	-	-	-
IEEE [10]	46.4	47.1	58.5	64.2	61.3	87.8	21.0	41.0
LRMM [19]	52.3	53.4	64.6	73.2	78.6	96.7	36.7	49.7
TIENet [20]	54.4	54.4	66.3	71.1	-	-	-	-
CCNet [12]	-	-	-	-	77.2	96.3	36.4	55.2
UniCat* [15]	57.0	55.7	-	-	79.4	96.2	-	-
GraFT* [21]	-	-	-	-	76.6	94.3	-	-
PHT* [14]	-	-	-	-	79.9	92.7	-	-
HTT* [22]	71.1	73.4	83.1	87.3	75.7	92.6	-	-
TOP-ReID* [7]	72.3	76.6	84.7	89.4	81.2	96.4	35.9	44.6
FACENet* [23]	-	-	-	-	81.5	96.9	36.2	54.1
EDITOR* [16]	66.5	68.3	81.1	88.2	82.1	96.4	39.0	49.3
WTSF-ReID* [24]	67.9	72.2	83.4	89.7	82.2	96.5	39.2	49.1
RSCNet* [25]	68.2	72.5	-	-	82.3	96.6	39.5	49.6
PromptMA† [26]	78.4	80.9	87.0	88.9	85.3	97.4	55.2	**64.5**
MambaPro† [18]	78.9	83.4	89.8	91.9	83.9	94.7	47.0	56.5
DeMo† [9]	79.0	82.3	88.8	92.0	86.2	97.6	49.2	59.8
IDEA† [8]	80.2	82.1	90.0	93.3	**87.2**	96.5	47.0	62.4
AdapDeFormer†	**81.5**	**84.1**	**91.3**	**95.5**	86.6	**98.1**	**56.0**	**64.5**

B. Implementation Details

We employ the pre-trained CLIP-ViT-B/16 [27] as the visual encoder backbone. Images are resized to 256×128 for RGBNT201 and 128×256 for RGBNT100/MSVR310. The mini-batch size is set to 64 with 8 images per identity for RGBNT201/MSVR310 and 128 with 16 images per identity for RGBNT100. For the EDA module, we set the deformable kernel size $k = 4$ with stride 2, offset range factor 5.0 and single-head attention ($n_{heads} = 1$) with 512 channels. For the HAQN module, we employ 16 base queries with 4 hierarchical layers. We fine-tune the proposed modules using Adam optimizer with a learning rate of 3.5×10^{-4} and apply a smaller learning rate of 5×10^{-6} for the visual backbone.

C. Comparison with State-of-the-Art Methods

Multi-modal Person ReID. Tab. I reports results on RGBNT201. AdapDeFormer achieves 81.5% mAP and 84.1% Rank-1 accuracy, outperforming the previous best CLIP-based method IDEA [8] by 1.3% mAP and 2.0% Rank-1. Moreover, compared to the ViT-based TOP-ReID [7], we obtain substantial gains of 9.2% mAP and 7.5% Rank-1, highlighting the benefits of EDA and HAQN under cross-modal reliability variation and spatial misalignment.

Multi-modal Vehicle ReID. Tab. I summarizes the vehicle ReID results. On RGBNT100, we achieve 86.6% mAP and

TABLE II
COMPARISON OF PARAMETERS AND FLOPS.

Methods	Params	FLOPs	RGBNT201		RGBNT100		MSVR310	
	(M)	(G)	mAP	R-1	mAP	R-1	mAP	R-1
TOP-ReID [7]	324.5	35.5	72.3	76.6	81.2	96.4	35.9	44.6
EDITOR [16]	119.3	40.8	66.5	68.3	82.1	96.4	39.0	49.3
PromptMA [26]	107.9	36.2	78.4	80.9	85.3	97.4	55.2	**64.5**
MambaPro [18]	74.8	52.4	78.9	83.4	83.9	94.7	47.0	56.5
DeMo [9]	98.8	35.1	79.0	82.3	86.2	97.6	49.2	59.8
IDEA [8]	91.7	43.7	80.2	82.1	**87.2**	96.5	47.0	62.4
AdapDeFormer	108.0	35.4	**81.5**	**84.1**	86.6	**98.1**	**56.0**	**64.5**

TABLE III
ABLATION STUDY OF EDA AND HAQN COMPONENTS.

EDA	Components		RGBNT201		MSVR310	
	AMW	MSOF	mAP	R-1	mAP	R-1
A	×	×	78.9(-2.6)	81.3(-2.8)	52.1(-3.9)	60.1(-4.4)
B	×	✓	79.7(-1.8)	82.2(-1.9)	53.3(-2.7)	61.5(-3.0)
C	✓	×	80.3(-1.2)	82.1(-2.0)	54.2(-1.8)	61.8(-2.7)
D	✓	✓	**81.5**	**84.1**	**56.0**	**64.5**

HAQN	HQ	QP	MS	AF	mAP	R-1	mAP	R-1
A	×	✓	✓	✓	80.1(-1.4)	83.3(-0.8)	54.4(-1.6)	63.4(-1.1)
B	✓	×	✓	✓	78.5(-3.0)	81.9(-2.2)	54.0(-2.0)	62.6(-1.9)
C	✓	✓	×	✓	79.7(-1.8)	82.8(-1.3)	54.8(-1.2)	61.6(-2.9)
D	✓	✓	✓	×	81.1(-0.4)	83.2(-0.9)	55.6(-0.4)	63.4(-1.1)
E	✓	✓	✓	✓	**81.5**	**84.1**	**56.0**	**64.5**

98.1% Rank-1 accuracy, improving Rank-1 by 1.6% over IDEA. On the MSVR310 dataset, AdapDeFormer delivers 56.0% mAP and 64.5% Rank-1, demonstrating strong generalization across scenarios and scales.

Efficiency Analysis. Tab. II compares parameters and FLOPs. AdapDeFormer achieves state-of-the-art accuracy with competitive efficiency across all benchmarks.

D. Ablation Studies

Effects of EDA Components. The upper part of Tab. III reports an ablation of the proposed EDA components. Starting from the baseline without EDA (Model A), we obtain 78.9% mAP on RGBNT201. Introducing MSOF alone (Model B) yields a 0.8% mAP gain, supporting the role of multi-scale deformable sampling in improving spatial alignment. Similarly, AMW alone (Model C) improves performance by 1.4% mAP, indicating the benefit of dynamically balancing modality contributions. Finally, combining AMW and MSOF (Model D) achieves the best results, reaching 81.5% mAP on RGBNT201.

Effects of HAQN Components. The lower part of Tab. III further investigates the contribution of each HAQN component. Removing HQ (Model A) reduces performance by 1.4% mAP, while removing QP (Model B) causes the largest drop (3.0% mAP), highlighting the importance of cross-layer information flow. In addition, removing the modality-specific branch design (Model C) particularly degrades MSVR310, with a 2.9% Rank-1 decrease. With all components enabled, the complete framework (Model E) achieves the best overall performance, reaching 81.5% mAP on RGBNT201.

Robustness to Missing Modality. Tab. IV evaluates robustness under missing-modality scenarios on RGBNT201. AdapDeFormer consistently outperforms existing methods when a single modality is unavailable at inference. Notably, missing TIR causes the largest drop across all methods, indicating TIR's critical role. With TIR missing, our method achieves 61.2% mAP, surpassing DeMo by 5.0% and TOP-ReID by 9.3%, demonstrating the effectiveness of AMW in compensating for missing modalities.

TABLE IV
PERFORMANCE COMPARISON UNDER MISSING MODALITY ON RGBNT201. M(X) DENOTES MISSING MODALITY X.

Method	Full	M(RGB)	M(NIR)	M(TIR)
	mAP	mAP	mAP	mAP
TOP-ReID [7]	72.3	54.4(-17.9)	64.3(-8.0)	51.9(-20.4)
DeMo [9]	79.0	63.3(-15.7)	72.6(-6.4)	56.2(-22.8)
IDEA [8]	80.2	62.9(-17.3)	71.5(-8.7)	58.4(-21.8)
AdapDeFormer	**81.5**	**66.9**(-14.6)	**76.0**(-5.5)	**61.2**(-20.3)

TABLE V
HYPERPARAMETER ANALYSIS ON RGBNT201.

K	mAP	R-1
3	79.4(-2.1)	81.1(-3.0)
4	**81.5**	**84.1**
5	78.7(-2.8)	79.8(-4.3)

M	mAP	R-1
8	79.3(-2.2)	80.7(-3.4)
16	**81.5**	**84.1**
32	77.8(-3.7)	80.0(-4.1)

Hyperparameter Analysis. Tab. V analyzes the impact of key hyperparameters on RGBNT201. For the MSOF kernel size K, K=4 achieves the best trade-off. A smaller K=3 limits receptive-field coverage for cross-modal alignment, resulting in a 2.1% mAP drop, whereas a larger K=5 introduces more background noise and degrades mAP by 2.8%. For the query number M, setting M=16 balances capacity and efficiency. Reducing it to M=8 limits robustness and decreases mAP by 2.2%, while increasing it to M=32 introduces feature redundancy and leads to a 3.7% mAP reduction.

E. Visualization Analysis

Feature Distributions. Fig. 3 presents t-SNE visualizations of learned feature embeddings. Compared with the baseline (A), both EDA (B) and HAQN (C) lead to clearer inter-class separation. Moreover, the full model EDA+HAQN (D) yields the most compact intra-class clusters and clearest inter-class boundaries, confirming that EDA's alignment-aware fusion and HAQN's hierarchical aggregation are complementary for discriminative identity learning.

Cosine Similarity Distributions. Fig. 4 visualizes the cosine similarity distributions of positive and negative pairs. The overlap between the two distributions progressively decreases from the baseline to our full model, indicating enlarged margins for more discriminative representations.

Attention Maps. Fig. 5 visualizes activation patterns across seven modality groups. Single-modal features focus on fine-grained local regions, dual-modal capture intermediate patterns, and triple-modal (RNT) emphasizes global structure. These complementary patterns validate our design.

V. CONCLUSION

In this paper, we propose AdapDeFormer, an adaptive deformable attention aggregation framework for multi-modal object ReID. Specifically, EDA enhances cross-modal feature

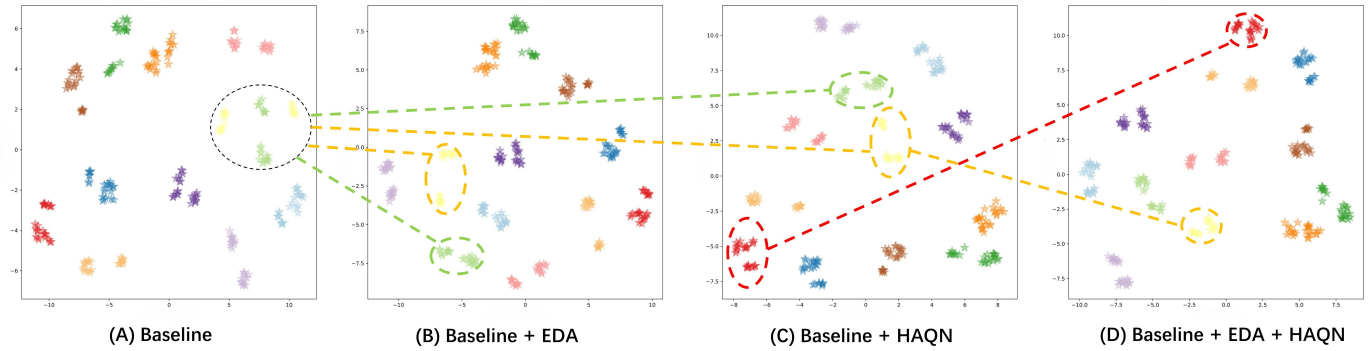

Fig. 3. t-SNE visualization of learned feature embeddings. Different colors denote different identities.

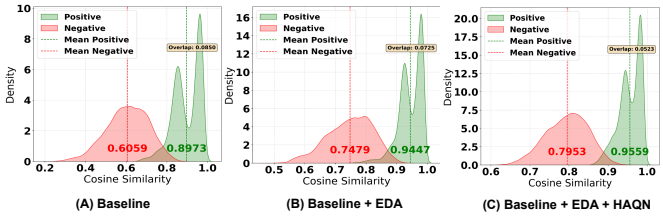

Fig. 4. Cosine similarity distributions of positive and negative pairs.

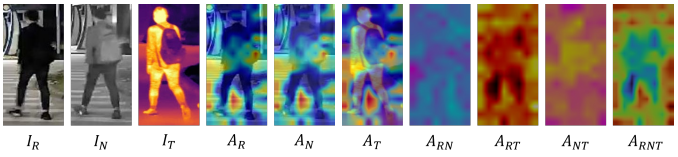

Fig. 5. Activation maps of seven modality groups, illustrating complementary attention patterns across modality combinations.

alignment and fusion by jointly leveraging AMW and multi-scale deformable sampling, while HAQN learns hierarchical, identity-aware representations via query-based modeling with cross-layer propagation. Extensive experiments on three benchmarks demonstrate consistent improvements over prior methods, together with strong robustness to missing-modality scenarios and competitive efficiency.

REFERENCES

[1] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *ICCV*, 2021.

[2] G. Zhang, P. Zhang, J. Qi, and H. Lu, "Hat: Hierarchical aggregation transformers for person re-identification," in *ACM MM*, 2021.

[3] X. Liu, P. Zhang, C. Yu, H. Lu, and X. Yang, "Watching you: Global-guided reciprocal learning for video-based person re-identification," in *CVPR*, 2021.

[4] X. Liu, C. Yu, P. Zhang, and H. Lu, "Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification," *TNNLS*, 2023.

[5] A. Zheng, Z. Wang, Z. Chen, C. Li, and J. Tang, "Robust multi-modality person re-identification," in *AAAI*, 2021.

[6] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, "Multi-spectral vehicle re-identification: A challenge," in *AAAI*, 2020.

[7] Y. Wang, X. Liu, P. Zhang, H. Lu, Z. Tu, and H. Lu, "Top-reid: Multi-spectral object re-identification with token permutation," in *AAAI*, 2024.

[8] Y. Wang, Y. Lv, P. Zhang, and H. Lu, "Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification," in *CVPR*, 2025.

[9] Y. Wang, Y. Liu, A. Zheng, and P. Zhang, "Decoupled feature-based mixture of experts for multi-modal object re-identification," in *AAAI*, 2025.

[10] Z. Wang, C. Li, A. Zheng, R. He, and J. Tang, "Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification," in *AAAI*, 2022.

[11] A. Zheng, Z. He, Z. Wang, C. Li, and J. Tang, "Dynamic enhancement network for partial multi-modality person re-identification," *arXiv preprint arXiv:2305.15762*, 2023.

[12] A. Zheng, X. Zhu, Z. Ma, C. Li, J. Tang, and J. Ma, "Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark," *Information Fusion*, 2023.

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

[14] W. Pan, L. Huang, J. Liang, L. Hong, and J. Zhu, "Progressively hybrid transformer for multi-modal vehicle re-identification," *Sensors*, 2023.

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

[16] P. Zhang, Y. Wang, Y. Liu, Z. Tu, and H. Lu, "Magic tokens: Select diverse tokens for multi-modal object re-identification," in *CVPR*, 2024.

[17] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *NeurIPS*, 2021.

[18] Y. Wang, X. Liu, T. Yan, Y. Liu, A. Zheng, P. Zhang, and H. Lu, "Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt," in *AAAI*, 2025.

[19] D. Wu, Z. Liu, Z. Chen, S. Gan, K. Tan, Q. Wan, and Y. Wang, "Lrmm: Low rank multi-scale multi-modal fusion for person re-identification based on rgb-ni-ti," *ESWA*, 2025.

[20] X. Yang, W. Dong, D. Cheng, N. Wang, and X. Gao, "Tienet: A tri-interaction enhancement network for multimodal person reidentification," *TNNLS*, 2025.

[21] H. Yin, J. Li, E. Schiller, L. McDermott, and D. Cummings, "Graft: Gradual fusion transformer for multimodal re-identification," *arXiv preprint arXiv:2310.16856*, 2023.

[22] Z. Wang, H. Huang, A. Zheng, and R. He, "Heterogeneous test-time training for multi-modal person re-identification," in *AAAI*, 2024.

[23] A. Zheng, Z. Ma, Y. Sun, Z. Wang, C. Li, and J. Tang, "Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification," *Information Fusion*, 2025.

[24] Z. Yu, Z. Huang, M. Hou, Y. Yan, and Y. Liu, "Wtsf-reid: Depth-driven window-oriented token selection and fusion for multi-modality vehicle re-identification with knowledge consistency constraint," *ESWA*, 2025.

[25] Z. Yu, Z. Huang, M. Hou, J. Pei, Y. Yan, Y. Liu, and D. Sun, "Representation selective coupling via token sparsification for multi-spectral object re-identification," *TCSVT*, 2024.

[26] S. Zhang, W. Luo, D. Cheng, Y. Xing, G. Liang, P. Wang, and Y. Zhang, "Prompt-based modality alignment for effective multi-modal object re-identification," *IEEE TIP*, 2025.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.