Signal: Selective Interaction and Global-local Alignment for Multi-Modal Object Re-Identification

Yangyang Liu[1*], **Yuhao Wang**[1*], **Pingping Zhang**[1,2†]

[1]School of Future Technology, Dalian University of Technology

[2] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Xi'an, China

{yyliu, 924973292}@mail.dlut.edu.cn,zhpp@dlut.edu.cn,

Abstract

Multi-modal object Re-IDentification (ReID) is devoted to retrieving specific objects through the exploitation of complementary multi-modal image information. Existing methods mainly concentrate on the fusion of multi-modal features, yet neglecting the background interference. Besides, current multi-modal fusion methods often focus on aligning modality pairs but suffer from multi-modal consistency alignment. To address these issues, we propose a novel selective interaction and global-local alignment framework called **Signal** for multi-modal object ReID. Specifically, we first propose a Selective Interaction Module (SIM) to select important patch tokens with intra-modal and inter-modal information. These important patch tokens engage in the interaction with class tokens, thereby yielding more discriminative features. Then, we propose a Global Alignment Module (GAM) to simultaneously align multi-modal features by minimizing the volume of 3D polyhedra in the gramian space. Meanwhile, we propose a Local Alignment Module (LAM) to align local features in a shift-aware manner. With these modules, our proposed framework could extract more discriminative features for object ReID. Extensive experiments on three multi-modal object ReID benchmarks (i.e., RGBNT201, RGBNT100, MSVR310) validate the effectiveness of our method.

Code — https://github.com/010129/Signal

Introduction

Object Re-IDentification (ReID) aims to retrieve identical objects across non-overlapping cameras. Initially, researchers focus on single-modal object ReID (He et al. 2021; Zhang et al. 2021; Liu et al. 2021) mainly based on RGB images. However, adverse environments such as darkness and strong light can cause blurry details in RGB images. Later, researchers find that Near Infrared (NIR) and Thermal Infrared (TIR) images exhibit strong robustness in harsh visual environments. With complementary information from different modalities, existing multi-modal object ReID methods (Lin et al. 2025; Wan et al. 2025b; Li et al. 2025c; Feng et al. 2025; Li et al. 2025b; Wan et al. 2025a) achieve outstanding performance. However, they ignore background interference in each modality. As shown in Fig. 1 (a), irrelevant background information may introduce noise and affect the extraction of discriminative information. Meanwhile, recent multi-modal fusion methods (Wang et al. 2023; Yu and Song 2024; Li et al. 2024) focus on aligning different modalities in a pairwise manner through contrastive learning. As illustrated in the upper part of Fig. 1 (b), these methods typically select one modality as the anchor and align the remaining modalities to it. However, this pairwise alignment strategy becomes less effective when scaling to more than two modalities, as it fails to capture the complex relationships among all modality pairs. Thus, as shown in the lower part of Fig. 1 (b), simultaneously aligning multiple modalities without relying on a fixed anchor modality offers a more flexible solution for multi-modal alignment. Besides, current multi-modal imaging sensors often struggle to ensure precise pixel-level alignment. As shown in Fig. 1 (c), pixel-level misalignment commonly exists across different modalities, leading to semantic inconsistency in multi-modal fusion.

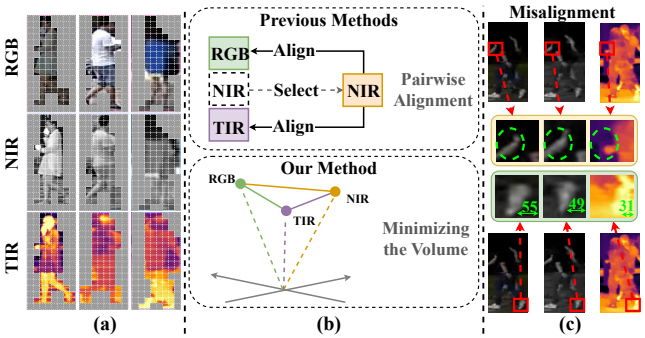

Figure 1: Motivations of our framework. (a) Background interferences in multi-modal images. (b) Comparison between previous pairwise alignment and our simultaneous alignment. (c) Misalignments exist across different modalities.

Motivated by the aforementioned observations, we propose **Signal**, a novel selective interaction and global-local alignment framework for multi-modal object ReID. Our proposed framework comprises three components: the Selective Interaction Module (SIM), the Global Alignment Mod-

*These authors contributed equally.

†Corresponding author.

ule (GAM) and the Local Alignment Module (LAM). First, SIM selects important patch tokens from multi-modal features by evaluating their significance both within and across modalities. This is achieved by computing intra-modal and inter-modal attention scores to guide the selection process. Second, GAM enables the simultaneous alignment of multi-modal features by minimizing the volume of 3D polyhedra in the gramian space (Cicchetti et al. 2025), as shown in the lower part of Fig. 1 (b). Unlike existing pairwise alignment methods, our method eliminates the need for a fixed anchor modality, enabling a more flexible and effective alignment across multiple modalities. Third, LAM further refines this process by focusing on fine-grained alignment at the local feature level. Leveraging deformable sampling, LAM adaptively aligns local details across modalities in a shift-aware manner, mitigating semantic inconsistency caused by pixel-level misalignment. With the above components, our proposed framework effectively addresses the challenges of background interference and multi-modal misalignment for robust multi-modal feature learning. Extensive experiments on three multi-modal object ReID datasets validate our method's effectiveness.

Our main contributions are summarized as follows:

- We propose a novel selective interaction and global-local alignment framework named Signal for multi-modal object ReID, which effectively addresses the challenges of background interference and multi-modal misalignment.
- We propose the Selective Interaction Module (SIM) to leverage inter-modal and intra-modal attention scores for selecting important patch tokens, thereby mitigating background interference in multi-modal fusion.
- We propose the Global Alignment Module (GAM) to simultaneously align multi-modal features through minimizing the volume of 3D polyhedra in the gramian space.
- We propose the Local Alignment Module (LAM) to align local features in a shift-aware manner, effectively addressing pixel-level misalignment across modalities.
- Extensive experiments on three multi-modal object ReID datasets validate the effectiveness of our method.

Related Work

Multi-Modal Object Re-Identification

Multi-modal object ReID is devoted to retrieving specific objects through the exploitation of multi-modal inputs. Existing methods focus on learning complementary image features. For example, Wang *et al.* (Wang et al. 2024b) propose a cyclic token permutation framework to reduce the distribution gap across different modalities. Feng *et al.* (Feng et al. 2025) integrate pixel-level interaction to balance modality-specific features. Wang *et al.* (Wang et al. 2025b) introduce the mixture of experts for adaptive weighting decoupled features. Besides, researchers find that graph-based models exhibit superior capabilities in modeling complex relational structures. Thus, Wan *et al.* (Wan et al. 2025b) introduce graph inference with modality awareness for improving feature robustness. Wan *et al.* (Wan et al. 2025a) further quantify uncertainty through graph models. Recently, researchers start to explore the use of Multi-modal Large Language Models (MLLMs) to enhance multi-modal feature learning. For instance, Wang *et al.* (Wang et al. 2025c) integrate semantic guidance from inverted texts generated by MLLMs. Li *et al.* (Li et al. 2025b) introduce text-modulated and context-shared experts to enhance feature robustness. However, these methods primarily focus on feature fusion and ignore the background interference in multi-modal object ReID. To address this issue, Zhang *et al.* (Zhang et al. 2024a) propose the object-centric feature refinement to mitigate background interference. Zhang *et al.* (Zhang et al. 2025) introduce token selection to filter out the irrelevant background noise. Although the above methods achieve remarkable performance, they typically select tokens within each modality separately, ignoring the importance of tokens across modalities. Meanwhile, previous methods (Wang et al. 2024b; Zhang et al. 2024a) mainly focus on pairwise alignment, which exhibits limitations in complex multi-modal scenarios. Thus, we introduce the selective interaction with intra-modal and inter-modal attention scores to mitigate background interference. In addition, we perform multi-modal alignment in the gramian space, which offers a great flexibility compared with pairwise alignments.

Multi-Modal Feature Fusion

Multi-modal feature fusion leverages complementary information from different modalities to enhance feature robustness. For example, Li *et al.* (Li et al. 2025a) propose a learnable modality dictionary to preserve consistency between individual modality features. Dai *et al.* (Dai et al. 2025) enhance cross-modal feature fusion through contrastive learning and reduce redundancy by utilizing visual sequence compression. Additionally, Nagrani *et al.* (Nagrani et al. 2021) employ fusion bottlenecks to facilitate modality information aggregation. In multi-modal object ReID, Zhang *et al.* (Zhang et al. 2024b) enhance feature discrimination by integrating inter-modality information with shallow and deep features through dense connections. Wang *et al.* (Wang et al. 2025a) propose a synergistic residual prompt to guide the joint learning of multi-modal features. Following this direction, later studies increasingly focus on effective feature alignment across modalities. For instance, Wang *et al.* (Wang et al. 2024b) utilize complementary reconstruction to minimize the distribution gap across different modalities. Zhang *et al.* (Zhang et al. 2024a) introduce a pairwise background consistency constraint to align background features across modalities for improved feature representation. Although these methods achieve remarkable performance, the simultaneous alignment of multi-modal features remains an under-explored area in current work. To bridge this gap, we propose a novel multi-modal alignment in the gramian space. It enables holistic and anchor-free alignment by modeling the global interactions among all modalities simultaneously in a unified space.

Methodology

As shown in Fig. 2, our proposed framework is consist of the Selective Interaction Module (SIM), the Global Alignment

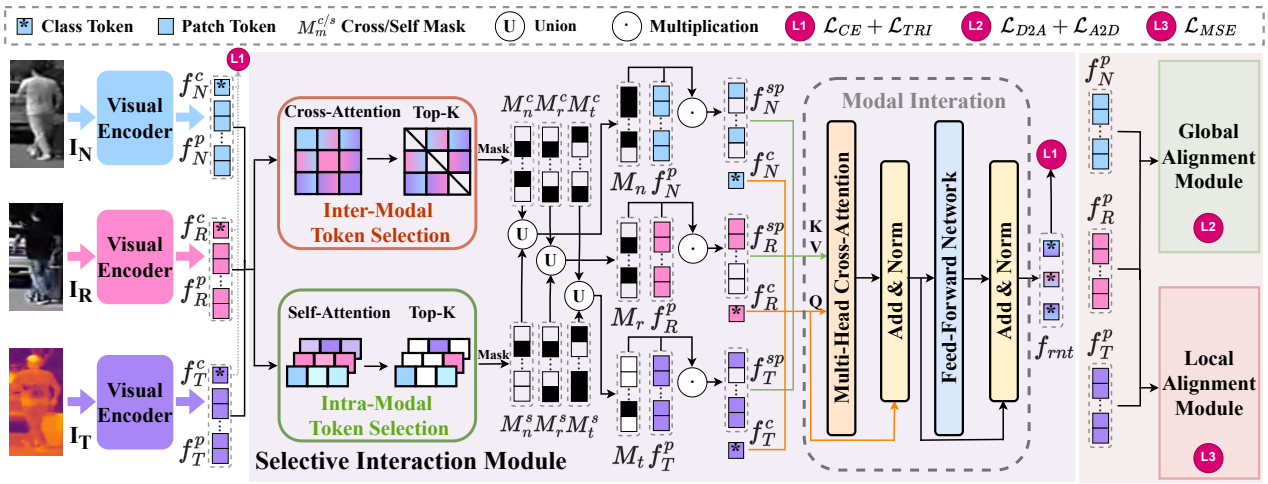

Figure 2: Illustration of our proposed framework.

Module (GAM) and the Local Alignment Module (LAM). In this section, we will describe the details of each module.

Selective Interaction Module

Background information often interferes with multi-modal object ReID. Existing methods (Wang et al. 2025c; Li et al. 2025c; Wang et al. 2025a; Wan et al. 2025a) lack an effective way to eliminate the background interference. To address this issue, we propose a Selective Interaction Module (SIM) to select important patch tokens with intra-modal and inter-modal information. Specifically, a given image $I \in \mathbb{R}^{3\times H\times W}$ is first split into L patches, where H and W denote the height and width, respectively. Then, the patch tokens are fed into a visual encoder to extract modality-specific features. As a result, the image features $\mathcal{F}_m \in \mathbb{R}^{(L+1)\times D}$ for each modality can be expressed as follows:

$$\mathcal{F}_m = \{f_m^{\text{cls}}; f_m^1, f_m^2, \cdots, f_m^L\}, \quad m \in \{N, R, T\}, \tag{1}$$

where f_m^{cls} denotes the [CLS] token and f_m^i represents the token of the i-th patch. Here, N, R and T correspond to the NIR, RGB and TIR modalities, respectively. To facilitate better explanations, we denote the features separately as:

$$f_m^c = f_m^{cls}, \tag{2}$$

$$f_m^p = \{f_m^1, f_m^2, \cdots, f_m^L\}. \tag{3}$$

Here, $f_m^c \in \mathbb{R}^D$, $f_m^p \in \mathbb{R}^{L\times D}$ and D is the embedding dimension. Then, to facilitate the selection of important patches within each modality, we propose the Intra-Modal Token Selection and Inter-Modal Token Selection.

Intra-Modal Token Selection. To assess the importance of patches within each modality, we introduce the Intra-Modal Token Selection. It leverages self-attention scores to preliminarily select and retain the most informative patches. More specifically, we first compute the attention scores of all patch tokens within each modality as follows:

$$Q_m = f_m^c W_q, K_m = f_m^p W_k, \tag{4}$$

$$\mathcal{S}_m = \text{Softmax}\left(\frac{Q_m K_m^\top}{\sqrt{D}}\right), \tag{5}$$

where W_q and W_k are identity matrices and they yield superior performance. Here, $\mathcal{S}_m \in \mathbb{R}^{1\times L}$ denotes the self-attention score of each patch token in modality m. Next, we select the high-similarity patches within each modality from $\mathcal{S}_m$ as follows:

$$\Theta_m = TopK(\mathcal{S}_m, k_1). \tag{6}$$

Here, k_1 denotes the number of important patch tokens to be selected. Θ_m is the index set of the top-k_1 patches selected within each modality. We then construct a binary mask using Θ_m to retain the selected patch tokens:

$$M_m^s = \Psi(\Theta_m). \tag{7}$$

where Ψ denotes a binary masking operation. Each element in M_m^s indicates whether the patch token is selected (1) or discarded (0). Through this process, we obtain a preliminary set of intra-modality important patches for each modality.

Inter-Modal Token Selection. To assess the inter-modal significance of each patch, we introduce the Inter-Modal Token Selection. It leverages attention scores from a cross-attention mechanism to identify informative tokens across modalities. The key insight is to measure the relevance of each patch based on the attention it receives from the remaining modalities. Specifically, class tokens from all three modalities are concatenated to form the query tokens, while patch tokens are concatenated to form the key tokens. Then, we can compute the importance scores as follows:

$$Q = \mathcal{T}[f_R^c, f_N^c, f_T^c], \tag{8}$$

$$K = \mathcal{C}[f_R^p, f_N^p, f_T^p], \tag{9}$$

$$\mathcal{S} = \text{Softmax}\left(\frac{QK^T}{\sqrt{D}}\right), \tag{10}$$

where $Q \in \mathbb{R}^{3\times D}$, $K \in \mathbb{R}^{3L\times D}$ and $\mathcal{S} \in \mathbb{R}^{3\times 3L}$. Here, $\mathcal{T}$ denotes a stacking operation followed by a linear projection,

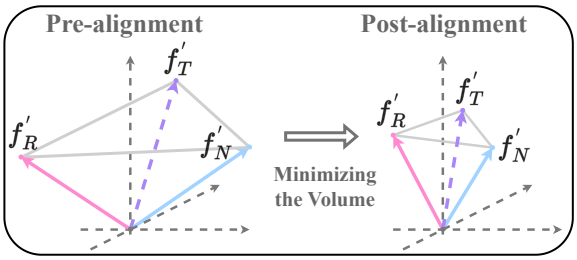

Figure 3: Details of Global Alignment Module.

and $\mathcal{C}$ denotes the concatenation followed by a linear layer. Then, we separate the scores of each modality from $\mathcal{S}$:

$$\mathcal{D}_m = \bar{\mathcal{C}}[\mathcal{S}[u \neq m]], \tag{11}$$

where $u \in \{N, R, T\}$, $\bar{\mathcal{C}}$ denotes the concatenation. $\mathcal{D}_m$ aggregates cross-attention scores by excluding self-modality tokens, indicating the relevance degree of each patch receives from other modalities. Based on $\mathcal{D}_m$, we can select the most relevant patches from the other modalities as:

$$\bar{\Theta}_m = TopK(\mathcal{D}_m, k_2). \tag{12}$$

Here, k_2 denotes the number of important patches selected based on cross-modal information and $\bar{\Theta}_m$ represents the corresponding index set. We then aggregate the patch indices selected by other modalities. Finally, masks for each modality are constructed as follows:

$$M_m^c = \Psi(\bar{\Theta}_m). \tag{13}$$

To select patches important to both their own modality and others, we combine the intra-modal and inter-modal selection masks via a union operation as follows:

$$M_m = M_m^c \cup M_m^s, \tag{14}$$

where $\cup$ denotes the union operation. Finally, we apply the mask M^m to the patch tokens f_m^p to select tokens as follows:

$$f_m^{sp} = M_m \odot f_m^p. \tag{15}$$

Here, $\odot$ means the element-wise multiplication. Through the above steps, we obtain the selected patch tokens f_m^{sp} for each modality, which effectively mitigates background interference with both intra-modal and inter-modal information.

Modal Interaction. To further reduce the background interference, we propose a modal interaction module that highlights informative features. It utilizes multi-head cross-attention (Vaswani et al. 2017) and a feed-forward network to model interactions between the selected tokens and class tokens, thereby extracting more discriminative representations. Specifically, we concatenate the class tokens f_m^c and selected patch tokens f_m^{sp} to form a query $\bar{Q}$ and a key $\bar{K}$:

$$\bar{Q} = \mathcal{T}[f_R^c, f_N^c, f_T^c], \bar{K} = \mathcal{C}[f_R^{sp}, f_N^{sp}, f_T^{sp}], \tag{16}$$

where $\bar{Q} \in \mathbb{R}^{3\times D}$ and $\bar{K} \in \mathbb{R}^{3L\times D}$. Then, we apply multi-head cross-attention to enhance the interaction between the class tokens and selected patch tokens as follows:

$$\bar{Q}' = LN(\bar{Q} + MHCA(\bar{Q}, \bar{K}, \bar{K})), \tag{17}$$

$$f_{rnt} = LN(\bar{Q}' + \phi(\bar{Q}')). \tag{18}$$

where $MHCA$ represents the multi-head cross-attention. LN represents the layer normalization (Ba, Kiros, and Hinton 2016). $\phi(\cdot)$ means the feed-forward network. Ultimately, we obtain the modality interaction feature $f_{rnt} \in \mathbb{R}^{3D}$, which aggregates the discriminative information from the selected patch tokens and class tokens across all modalities.

Global Alignment Module

Multi-modal alignment encourages consistent semantic representations across modalities, reducing cross-modal conflicts. However, traditional alignment methods (Ruan et al. 2023; Girdhar et al. 2023; Chen et al. 2023) are difficult to extend to multiple modalities. Otherwise it leads to extremely high complexity. Meanwhile, simultaneous alignment across multiple modalities remains unexplored in the context of multi-modal object ReID. Motivated by these observations, we introduce Global Alignment Module (GAM) based on multi-modal representation learning in the gramian space (Cicchetti et al. 2025). It addresses these limitations by ensuring that all modalities are aligned with each another, rather than merely aligning each modality to a designated anchor. Specifically, we preprocess patch tokens as follows:

$$f_m = Mean(f_m^p), \tag{19}$$

$$f_m' = \frac{f_m}{||f_m||_2}. \tag{20}$$

Here, f_m denotes the average feature vector. f_m' is the normalized vector. As shown in Fig. 3, we consider the volume of a 3D polyhedron composed of three vectors (i.e., f_R', f_N' and f_T') as a measure of the alignment degree of the three modalities. A larger volume indicates a worse alignment of the three modalities, while a smaller volume indicates a better alignment. We interpret the alignment quality of the modalities through this volume metric. Therefore, we can calculate the volume to align the three vectors. More specifically, f_R', f_N' and f_T' can be arranged into columns of a matrix $\mathbf{A} = (f_R', f_N', f_T')$. The Gram matrix $\mathbf{G}(f_R', f_N', f_T')$ is defined as:

$$\mathbf{G}(f_R', f_N', f_T') = \mathbf{A}^\top \mathbf{A}, \tag{21}$$

$$= \begin{bmatrix} \langle \mathbf{f_R'}, \mathbf{f_R'} \rangle & \langle \mathbf{f_R'}, \mathbf{f_N'} \rangle & \langle \mathbf{f_R'}, \mathbf{f_T'} \rangle \\ \langle \mathbf{f_N'}, \mathbf{f_R'} \rangle & \langle \mathbf{f_N'}, \mathbf{f_N'} \rangle & \langle \mathbf{f_N'}, \mathbf{f_T'} \rangle \\ \langle \mathbf{f_T'}, \mathbf{f_R'} \rangle & \langle \mathbf{f_T'}, \mathbf{f_N'} \rangle & \langle \mathbf{f_T'}, \mathbf{f_T'} \rangle \end{bmatrix}, \tag{22}$$

where $\mathbf{G}(f_R', f_N', f_T') \in \mathbb{R}^{3\times 3}$ is the geometric relationship among the three vectors. The volume of the 3D polyhedron spanned by these vectors is:

$$\mathrm{Vol}(f_R', f_N', f_T') = \sqrt{\det \mathbf{G}(f_R', f_N', f_T')}. \tag{23}$$

Here, Vol is the volume and $\det \mathbf{G}$ represents the determinant of matrix $\mathbf{G}$. Minimizing this volume enables simultaneous alignment of the tri-modal features, thereby achieving global alignment. In contrast to traditional alignment methods, GAM exhibits superior efficiency in the simultaneous alignment of an arbitrary number of modalities.

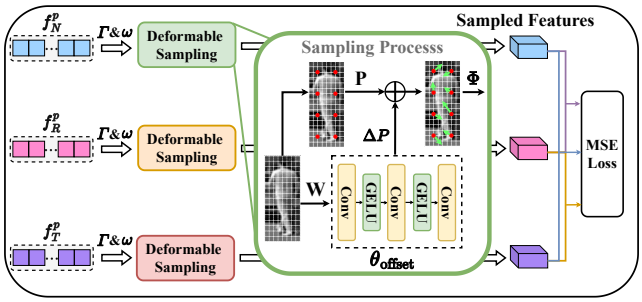

Figure 4: Mechanism of Local Alignment Module.

Local Alignment Module

The GAM achieves global alignment among the three modalities, while neglecting the issue of pixel misalignment in multi-modal imaging. To address this issue, we propose the Local Alignment Module (LAM). Unlike traditional local alignment methods (Wang et al. 2024a; Li et al. 2025d; Wang et al. 2022a), LAM emphasizes adaptive offset sampling and focuses on key details. As shown in Fig. 4, we utilize the advantages of deformable attention (Xia et al. 2022) by learning the offset to automate the correction of pixel offset errors. Specifically, we reshape the patch tokens into the spatial manner and generate uniform grid points as follows:

$$P = \omega(\Gamma(f_m^p)), \tag{24}$$

where Γ represents the reshape operation. ω denotes generating a uniform grid of points and $P \in \mathbb{R}^{H_g \times W_g \times 2}$ are reference points. The grid size is down-sampled by a factor r, $H_g = \frac{H}{r}, W_g = \frac{W}{r}$. The values of the reference points are linearly spaced at 2D coordinates $\{(0,0), \ldots, (H_g-1, W_g-1)\}$ and then normalized to the range $[-1, +1]$ based on the grid shape $H_g \times W_g$, where $(-1, -1)$ represents the top-left corner and $(+1, +1)$ represents the bottom-right corner. To obtain the offset, we perform the following operation on f_m^p:

$$\Delta P = \theta_{\text{offset}}(f_m^p W), \tag{25}$$

where W represents a linear projection layer, $\theta_{\text{offset}}(\cdot)$ is composed of multiple convolutional layers. Here, ΔP denotes the sampling offset, which adjusts the reference positions P in the feature map. The sampled features $\tilde{f}_m^p$ are then obtained using a bilinear interpolation function $\Phi(\cdot)$, which extracts features from positions $P + \Delta P$ as follows:

$$\tilde{f}_m^p = \Phi(f_m^p; P + \Delta P). \tag{26}$$

After obtaining the sampled features of three modalities, we align these features using the MSE loss to facilitate cross-modal feature alignment. Through the integration of LAM, our model exhibits the capability of mitigating pixel-level misalignment across different modalities, thereby enhancing the semantic consistency of multi-modal features.

Objective Functions

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

$$\mathcal{L}_g = \mathcal{L}_{CE} + \mathcal{L}_{TRI}, \tag{27}$$

where $\mathcal{L}_{CE}$ is the label smoothing cross-entropy loss. $\mathcal{L}_{TRI}$ is the triplet loss. Features after GAM are supervised by the gram multi-modal contrastive loss (Cicchetti et al. 2025):

$$\mathcal{L}_{D2A} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(-\text{Vol}(\mathbf{a}_i, \mathbf{m}_{2i}, \mathbf{m}_{3i})/\tau)}{\sum_{j=1}^{K} \exp(-\text{Vol}(\mathbf{a}_j, \mathbf{m}_{2i}, \mathbf{m}_{3i})/\tau)}, \tag{28}$$

$$\mathcal{L}_{A2D} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(-\text{Vol}(\mathbf{a}_i, \mathbf{m}_{2i}, \mathbf{m}_{3i})/\tau)}{\sum_{j=1}^{K} \exp(-\text{Vol}(\mathbf{a}_i, \mathbf{m}_{2j}, \mathbf{m}_{3j})/\tau)}. \tag{29}$$

Here, B is the batch size, K is the number of modalities, a_x refers to the embeddings of the anchor modality of the x-th sample in the batch, while m_{xy} refers to the embedding of the x-th modality of the j-th sample in the batch, τ is a learnable scaling parameter. As for LAM, the output features are supervised by the MSE loss as follows:

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2. \tag{30}$$

Here, x_i is the true value, $\hat{x}_i$ is the predict value. Finally, the overall loss $\mathcal{L}$ for our framework can be given by:

$$\mathcal{L} = \mathcal{L}_g + \alpha(\mathcal{L}_{D2A} + \mathcal{L}_{A2D}) + \beta\mathcal{L}_{MSE}. \tag{31}$$

Experiments

Datasets and Evaluation Metrics

Datasets. We evaluate the proposed method on three multi-modal object ReID benchmarks. RGBNT201 (Zheng et al. 2021) is a multi-modal person ReID dataset. It contains 4,787 RGB, NIR and TIR image triples, captured from 201 distinct identities. RGBNT100 (Li et al. 2020) is a large scale multi-modal vehicle ReID dataset. It comprises 17,250 image triples and encompasses a broad spectrum of visual scenarios. MSVR310 (Zheng et al. 2022) serves as a small scale multi-modal vehicle ReID dataset. It has 2,087 high quality image triples taken in various environments.

Evaluation Metrics. To evaluate the performance, we utilize the mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-K (K= $1, 5, 10$).

Implementation Details

The proposed model is implemented in PyTorch and trained using two NVIDIA GeForce RTX 3090 GPUs. We use the pre-trained CLIP (Radford et al. 2021) as the visual encoder. Images in triples are resized to 256×128 for RGBNT201, 128×256 for RGBNT100 and 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 and 4 images per identity respectively. For RGBNT100, the mini-batch size is 128 with 16 images per identity. We use the Adam optimizer (Kinga, Adam et al. 2015) to fine-tune the proposed modules with a learning rate of 3.5e^{-4} and the visual encoder with a relatively low learning rate of 5e^{-6}. We train the model for 50 epochs.

Methods	mAP	R-1	R-5	R-10
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. 2022b)	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. 2024c)	71.1	73.4	83.1	87.3
TOP-ReID* (Wang et al. 2024b)	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. 2024)	68.2	72.5	-	-
DeMo† (Wang et al. 2025b)	79.0	82.3	88.8	92.0
IDEA† (Wang et al. 2025c)	80.2	82.1	90.0	93.3
PromptMA† (Zhang et al. 2025)	78.4	80.9	87.0	88.9
Signal† (Ours)	**80.3**	**85.2**	**91.4**	**93.7**

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
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	55.2
LRMM (Wu et al. 2025)	78.6	96.7	36.7	49.7
PHT* (Pan et al. 2023)	79.9	92.7	-	-
HTT* (Wang et al. 2024c)	75.7	92.6	-	-
TOP-ReID* (Wang et al. 2024b)	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. 2024)	82.3	96.6	39.5	49.6
DeMo† (Wang et al. 2025b)	86.2	97.6	49.2	59.8
IDEA† (Wang et al. 2025c)	**87.2**	96.5	47.0	62.4
PromptMA† (Zhang et al. 2025)	85.3	97.4	**55.2**	64.5
Signal† (Ours)	86.3	**97.6**	53.6	**71.9**

Table 2: Performance on RGBNT100 and MSVR310.

Comparison with State-of-the-Art Methods

Multi-modal Person ReID. In Tab. 1, we compare our method with other methods on RGBNT201. Generally, multi-modal methods show a considerable improvement over single-modal methods by incorporating complementary information. Among these methods, models based on CLIP perform better. Specifically, our framework improves by 3.1% in Rank-1 accuracy compared to IDEA. This highlights the effectiveness of hierarchical alignment of different modalities. Besides, compared to DeMo, our method improves by 1.3% in mAP. These results confirm the effectiveness of our framework for multi-modal person ReID.

Multi-modal Vehicle ReID. In Tab. 2, we compare our method with other methods on the RGBNT100 and MSVR310 datasets. on RGBNT100, our method improves Rank-1 by 1.1% compared to IDEA. In comparison with EDITOR, our method achieves a 4.2% improvement in mAP and a 1.2% improvement in Rank-1. On MSVR310, our method improves mAP by 6.6% and Rank-1 by 9.5% compared to IDEA. These results indicate the effectiveness of our framework for multi-modal vehicle ReID.

Models	Modules			Metrics		Params	
	SIM	GAM	LAM	mAP	R-1	M	↑%
A	×	×	×	70.3	71.8	86.41	-
B	✓	×	×	77.0	80.6	89.56	3.65
C	✓	✓	×	79.0	82.8	89.56	3.65
D	✓	✓	✓	**80.3**	**85.2**	**91.17**	**5.51**

Table 3: Comparison with different modules on RGBNT201.

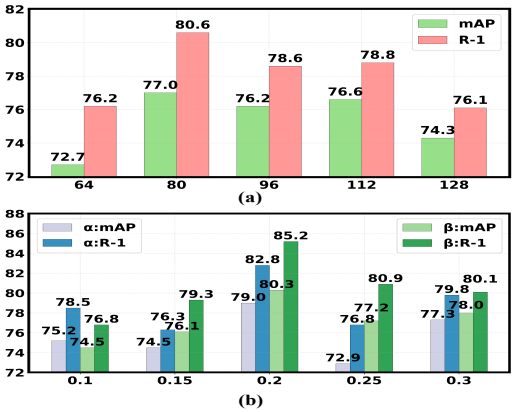

Figure 5: (a) Performance with different numbers of reserved tokens. (b) Performance with different α and β.

Ablation Studies

We evaluate the effectiveness of different modules on the RGBNT201 dataset. Our baseline employs a concatenated feature of tri-modal class tokens from the visual encoder.

Effects of Key Modules. Tab. 3 shows the performance comparison of different modules. Model A is the baseline model, achieving 70.3% mAP and 71.8% Rank-1 accuracy. After adding SIM, the performance of Model B improves to 77.0% mAP and 80.6% Rank-1. This indicates that removing useless image backgrounds is of great significance. Model C further integrates GAM, increasing mAP to 79.0% and Rank-1 to 82.8%, demonstrating the effectiveness of the modality alignment. Finally, Model D combines all modules and achieves the best results, with 80.3% mAP and 85.2% Rank-1. The complexity analysis shows that our proposed modules add only **4.76MB** learnable parameters in total. It achieves great performance gains with fewer parameters.

Effects of the Number of Reserved Tokens. Fig. 5 (a) demonstrates how the number of reserved patch tokens (k_1) affects the retrieval performance. The best result is observed when $k_1 = 80$. Thus, we set $k_1 = 80$ as default.

Effects of Loss Weights on GAM and LAM. Fig. 5 (b) presents ablation results of loss weights. When α is 0.2, adding GAM increases the mAP of the model to 79.0% and Rank-1 to 82.8%. When β is 0.2, adding LAM increases the mAP of the model to 80.3% and Rank-1 to 85.2%. Thus, we utilize these optimal weights as default settings.

Effect of Mask Intersection and Union. Tab. 4 presents a comparison investigation for the intersection and union of M_m^c and M_m^s within SIM. The results reveal that the union operation yields higher performances. The union operation delivers 1.1% higher in mAP than the intersection operation.

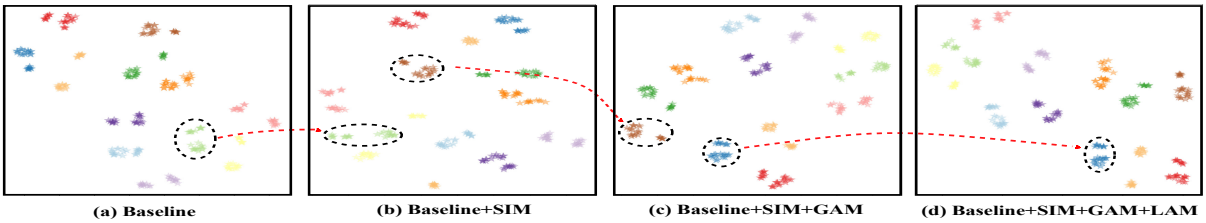

Figure 6: Visualization of the feature distributions with t-SNE. Different colors stand for different identities.

Methods	mAP	R-1	R-5	R-10
Intersection	79.2	82.3	89.5	93.7
Union	**80.3**	**85.2**	**91.4**	**93.7**

Table 4: Comparison of the intersection and union in SIM.

Methods	mAP	R-1	R-5	R-10
Sharing	74.6	78.3	86.6	92.0
Non-Sharing	**80.3**	**85.2**	**91.4**	**93.7**

Table 5: Comparison of whether offset is shared in LAM.

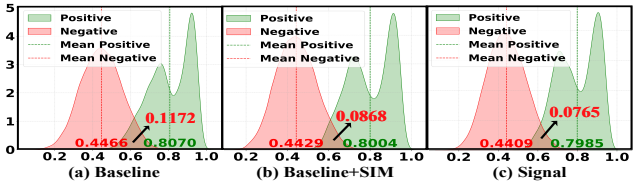

Figure 7: Visualization of the cosine similarity distribution.

Effect of Offset Sharing and Non-Sharing. Tab. 5 presents a comparative analysis with the offset sharing and non-sharing among RGB, NIR and TIR modalities in the LAM. Compared with the sharing offset, the non-sharing offset yields a 5.7% improvement in mAP and a 6.9% improvement in Rank-1. This indicates that the utilization of independent offsets for each modality exhibits greater efficacy.

Visualization Analysis

Multi-modal Feature Distributions with t-SNE. Fig. 6 shows the feature distribution with different modules. Comparing Fig. 6 (a) and (b), as instances with the same ID become more compact, removing redundant patches improves the feature discrimination ability. In Fig. 6 (c), by using GAM, the feature distribution is more compact than the one in Fig. 6 (b). In Fig. 6 (d), the feature distribution is more compact than the one in Fig. 6 (c), indicating that LAM enhances feature discrimination. These results demonstrate the effectiveness of our proposed modules.

Cosine Similarity Distributions. Fig. 7 shows the distributions of cosine similarities among test features. As observed, the intersected area of the two feature distributions is decreasing. It indicates our framework further amplifies the discrepancy between positive and negative samples.

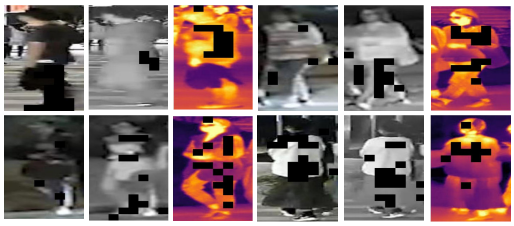

Figure 8: Visualization of token selection. Black blocks denote the removed image content.

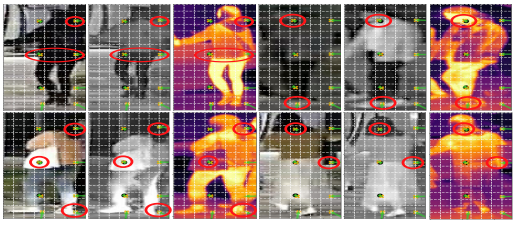

Figure 9: The visualization of generated offsets.

Token Selection in SIM. As shown in Fig. 8, some patches are removed from each modality. This indicates that intra-modal and inter-modal token selection achieve the expected goal of retaining important patch tokens.

Visualization of Generated Offsets. Fig. 9 demonstrates the alignment of local details across image triples for each object. The first one in the second row shows the model's effectiveness in aligning fine-grained details such as hair, shoulder bags and heels across modalities. It solves the problem of pixel-level shift. This validates the model's capability to learn local detail alignment.

Conclusion

In this paper, we propose a novel framework named **Signal** for multi-modal object ReID. Our method first uses a Selective Interaction Module (SIM) to select important patch tokens from multi-modal images. Then, we introduce the Global Alignment Module (GAM) to achieve feature alignment across multiple modalities. Finally, the Local Alignment Module (LAM) aligns important details within each modality in a shift-aware manner. As a result, our framework can extract more effective features for multi-modal object ReID. Extensive experiments on three benchmark datasets validate the effectiveness of our proposed method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62576069, 62506272), Natural Science Foundation of Liaoning Province (No.2025-MS-025) and Dalian Science and Technology Innovation Fund (No.2023JJ11CG001).

References

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

Chen, S.; Li, H.; Wang, Q.; Zhao, Z.; Sun, M.; Zhu, X.; and Liu, J. 2023. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *NeurIPS*, 36: 72842–72866.

Cicchetti, G.; Grassucci, E.; Sigillo, L.; Comminiello, D.; et al. 2025. Gramian multimodal representation learning and alignment. In *ICLR*.

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, W.; Zheng, D.; Yu, F.; Zhang, Y.; and Hou, Y. 2025. A novel approach to for multimodal emotion recognition: Multimodal semantic information fusion. *arXiv preprint arXiv:2502.08573*.

Feng, Y.; Li, J.; Xie, C.; Tan, L.; and Ji, J. 2025. Multimodal object re-identification via sparse mixture-of-experts. In *ICML*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *CVPR*, 15180–15190.

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. *IEEE TITS*, 24(11): 12431–12447.

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*.

Kinga, D.; Adam, J. B.; et al. 2015. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5. California;.

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

Li, H.; Yang, Z.; Zhang, Y.; Jia, W.; Yu, Z.; and Liu, Y. 2025a. MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE TPAMI*.

Li, S.; Li, C.; Zheng, A.; Lu, A.; Tang, J.; and Ma, J. 2025b. NEXT: Multi-grained mixture of experts via text-modulation for multi-modal object re-id. *arXiv preprint arXiv:2505.20001*.

Li, S.; Li, C.; Zheng, A.; Tang, J.; and Luo, B. 2025c. ICPL-ReID: Identity-conditional prompt learning for multi-spectral object re-identification. *arXiv preprint arXiv:2505.17821*.

Li, X.; Wu, Y.; Jiang, X.; Guo, Z.; Gong, M.; Cao, H.; Liu, Y.; Jiang, D.; and Sun, X. 2024. Enhancing visual document understanding with contrastive learning in large visual-language models. In *CVPR*, 15546–15555.

Li, Y.; Xing, Y.; Lan, X.; Li, X.; Chen, H.; and Jiang, D. 2025d. AlignMamba: Enhancing Multimodal Mamba with Local and Global Cross-modal Alignment. In *CVPR*, 24774–24784.

Lin, M.; Wang, S.; Wang, X.; Tang, J.; Fu, L.; Zuo, Z.; and Sang, N. 2025. DMPT: Decoupled modality-aware prompt tuning for multi-modal object re-identification. In *WACV*, 2103–2112.

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.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *NeurIPS*, 34: 14200–14213.

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.

Ruan, L.; Hu, A.; Song, Y.; Zhang, L.; Zheng, S.; and Jin, Q. 2023. Accommodating audio modality in clip for multimodal processing. In *AAAI*, 9641–9649.

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

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.

Wan, X.; Zheng, A.; Jiang, B.; Wang, B.; Li, C.; and Tang, J. 2025a. UGG-ReID: Uncertainty-guided graph model for multi-modal object te-identification. *arXiv preprint arXiv:2507.04638*.

Wan, X.; Zheng, A.; Wang, Z.; Jiang, B.; Tang, J.; and Ma, J. 2025b. Reliable multi-modal object re-identification via modality-aware graph reasoning. *arXiv preprint arXiv:2504.14847*.

Wang, F.; Zhou, Y.; Wang, S.; Vardhanabhuti, V.; and Yu, L. 2022a. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *NeurIPS*, 35: 33536–33549.

Wang, X.; Liu, F.; Li, Z.; and Guo, C. 2024a. Attribute-aware implicit modality alignment for text attribute person search. *arXiv preprint arXiv:2406.03721*.

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. 2024b. 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, 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. 2022b. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In *AAAI*, 2633–2641.

Wang, Z.; Zhao, Y.; Huang, H.; Liu, J.; Yin, A.; Tang, L.; Li, L.; Wang, Y.; Zhang, Z.; and Zhao, Z. 2023. Connecting multi-modal contrastive representations. *NeurIPS*, 36: 22099–22114.

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.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision transformer with deformable attention. In *CVPR*, 4794–4803.

Yu, H.-T.; and Song, M. 2024. Mm-point: Multi-view information-enhanced multi-modal self-supervised 3d point cloud understanding. In *AAAI*, 6773–6781.

Yu, Z.; Huang, Z.; Hou, M.; Pei, J.; Yan, Y.; Liu, Y.; and Sun, D. 2024. Representation selective coupling via token sparsification for multi-spectral object re-identification. *IEEE 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, R.; Xu, L.; Yang, S.; and Wang, L. 2024b. MambaReID: Exploiting vision mamba for multi-modal object re-identification. *Sensors*, 24(14): 4639.

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

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*, 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*, 13001–13008.