Magic Tokens: Select Diverse Tokens for Multi-modal Object Re-Identification

Pingping Zhang[1,2]*, Yuhao Wang[1], Yang Liu[1], Zhengzheng Tu[2,3] and Huchuan Lu[1]
[1] School of Future Technology, School of Artificial Intelligence, Dalian University of Technology, China
[2] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, China
[3] School of Computer Science and Technology, Anhui University, China
{zhpp,ly,lhchuan}@dlut.edu.cn, 924973292@mail.dlut.edu.cn, zhengzhengahu@163.com

Abstract

*Single-modal object re-identification (ReID) faces great challenges in maintaining robustness within complex visual scenarios. In contrast, multi-modal object ReID utilizes complementary information from diverse modalities, showing great potentials for practical applications. However, previous methods may be easily affected by irrelevant backgrounds and usually ignore the modality gaps. To address above issues, we propose a novel learning framework named **EDITOR** to select diverse tokens from vision Transformers for multi-modal object ReID. We begin with a shared vision Transformer to extract tokenized features from different input modalities. Then, we introduce a Spatial-Frequency Token Selection (SFTS) module to adaptively select object-centric tokens with both spatial and frequency information. Afterwards, we employ a Hierarchical Masked Aggregation (HMA) module to facilitate feature interactions within and across modalities. Finally, to further reduce the effect of backgrounds, we propose a Background Consistency Constraint (BCC) and an Object-Centric Feature Refinement (OCFR). They are formulated as two new loss functions, which improve the feature discrimination with background suppression. As a result, our framework can generate more discriminative features for multi-modal object ReID. Extensive experiments on three multi-modal ReID benchmarks verify the effectiveness of our methods. The code is available at https://github.com/924973292/EDITOR.*

1. Introduction

Object re-identification (ReID) aims to retrieve specific objects (e.g., person, vehicle) across non-overlapping cameras. Over the past few decades, object ReID has advanced significantly. However, traditional object ReID with single-modal input encounters substantial challenges [17], particularly in complex visual scenarios, such as extreme illumination, thick fog and low image resolution. It can result in noticeable distortions in critical object regions, leading to disruptions during the retrieval process [53]. Therefore, there has been a notable shift toward multi-modal approaches in recent years, capitalizing on diverse data sources to enhance the feature robustness for practical applications [43, 44, 53]. However, as illustrated in Fig. 1, previous multi-modal ReID methods typically extract global features from all regions of images in different modalities and subsequently aggregate them. Nevertheless, these

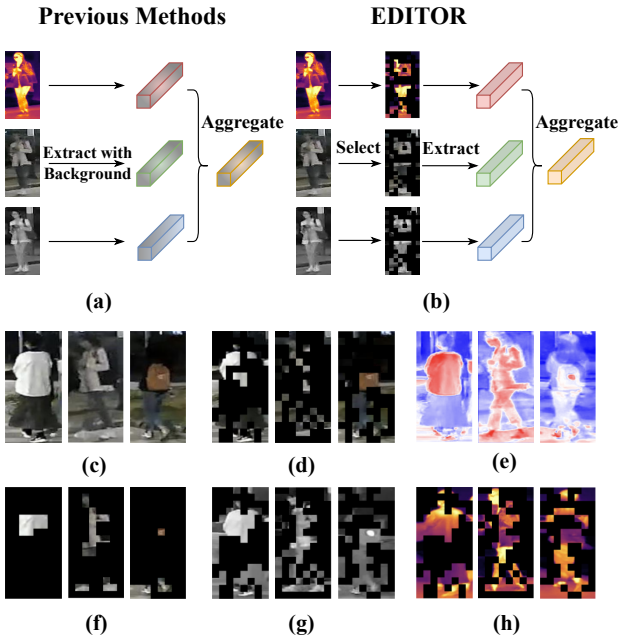

Figure 1. Comparison of different methods and token selections. (a) Framework of previous methods; (b) Framework of our proposed EDITOR; (c) RGB images; (d) Spatial-based token selection; (e) Multi-modal frequency transform; (f) Frequency-based token selection; (g) Selected tokens in the NIR modality; (h) Selected tokens in the TIR modality.

*Corresponding author

methods present two key limitations: (1) Within individual modalities, backgrounds introduce additional noise [37], especially in challenging visual scenarios. (2) Across different modalities, backgrounds introduce overhead in reducing modality gaps, which may amplify the difficulty in aggregating features [15]. Hence, our method prioritizes the selection of object-centric information, aiming to preserve the diverse features of different modalities while minimizing background interference.

To address above issues, we propose a novel feature learning framework named EDITOR to select diverse tokens for multi-modal object ReID. Our EDITOR comprises two key modules: Spatial-Frequency Token Selection (SFTS) and Hierarchical Masked Aggregation (HMA). Technically, we begin with a shared vision Transformer (ViT) [7] to extract tokenized features from different input modalities. Then, SFTS employs a dual approach to select object-centric tokens from both spatial and frequency perspectives. In the spatial-based token selection, we combine all spatial indices selected by various heads in multi-head self-attention [7] within each modality. Afterwards, we further combine the indices from different modalities to enhance the token diversity across modalities. However, as shown in Fig. 1 (d), the spatial-based token selection may not fully capture all object-centric tokens. Therefore, we incorporate a frequency-based token selection to collaboratively extract the most salient tokens, as shown in Fig. 1 (e)-(f). With the selected tokens, we introduce HMA to effectively aggregate object-centric tokens within and across modalities. To further reduce the effect of backgrounds, we propose a Background Consistency Constraint (BCC) and an Object-Centric Feature Refinement (OCFR). They are formulated as two new loss functions, which improve the feature discrimination with background suppressions. With the proposed modules, our framework can extract more discriminative features for multi-modal object ReID. Experiments on the three multi-modal object ReID benchmarks, i.e., RGBNT201, RGBNT100 and MSVR310 demonstrate the effectiveness of our proposed EDITOR.

In summary, our contributions are as follows:

- We introduce EDITOR, a novel feature learning framework for multi-modal object ReID. To our best knowledge, it is the first attempt to enhance multi-modal object ReID through object-centric token selection.
- We propose a Spatial-Frequency Token Selection (SFTS) module and a Hierarchical Masked Aggregation (HMA) module. These modules effectively facilitate the selection and aggregation of multi-modal tokenized features.
- We propose two new loss functions with a Background Consistency Constraint (BCC) and an Object-Centric Feature Refinement (OCFR) to improve the feature discrimination with background suppressions.
- Extensive experiments are performed on three multi-modal object ReID benchmarks. The results fully validate the effectiveness of our proposed methods.

2. Related Work

2.1. Single-modal Object ReID

Single-modal object ReID extracts discriminative features from single-modality inputs, such as RGB, Near Infrared (NIR), Thermal Infrared (TIR), or depth images. Most of existing object ReID methods are based on Convolutional Neural Networks (CNNs) or Transformers [39]. Regarding CNN-based methods, PCB [34] and MGN [40] employ a part-based image partitioning approach to extract features at multiple levels of granularity. In addition, with a unified aggregation gate mechanism, OSNet [57] dynamically fuses features across omni-scales. DMML [3] offers a meta-level view of metric learning, demonstrating the alignment of softmax and triplet losses in the meta space. Circle loss [35] introduces a novel approach to re-weight similarity scores and achieve a more flexible optimization. AGW [47] extracts fine-grained features with non-local attention mechanisms. However, CNN-based methods [5, 20, 22, 42] may not be sufficiently robust in complex scenarios due to their limited receptive field. Drawing inspiration from the success of ViT [7], the first pure Transformer-based method named TransReID [13] is proposed with the adaptive modeling of image patches, yielding competitive results. Furthermore, AAformer [59] introduces an automated alignment strategy to extract local features. DCAL [58] proposes a dual cross-attention method, which enhances self-attention with global-local and pairwise cross-attentions. PHA [51] improves ViTs by enhancing high-frequency feature representations through a patch-wise contrastive loss. Moreover, a multitude of Transformer-based approaches, as presented in works [19, 21, 25, 41, 45, 49, 50, 60], showcase their benefits in object ReID. Nevertheless, these approaches rely on single-modal input, which offer limited representation capabilities, especially in complex scenarios. In contrast, our proposed EDITOR integrates diverse modalities and leverages token selections, enabling to capture more fine-grained features in a variety of scenarios.

2.2. Multi-modal Object ReID

Aggregating robust representations from multi-modal data has attracted considerable attention in recent years. In the field of multi-modal person ReID, Zheng *et al.* [53] first design a PFNet to learn robust features with progressive fusion. Wang *et al.* [44] advance the field further with their IEEE framework, which employs three learning strategies to enhance modality-specific representations. Then, Zheng *et al.* [55] introduce the pixel-level reconstruction to address the modal-missing problem. For multi-modal vehicle ReID, Li *et al.* [17] propose a HAMNet to fuse dif-

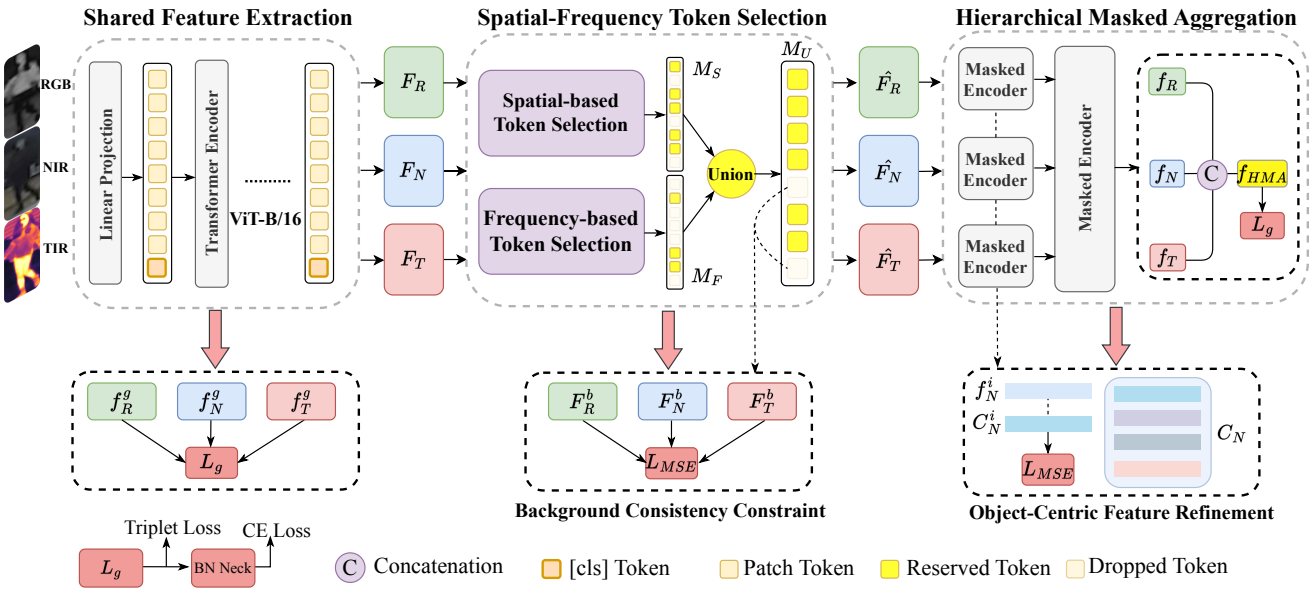

Figure 2. An illustration of our proposed EDITOR. First, features from different input modalities are extracted by using the shared ViT-B/16 backbone. Then, a Spatial-Frequency Token Selection (SFTS) is utilized to select diverse tokens with object-centric features. Meanwhile, the Background Consistency Constraint (BCC) loss is designed for stabilizing the selection process. After that, a Hierarchical Masked Aggregation (HMA) is grafted to aggregate the selected tokens. Finally, combined with the Object-Centric Feature Refinement (OCFR) loss, the whole framework can obtain more discriminative features for multi-modal object ReID.

ferent modal features with a heterogeneous score coherence loss. Then, Zheng *et al.* [54] reduce the discrepancies from sample and modality aspects. From the perspective of generating modalities, Guo *et al.* [9] propose a GAFNet to fuse the multiple data sources. He *et al.* [12] propose a GPFNet to adaptively fuse multi-modal features with graph learning. With Transformers, Pan *et al.* [29] introduce a PHT, employing a feature hybrid mechanism to balance modal-specific and modal-shared information. Jennifer *et al.* [4] provide a UniCat by analyzing the issue of modality laziness. Very recently, Wang *et al.* [43] propose a novel token permutation mechanism for robust multi-modal object ReID. While contributing to the multi-modal object ReID, they commonly overlook the influence of irrelevant backgrounds on the aggregation of features across different modalities. In contrast, our proposed EDITOR explicitly addresses the influence of irrelevant backgrounds on multi-modal feature aggregation. Our approach effectively identifies critical regions within each modality while fostering inter-modal collaboration. Furthermore, the incorporation of BCC and OCRF losses, along with the innovative SFTS and HMA modules, distinguishes our work as a promising avenue for improved performance in complex scenarios.

2.3. Token Selection in Transformer

With the increasing adoption of Transformers [16, 24, 31], token selection has gained significant attention [1, 8, 10, 11, 23, 28, 33, 46], due to its ability to focus on essential objects and reduce computational overhead. In vision tasks, such as ReID, where fine-grained features are crucial, the extraction of key regions becomes particularly important. For example, TransFG [11] utilizes the multi-head self-attention of ViT to select representative local patches, achieving outstanding performance in fine-grained classification tasks. DynamicViT [33] employs gating mechanisms to dynamically accelerate both training and inference. TVTR [46] extends token selection to cross-modal ReID, aligning features by selecting the top-K salient tokens. However, our method differs from them in the following ways: (1) Our selection is **instance-level**, where for different input images, the model dynamically selects different numbers of object-centirc tokens. Unlike previous methods, which specify the fixed top-K local regions for feature aggregation, our approach allows the model to adapt more flexibly to various inputs. (2) Previous methods do not consider the impact of **distracted backgrounds** during the early selection process. With our proposed losses, we effectively stabilize the selection process, achieving dynamic distribution alignments. Thus, we provide a more flexible framework, ultimately enhancing ReID performance in complex scenarios.

3. Proposed Method

As illustrated in Fig. 2, our proposed EDITOR comprises three key components: Shared Feature Extraction, Spatial-Frequency Token Selection (SFTS) and Hierarchical Masked Aggregation (HMA). In addition, we incorporate the Background Consistency Constraint (BCC) and

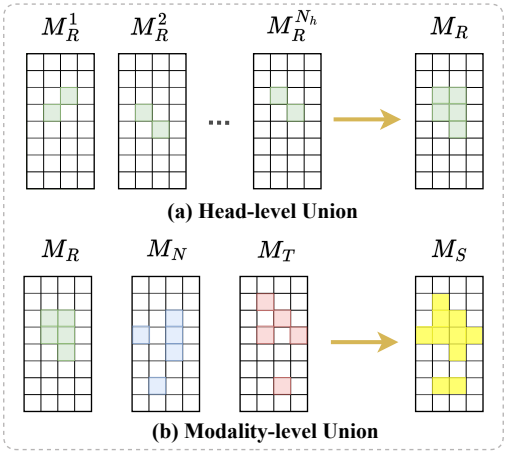

Figure 3. Illustration of spatial-based token selection.

Object-Centric Feature Refinement (OCFR) to further reduce the effect of irrelevant backgrounds. We will describe these key modules in the following subsections.

3.1. Shared Feature Extraction

To extract multi-modal features while reducing the model parameters, we deploy a shared vision Transformer (ViT) for multi-modal inputs. Without loss of generality, for the RGB, NIR and TIR modalities, the multi-modal tokenized features can be expressed as:

$$F_R = \mathrm{ViT}\left(I_R\right), F_N = \mathrm{ViT}\left(I_N\right), F_T = \mathrm{ViT}\left(I_T\right), \quad (1)$$

where I_R, I_N and I_T represent the input RGB, NIR and TIR images, respectively. The tokenized features F_R, F_N and F_T, each of which has a shape of $\mathbb{R}^{D\times(N_p+1)}$, are extracted from the last layer of ViT. Here, we follow previous works and employ additional learnable class tokens f_R^g, f_N^g and f_T^g for corresponding modalities. N_p means the number of patch tokens while D is the embedding dimension.

3.2. Spatial-Frequency Token Selection

To preserve diverse information within and across modalities while eliminating the influence of irrelevant backgrounds, we propose the Spatial-Frequency Token Selection (SFTS) module. It consists of spatial-based token selection and frequency-based token selection. Through the collaboration of these two selection methods, our EDITOR can focus on the critical regions of the object.

Spatial-based Token Selection. As shown in Fig. 3, our spatial-based token selection is enhanced by both head-level union and modality-level union. This kind of combinations facilitate a dynamic selection of instance-level tokens, preserving diverse information across different input modalities. Technically, the spatial-based token selection takes tokens from the three modalities and ultimately produces a selection mask M_S for all three modalities. Taking the RGB modality as an example, assuming there are a total of K layers in the backbone network, and there are N_h heads in the self-attention layer, the attention weights of the k-th layer for the RGB modality can be represented as follows:

$$a_k^i = \left[a_k^{i_0}, a_k^{i_1}, a_k^{i_2}, \cdots, a_k^{i_{N_p}}\right], \quad i \in 1, 2, \cdots, N_h, \quad (2)$$

where a_k^i is the attention weight in the i-th head of the k-th layer. $a_k^{i_0}$ is the corresponding weight of the class token. Thus, the attention weights of all layers are organized as:

$$A_k = \left[a_k^1, a_k^2, a_k^3, \cdots, a_k^{N_h}\right], \quad k \in 1, 2, \cdots, K. \quad (3)$$

To further concentrate attention on objects, we follow [11] to integrate attention weights from all the preceding layers. Specifically, the attention score is iteratively computed through a matrix multiplication in the following manner:

$$A_{score} = \prod_{k=1}^{K} A_k. \quad (4)$$

Here, A_{score} represents the comprehensive relationships between patches. Then, we extract the weights associated with the class token $a_{score}^{i_0}$ from each head in A_{score}. For each head, we retain the crucial tokens and generate the mask M_R^i. This process can be formalized as:

$$M_R^i = \mathrm{Mask}(\mathrm{Top_s}(a_{score}^{i_0})), \quad i \in 1, 2, \cdots, N_h, \quad (5)$$

where Mask represents transforming the selected tokens into a mask form, and $\mathrm{Top_s}$ retains the top s important tokens ($s \in N^+$). In multi-head self-attention, different heads focus on different aspects. To capture more details within modality, we employ **head-level union** to combine the selected tokens from different heads. Finally, we obtain the mask of RGB modality M_R, which can be formulated as:

$$M_R = \bigcup_{i=1}^{N_h} M_R^i. \quad (6)$$

For other modalities, we execute similar operations as:

$$M_N = \bigcup_{i=1}^{N_h} M_N^i, M_T = \bigcup_{i=1}^{N_h} M_T^i. \quad (7)$$

Based on above operations, we complete the selection process within individual modalities. As a result, each modality chooses tokens that focus on objects and eliminate most background interferences. However, the significant variations across modalities lead to challenges when directly aggregating these tokenized features. To address this issue and facilitate the modality complementary, we further introduce **modality-level union**. It can be expressed as follows:

$$M_S = M_R \cup M_N \cup M_T, \quad (8)$$

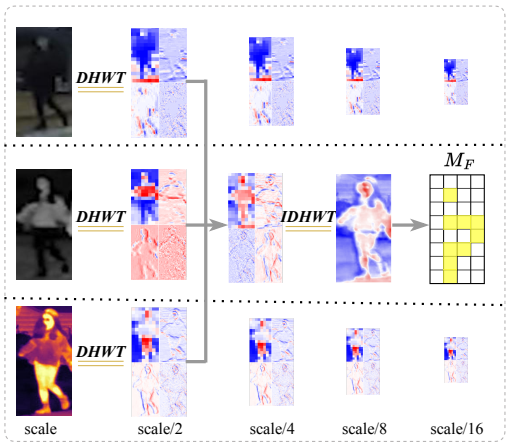

Figure 4. Illustration of frequency-based token selection.

where M_S means the final mask from spatial-based token selection. Through the head-level union and modality-level union, we achieve an instance-level token selection strategy, providing diverse tokenized features for modal fusion.

Frequency-based Token Selection. As shown in Fig. 1 (d), the spatial-based token selection may result in the neglect of some salient tokens. Considering the frequency information could provide a structural perception of images, we introduce a frequency-based token selection to mine more salient tokens. As shown in Fig. 4, we apply the Discrete Haar Wavelet Transform (DHWT) [27] to the images of different modalities. Taking the RGB modality as an example, we obtain four frequency components:

$$I_R^{ll}, I_R^{lh}, I_R^{hl}, I_R^{hh} = \mathrm{DHWT}(I_R), \tag{9}$$

where I_R^{ll} is the low-frequency component, while the other three terms are the high-frequency components. The same operations are carried out for other modalities. It can be observed that the decomposition results of different modalities exhibit significant frequency differences. Technically, we first sum up the decomposition results from different modalities at each scale. Then, we perform the Inverse Discrete Haar Wavelet Transform (IDHWT) with the summed results to obtain the inverse transformed image. As shown in the middle of Fig. 4, the inverse transformed image highlights salient regions. Finally, we use a sliding window to count the pixel values within each patch. The tokens of top f values are selected as the frequency-based selection result, denoted as M_F. As a result, by uniting spatial mask M_S and frequency mask M_F, we obtain the final mask M_U.

Background Consistency Constraint. Most of previous methods [11, 46] directly discard non-selected tokens, treating them as backgrounds. However, this may lead to the loss of important information. Therefore, we step further to impose consistency constraints on these non-selected tokens from different modalities. Formally, the background mask $M_B \in \mathbb{R}^{N_p}$ is defined as:

$$M_B = 1 - M_U. \tag{10}$$

Then the background tokens from different modalities can be represented as follows:

$$F_R^b = F_R^p \odot M_B, F_N^b = F_N^p \odot M_B, F_T^b = F_T^p \odot M_B, \tag{11}$$

where F_R^p, F_N^p and F_T^p represents the patch features of F_R, F_N and F_T, respectively. M_B is the background indices. $\odot$ denotes the element-wise multiplication. Then, the background tokens from paired modalities are constrained by the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\mathrm{R2N}} = \sum_{i=1}^{N_r} ||F_R^b - F_N^b||_2^2, \tag{12}$$

$$\mathcal{L}_{\mathrm{R2T}} = \sum_{i=1}^{N_r} ||F_R^b - F_T^b||_2^2, \tag{13}$$

$$\mathcal{L}_{\mathrm{N2T}} = \sum_{i=1}^{N_r} ||F_N^b - F_T^b||_2^2. \tag{14}$$

The final consistency constraint loss can be formulated as:

$$\mathcal{L}_{BCC} = \frac{1}{N_r}(\mathcal{L}_{\mathrm{R2N}} + \mathcal{L}_{\mathrm{R2T}} + \mathcal{L}_{\mathrm{N2T}}), \tag{15}$$

where N_r is the number of reserved tokens. With $\mathcal{L}_{BCC}$, we can achieve dynamic alignments of backgrounds and stabilize the token selection process.

3.3. Hierarchical Masked Aggregation

For enhancing the feature robustness, we introduce the Hierarchical Masked Aggregation (HMA) to effectively aggregate selected diverse tokens from different modalities. More specifically, the HMA consists of independent aggregation and collaborative aggregation. In the independent aggregation stage, each modality interacts with its selected tokens, highlighting specific regions and improving the feature discrimination. In the collaborative aggregation stage, tokens from all modalities interact with each other, facilitating the exchange and fusion of multi-modal information.

Independent Aggregation. Without loss of generality, taking the RGB modality as an example, we first concatenate f_R^g with selected tokens to form $\hat{F_R} \in \mathbb{R}^{D\times(N_p+1)}$. Then, it is fed into a masked encoder for feature interaction:

$$\overline{F}_R = \Theta(\hat{F_R}), \hat{F_R} = [f_R^g, F_R^p \odot M_U], \tag{16}$$

where $[\cdot]$ is the concatenation operation. $\overline{F}_R \in \mathbb{R}^{D\times(N_p+1)}$ represents the aggregated features. The masked encoder Θ is essentially a Transformer block with a Multi-Head Self-Attention (MHSA) [7] and a Feed-forward Neural Network (FFN) [7]. Thus, we obtain tokenized features aligned with object-centric regions. Other modalities are processed as:

$$\overline{F}_T = \Theta(\hat{F_T}), \overline{F}_N = \Theta(\hat{F_N}). \tag{17}$$

As a result, global features of each modality will further focus on key tokens, obtaining object-centric features.

Collaborative Aggregation. To collaboratively aggregate tokens from different modalities, we first concatenate $\overline{F}_R$, $\overline{F}_N$ and $\overline{F}_T$ along the token dimension to form $\overline{F} \in \mathbb{R}^{D\times 3(N_p+1)}$. Then, it is fed into Θ for feature interaction:

$$F_{Agg} = \Theta(\overline{F}), \overline{F} = [\overline{F}_R, \overline{F}_N, \overline{F}_T], \tag{18}$$

where $F_{Agg} \in \mathbb{R}^{D\times 3(N_p+1)}$ is the aggregated feature from different modalities. Then, we extract class tokens f_R, f_N and f_T from F_{Agg}, and concatenate them as the output $f_{HMA} \in \mathbb{R}^{3D}$ of HMA. With HMA, we achieve an adaptive token aggregation within and across modalities, enhancing the discriminative ability of multi-modal features.

Object-Centric Feature Refinement. After suppressing background interferences, we propose the Object-Centric Feature Refinement to enhance the aggregation of intra-modal features. As shown in the bottom right corner of Fig. 2, we construct and update the feature center for the i-th ID. Without loss of generality, taking the NIR modality as an example, this is achieved by first computing the averaged feature belonging to the i-th ID in the current mini-batch:

$$f_N^i = \frac{1}{P}\sum_{y_i=i}(\overline{F}_N^{cls}[\text{label} = y_i]), \tag{19}$$

where y_i is the label of the current feature, and P is the number of instances in each ID in a mini-batch. $\overline{F}_N^{cls}$ is the class token of $\overline{F}_N$. Then, the updated center is:

$$C_N^i|_{iter} := \alpha f_N^i + (1-\alpha)C_N^i|_{iter-1}, \tag{20}$$

where α is the exponential decay rate, and $iter$ denotes the current iteration. Furthermore, by using a MSE loss, we ensure that features belonging to the same ID are pulled closer to the ID center. This can be represented as follows:

$$\mathcal{L}_N = \frac{1}{B}\sum_i\sum_{y_i=i}||\overline{F}_N^{cls}[\text{label} = y_i] - C_N^i||_2^2, \tag{21}$$

where B represents the batch size. Similarly, the features from the RGB and TIR modalities will align with their respective centers, resulting in the following loss:

$$\mathcal{L}_{OCFR} = \mathcal{L}_R + \mathcal{L}_N + \mathcal{L}_T. \tag{22}$$

3.4. Objective Function

As illustrated in Fig. 2, our objective function comprises four components: losses for the ViT backbone, HMA, BCC and OCFR. For the backbone and HMA, they are both supervised by the label smoothing cross-entropy loss [36] and triplet loss [14] with equal weights:

$$\mathcal{L}_g = \mathcal{L}_{ce} + \mathcal{L}_{tri}. \tag{23}$$

Finally, the total loss for our framework can be defined as:

$$\mathcal{L}_{total} = \mathcal{L}_g^{ViT} + \mathcal{L}_g^{HMA} + \mathcal{L}_{BCC} + \mathcal{L}_{OCFR}. \tag{24}$$

4. Experiments

4.1. Dataset and Evaluation Protocols

To evaluate the performance of our method, we employ three multi-modal object ReID benchmarks. More specifically, RGBNT201 [53] is the first multi-modal person ReID dataset encompassing RGB, NIR, and TIR modalities. RGBNT100 [17] is a large-scale multi-modal vehicle ReID dataset. MSVR310 [54] is a small-scale multi-modal vehicle ReID dataset with complex visual scenarios. As for evaluation metrics, we follow previous works and utilize the mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-K ($K = 1, 5, 10$).

4.2. Implementation Details

Our model is implemented by using the PyTorch toolbox. Experiments are conducted on two NVIDIA A100 GPUs. We employ pre-trained Transformers from the ImageNet classification dataset [6] as our backbones. For data processing, images are resized to 256×128 for RGBNT201 and 128×256 for RGBNT100/MSVR310. During the training process, we employ random horizontal flipping, cropping, and erasing [56] for data augmentation. The mini-batch size is set to 128, containing 8 randomly selected object identities, and 16 images sampled for each identity. To optimize our model, we use the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.0001. The learning rate is initialized at 0.001 and follows a warmup strategy with a cosine decay. In spatial-based token selection, s is set to 2, while in frequency-based token selection, f is set to 10. For the OCFR, we set α to 0.8.

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

We perform comparisons with state-of-the-art methods on three multi-modal ReID datasets. Our method demonstrates competitive results compared with previous methods.

Multi-modal Person ReID. As presented in Tab. 1, we compare EDITOR with both single-modal and multi-modal methods on RGBNT201. In general, single-modal methods tend to exhibit lower performance. Among the single-modal methods, PCB [34] stands out with an impressive mAP of 32.8%, showcasing the effectiveness of its part-based matching strategy. For multi-modal methods, TOP-ReID (B) [43] achieves a remarkable mAP of 64.6%. However, our EDITOR (B) with a mAP of 65.7%, outperforms TOP-ReID (B), delivering an 1.1% improvement. Moreover, there is a noticeable improvement in the rank metrics, indicating the effectiveness of our method in addressing the

Table 1. Performance comparison on three multi-modal object ReID benchmarks. The best and second results are in bold and underlined, respectively. *denotes Transformer-based methods, while the rest are CNN-based methods. Both single-modal and multi-modal methods are included. For the comparison between TOP-ReID and EDITOR, A and B means the AL setting and BL setting [43], respectively.

(a) Comparison on RGBNT201.

Methods		RGBNT201			
		mAP	R-1	R-5	R-10
Single	MUDeep [30]	23.8	19.7	33.1	44.3
	HACNN [18]	21.3	19.0	34.1	42.8
	MLFN [2]	26.1	24.2	35.9	44.1
	PCB [34]	32.8	28.1	37.4	46.9
	OSNet [57]	25.4	22.3	35.1	44.7
	CAL [32]	27.6	24.3	36.5	45.7
Multi	HAMNet [17]	27.7	26.3	41.5	51.7
	PFNet [53]	38.5	38.9	52.0	58.4
	IEEE [44]	49.5	48.4	59.1	65.6
	DENet [55]	42.4	42.2	55.3	64.5
	UniCat* [4]	57.0	55.7	-	-
	TOP-ReID (A)* [43]	**72.3**	**76.6**	**84.7**	**89.4**
	TOP-ReID (B)* [43]	64.6	64.6	77.4	82.4
	EDITOR (A)*	<u>66.5</u>	68.3	81.1	88.2
	EDITOR (B)*	65.7	<u>68.8</u>	<u>82.5</u>	<u>89.1</u>

(b) Comparison on RGBNT100 and MSVR310.

Methods		RGBNT100		MSVR310	
		mAP	R-1	mAP	R-1
Single	PCB [34]	57.2	83.5	23.2	42.9
	MGN [40]	58.1	83.1	26.2	44.3
	DMML [3]	58.5	82.0	19.1	31.1
	BoT [26]	78.0	95.1	23.5	38.4
	OSNet [57]	75.0	95.6	28.7	44.8
	Circle Loss [35]	59.4	81.7	22.7	34.2
	HRCN [52]	67.1	91.8	23.4	44.2
	AGW [47]	73.1	92.7	28.9	46.9
	TransReID* [13]	75.6	92.9	18.4	29.6
Multi	HAMNet [17]	74.5	93.3	27.1	42.3
	PFNet [53]	68.1	94.1	23.5	37.4
	GAFNet [9]	74.4	93.4	-	-
	CCNet [54]	77.2	96.3	<u>36.4</u>	**55.2**
	GraFT* [48]	76.6	94.3	-	-
	GPFNet [12]	75.0	94.5	-	-
	PHT* [29]	79.9	92.7	-	-
	UniCat* [4]	79.4	96.2	-	-
	TOP-ReID (A)* [43]	73.7	92.2	30.2	33.7
	TOP-ReID (B)* [43]	<u>81.2</u>	<u>96.4</u>	35.9	44.6
	EDITOR (A)*	79.8	93.9	35.8	43.1
	EDITOR (B)*	**82.1**	**96.4**	**39.0**	<u>49.3</u>

challenges of multi-modal person ReID. Although showing inferior performance than TOP-ReID (A), EDITOR (A) is more robust across different settings, potentially addressing the modality laziness problem [4]. Besides, EDITOR has fewer parameters than TOP-ReID, making it more efficient.

Multi-modal Vehicle ReID. As shown in Tab. 1, single-modal methods generally exhibit lower performance compared with multi-modal methods. In single-modal methods, CNN-based methods like AGW [47], OSNet [57] and BoT [26] consistently achieve better results across datasets. While Transformer-based methods, such as TransReID [13], exhibit slightly inferior performance, especially on smaller datasets like MSVR310, where they lag behind CNN-based methods. However, Transformer-based methods prove their effectiveness in integrating multi-modal data. Specifically, TOP-ReID (B) [43] achieves a mAP of 81.2% on RGBNT100. Our EDITOR (B) surpasses TOP-ReID (B) on RGBNT100, demonstrating a 0.9% higher mAP. Notably, our improvement over TransReID on the smaller dataset MSVR310 highlights our model's resilience to over-fitting. Meanwhile, our EDITOR (B) achieves a 2.6% higher mAP than CCNet. These results verify the effectiveness of our method in multi-modal vehicle ReID.

4.4. Ablation Studies

We conduct ablation studies on the RGBNT201 dataset to validate the proposed components. Our baseline utilizes a ViT-B/16 with camera embeddings, supervised by $\mathcal{L}_g^{ViT}$.

Table 2. Performance comparison with different components.

	Module		Loss		RGBNT201			
	SFTS	HMA	BCC	OCFR	mAP	R-1	R-5	R-10
A	×	×	×	×	54.0	53.5	70.2	78.8
B	×	✓	×	×	60.7	62.4	77.2	83.6
C	✓	✓	×	×	62.2	65.0	79.3	85.4
D	✓	✓	✓	×	65.2	65.9	82.2	87.1
E	✓	✓	×	✓	64.8	66.9	82.3	87.3
F	✓	✓	✓	✓	**65.7**	**68.8**	**82.5**	**89.1**

Effect of Key Components. Tab. 2 shows the performance comparison with different components. The model A is the baseline. Model B incorporates HMA, resulting in a 6.7% increase in mAP, demonstrating the effectiveness in aggregating multi-modal features. Furthermore, Model C introduces SFTS, achieving further performance improvement through object-centric token selection. The introduction of BCC effectively achieves dynamic alignments of multi-modal distributions, resulting in a 3% mAP improvement compared with Model C. Besides, Model E makes the feature distribution more compact, leading to robust improvements. By integrating all components, our model achieves the optimal performance. These results validate the effectiveness of our EDITOR in complex scenarios.

Table 3. Comparison between modality union and separation. The BCC and OCFR losses are not added here.

Methods	RGBNT201			
	mAP	R-1	R-5	R-10
w/o selection	60.7	62.4	77.2	83.6
w/ separation	57.7	58.5	75.4	82.5
w/ union	**62.2**	**65.0**	**79.3**	**85.4**

Table 4. Effect of different selection methods in SFTS.

Selection Methods	Reserved Tokens	RGBNT201	
	Average number	mAP	R-1
Modality	30.2	64.2	65.7
Spatial	55.0	65.0	66.8
Frequency	55.0	64.1	65.3
Spatial+Frequency	58.0	**65.7**	**68.8**

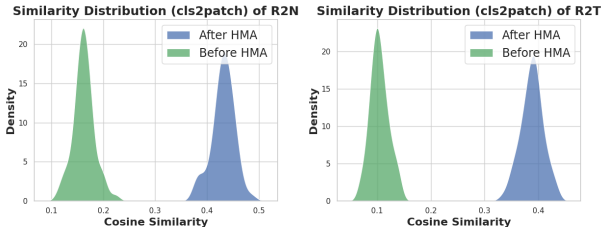

Figure 5. Alignment visualization in HMA with RGB modality.

Effect of Modality Union vs. Separation. As shown in Tab. 3, we verify the effect of using modality union. The results show that modality union significantly improves the performance. Selected tokens from different modalities vary significantly, potentially causing instability in the subsequent aggregation. This is evident in the second row of Tab. 3. Therefore, by establishing shared indices, our modality union enables a collaborative interaction among different modalities, providing a more stable aggregation.

Effect of Different Selection Methods. In Tab. 4, we validate different selection methods in SFTS. The first row is modality union, and the second row introduces head union, forming the complete spatial-based token selection. The introduction of head union increases retained tokens, leading to better performance. In contrast, frequency-based token selection shows inferior results. The best results are achieved by combining them.

4.5. Visualization

Feature Alignment of HMA. In Fig. 5, we measure the cosine similarity between class tokens of different modalities before and after HMA. The results show that, after HMA, the class token of the RGB modality effectively aligns with patch tokens of other modalities. Similar results can be observed for other modalities, confirming the effectiveness of HMA on feature alignment and aggregation. More visualizations are provided in the supplementary material.

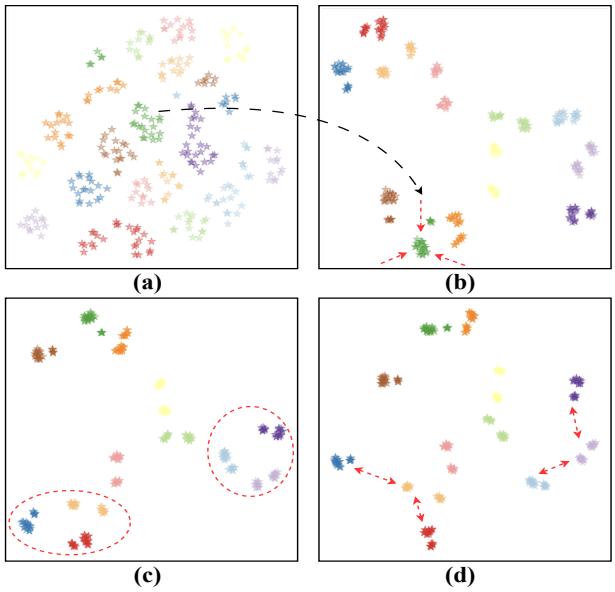

Figure 6. Comparison of feature distributions with t-SNE [38]. Different colors represent different identities. (a) Baseline; (b) Baseline + SFTS + HMA; (c) Baseline + SFTS + HMA + OCFR; (d) Baseline + SFTS + HMA + OCFR + BCC.

Feature Distributions. Fig. 6 shows the feature distributions with different components. When comparing Fig. 6(a) and Fig. 6(b), one can observe that SFTS and HMA can pull features to their ID clusters and increase gaps between different IDs. As shown in Fig. 6(c), with OCFR, it obtains more compact features within the same ID, effectively enhancing feature distinctiveness. Finally, with BCC, it enlarges gaps between different IDs. These visualizations vividly verify the effectiveness of different modules.

5. Conclusion

In this work, we propose EDITOR, a novel feature learning framework that selects diverse tokens from vision Transformers for multi-modal object ReID. Our framework integrates Spatial-Frequency Token Selection (SFTS) and Hierarchical Masked Aggregation (HMA), which select and aggregate multi-modal features, respectively. To reduce the effect of backgrounds, we introduce Background Consistency Constraint (BCC) and Object-Centric Feature Refinement (OCFR) losses. Extensive experiments on three benchmarks validate the effectiveness of our method.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China (No.62101092), Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No.MMC202102) and Fundamental Research Funds for the Central Universities (No.DUT22QN228 and No.DUT23BK050).

References

[1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *ICLR*, 2022. 3

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, pages 2109–2118, 2018. 7

[3] Guangyi Chen, Tianren Zhang, Jiwen Lu, and Jie Zhou. Deep meta metric learning. In *ICCV*, pages 9547–9556, 2019. 2, 7

[4] Jennifer Crawford, Haoli Yin, Luke McDermott, and Daniel Cummings. Unicat: Crafting a stronger fusion baseline for multimodal re-identification. *arXiv preprint arXiv:2310.18812*, 2023. 3, 7

[5] Ju Dai, Pingping Zhang, Dong Wang, Huchuan Lu, and Hongyu Wang. Video person re-identification by temporal residual learning. *TIP*, 28(3):1366–1377, 2019. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[8] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *ECCV*, pages 396–414. Springer, 2022. 3

[9] Jinbo Guo, Xiaojing Zhang, Zhengyi Liu, and Yuan Wang. Generative and attentive fusion for multi-spectral vehicle re-identification. In *ICSP*, pages 1565–1572, 2022. 3, 7

[10] Joakim Bruslund Haurum, Sergio Escalera, Graham W Taylor, and Thomas B Moeslund. Which tokens to use? investigating token reduction in vision transformers. In *CVPR*, pages 773–783, 2023. 3

[11] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *AAAI*, pages 852–860, 2022. 3, 4, 5

[12] Qiaolin He, Zefeng Lu, Zihan Wang, and Haifeng Hu. Graph-based progressive fusion network for multi-modality vehicle re-identification. *TITS*, pages 1–17, 2023. 3, 7

[13] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021. 2, 7

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 6

[15] Yan Huang, Qiang Wu, JingSong Xu, and Yi Zhong. Sbsgan: Suppression of inter-domain background shift for person re-identification. In *CVPR*, pages 9527–9536, 2019. 2

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *CVPR*, 2023. 3

[17] Hongchao Li, Chenglong Li, Xianpeng Zhu, Aihua Zheng, and Bin Luo. Multi-spectral vehicle re-identification: A challenge. In *AAAI*, pages 11345–11353, 2020. 1, 2, 6, 7

[18] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 7

[19] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, Xuesheng Qian, and Xiaoyun Yang. A video is worth three views: Trigeminal transformers for video-based person re-identification. *arXiv preprint arXiv:2104.01745*, 2021. 2

[20] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, pages 13334–13343, 2021. 2

[21] Xuehu Liu, Chenyang Yu, Pingping Zhang, and Huchuan Lu. Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification. *TNNLS*, 2023. 2

[22] Xuehu Liu, Pingping Zhang, and Huchuan Lu. Video-based person re-identification with long short-term representation learning. In *ICIG*, pages 55–67. Springer, 2023. 2

[23] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, pages 319–335. Springer, 2022. 3

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, pages 10012–10022, 2021. 3

[25] Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *AAAI*, pages 1835–1843, 2023. 2

[26] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, pages 1487–1495, 2019. 7

[27] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989. 5

[28] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers. *arXiv preprint arXiv:2110.03860*, 2021. 3

[29] Wenjie Pan, Linhan Huang, Jianbao Liang, Lan Hong, and Jianqing Zhu. Progressively hybrid transformer for multi-modal vehicle re-identification. *Sensors*, 23(9):4206, 2023. 3, 7

[30] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *ICCV*, pages 5399–5408, 2017. 7

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3

[32] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *ICCV*, pages 1025–1034, 2021. 7

[33] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *NIPS*, 34: 13937–13949, 2021. 3

[34] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 2, 6, 7

[35] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020. 2, 7

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 6

[37] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *CVPR*, pages 5794–5803, 2018. 2

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 8

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[40] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018. 2, 7

[41] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. Nformer: Robust person re-identification with neighbor transformer. In *CVPR*, pages 7297–7307, 2022. 2

[42] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *ICCV*, pages 12026–12035, 2021. 2

[43] Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. Top-reid: Multi-spectral object re-identification with token permutation. *arXiv preprint arXiv:2312.09612*, 2023. 1, 3, 6, 7

[44] Zi Wang, Chenglong Li, Aihua Zheng, Ran He, and Jin Tang. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In *AAAI*, pages 2633–2641, 2022. 1, 2, 7

[45] Peilei Yan, Xuehu Liu, Pingping Zhang, and Huchuan Lu. Learning convolutional multi-level transformers for image-based person re-identification. *Visual Intelligence*, 1(1):24, 2023. 2

[46] Bin Yang, Jun Chen, and Mang Ye. Top-k visual tokens transformer: Selecting tokens for visible-infrared person re-identification. In *ICASSP*, pages 1–5. IEEE, 2023. 3, 5

[47] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 44(6):2872–2893, 2021. 2, 7

[48] Haoli Yin, Jiayao Li, Eva Schiller, Luke McDermott, and Daniel Cummings. Graft: Gradual fusion transformer for multimodal re-identification. *arXiv preprint arXiv:2310.16856*, 2023. 7

[49] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. Tf-clip: Learning text-free clip for video-based person re-identification, 2023. 2

[50] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *ACM MM*, pages 516–525, 2021. 2

[51] Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *CVPR*, pages 14133–14142, 2023. 2

[52] Jiajian Zhao, Yifan Zhao, Jia Li, Ke Yan, and Yonghong Tian. Heterogeneous relational complement for vehicle re-identification. In *ICCV*, pages 205–214, 2021. 7

[53] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *AAAI*, pages 3529–3537, 2021. 1, 2, 6, 7

[54] Aihua Zheng, Xianpeng Zhu, Zhiqi Ma, Chenglong Li, Jin Tang, and Jixin Ma. Multi-spectral vehicle re-identification with cross-directional consistency network and a high-quality benchmark. *arXiv preprint arXiv:2208.00632*, 2022. 3, 6, 7

[55] Aihua Zheng, Ziling He, Zi Wang, Chenglong Li, and Jin Tang. Dynamic enhancement network for partial multi-modality person re-identification. *arXiv preprint arXiv:2305.15762*, 2023. 2, 7

[56] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020. 6

[57] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3702–3712, 2019. 2, 7

[58] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *CVPR*, pages 4692–4702, 2022. 2

[59] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Gaopan Huang, Honglin Qiao, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*, 2021. 2

[60] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *European Conference on Computer Vision*, pages 198–214. Springer, 2022. 2